目次

第 1 章

多変量解析の基礎

1.1　多変量解析手法の概要

一般的にいえば，3 変量以上を同時に考慮して分析する統計解析手法を**多変量解析**と呼ぶ。

多変量解析では，複数個の変数の重み付き合計値を用いる。この際，それぞれの解析の目的により重みの付けかたが異なる。

p 個の変数 x_1，x_2，$\cdots$，x_p が，重み w_1，w_2，$\cdots$，w_p で重み付けされた f を，**合成変数**と呼ぶ。

$$f = w_1\, x_1 + w_2\, x_2 + \cdots + w_p\, x_p \tag{1.1}$$

なお，全ての重みが等しい場合，すなわち $w_1 = w_2 = \cdots = w_p = 1$ のときは，

$$t = x_1 + x_2 + \cdots + x_p \tag{1.2}$$

となり，t は全ての変数の和（合計得点）であるが，これも合成変数の一つである。例えば入学試験において，受験科目の得点の合計点によって合否を決めるということが多いが，理数系の得点の重みを大きくしたり，文化系の得点の重みを大きくした重み付け合計得点を用いて合否を決めることもできる。それぞれの合計得点によって選抜された受験生は異なる特性を持つことになる。

多変量解析では，主観的に重みを決めるのではなく，目的に合う重みをデータに基づいて客観的に決めるのである。

例題 1.1

対象が A，B の 2 群に前もって分類されていて，それぞれの対象について 2 変数 x_1，x_2 の測定値がある。対象が，図 1.1 の ようにプロットされているとき，図中に 2 群を分ける直線を 1 本引き，その直線の式を求めよ。その式を（1.1）式のように変換できるか。

解

2 つの群を完全に分ける直線は引けないが，最もよく分離する直線は引ける。その直線の式は，傾き a，切片 b とすると $x_2 = a\, x_1 + b$ のように表せる。この式は $x_2 - a\, x_1 - b = 0$ と書き直せるので，$f = -a\, x_1 + x_2 - b$ とすれば（1.1）式で表される合成変数の形をしている。2 群を分割する直線の上の点では $f = 0$ となる。x_1，x_2 に各対象の実測値を代入し，f が正の値になるか負の値になるかによって，どちらの群であるかを予測することができる。

この例題のように 2 変数で対象数も少ないときには目分量で分割線を引く（すなわち各変数の重みを適当に決める）ことができるが，3 変数以上や対象数が多くなったときにはデータに基づいた客観的な重み付けをする必要がある。

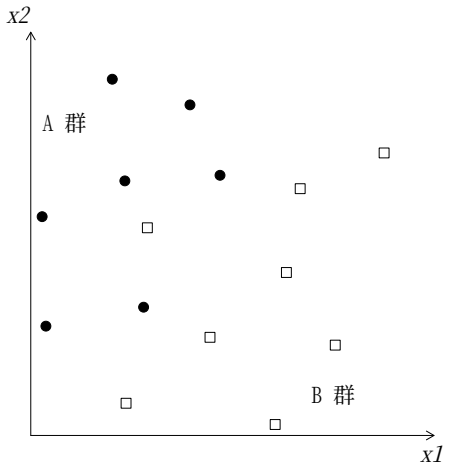

図 1.1　群を分ける直線を引く

1.2　目的による多変量解析手法の分類

多変量解析として分類されるものには多くの手法があるが，よく用いられるものを表 1.1 に挙げる。

表 1.1　多変量解析の手法

外的基準	説明変数	手法
量的基準	量的変数	重回帰分析
	質的変数	数量化 I 類
質的基準	量的変数	判別分析
	質的変数	数量化 II 類
なし	量的変数	主成分分析，因子分析
	質的変数	数量化 III 類

この表において，**外的基準**というのは複数個の**説明変数**によって予測される変数を意味する。

例えば，食物摂取量からその人の体重を予測したり（重回帰分析），健康診断時の検査結果からその人が健康か病気かを予測する（判別分析）ような場合，食物摂取量や健康診断時の検査結果は体重や健康状態を説明するものである。

通常，外的基準は何らかの方法で既に存在するものであるから，それを予測することにはどういう意味があるだろうか。一つには，外的基準とされる変数が測定しにくいものである場合に，簡単に測定できる説明変数に基づいて予測式のようなものが作れるとすれば，外的基準を直接測定する必要がなくなるというメリットがある。もう一つは，外的基準と説明変数の関係を表す式は，説明変数が外的基準をどのように規定しているかについての知識を与えてくれるという点である。外的基準がない場合（主成分分析，因子分析など）には分析に使用される変数は全て同等に扱われ，変数相互の関連を明らかにすることが目的になる。

1.3 測定データの特性別の分類

データは測定（観察) されたある特性を持つ。データ解析においては，何種類かの特性を持つデータを用いて解析を行う。特性別に表 1.2 のように分類できる。

表 1.2 各分析手法で使用される変数の種別

分析手法（目的）	使用される変数の種別
群の比較	群分け変数，分析対象変数（平均値などが取られる変数）
判別分析	群分け変数（外的基準），説明変数（独立変数）
回帰分析	従属変数（外的基準），独立変数（説明変数）
相関分析	分析対象変数（全ての変数が同格）
時系列分析（回帰分析）	時間変数（独立変数），分析対象変数（従属変数）
生存率解析（生命表）	時間変数，エンドポイント変数
生存率解析（回帰分析）	時間変数，エンドポイント変数，独立変数（説明変数，共変量）

分析対象変数 相関分析，例えば，主成分分析や因子分析（第 4，5 章参照）などでは，分析に用いられる変数は相互間の関係を明らかにするのが目的であるため，全ての変数は同格に扱われる。研究対象を解明できると期待される変数を網羅的に用意すればよい。例えば，食生活と身体状況の関連を知りたいような場合，各栄養素の摂取量，運動量，喫煙・飲酒，身体計測値，臨床検査値に関するデータを採取しデータファイルを作ればよい。

群を表す変数 統計解析で群の比較を行う手法としては，2 群の平均値の差の検定（t 検定），多群の平均値の差の検定（一元配置分散分析）などがある。例えば，男と女の体脂肪量の平均値に差があるかとか，40 歳代・50 歳代・60 歳代の収縮期血圧の平均値に差があるかどうかなどを知りたい時に，群ごとの例数・平均値・標準偏差をもとにして計算が行われる。データとしては各群に属するケースの測定値が必要であるのはいうまでもないが，個々のケースがどの群に属するのかを表す変数が必要になる。性別を表すには，男を 1，女を 2 で表すようにしておけばよい。年齢によって群分けを行う場合には，前もって 40 歳代のものに 1，50 歳代のものに 2，60 歳代のものに 3 という数値を与えるようにカテゴリー化しておいてもよい。

群を表す変数が必要になる分析手法はこれらの他に，判別分析（第 3 章参照）が挙げられる。例えば，治療効果の判定が医師によって，「著効」,「有効」,「不変」,「悪化」などのように表されていて，これらの予後が初診時の臨床所見によって判別できるかどうか，判別できるとすればどの要因が最も予後を左右するかを知りたいような場合である。各患者はその予後によって，群分けされていることになる。各患者について，「著効」を 1,「有効」を 2,「不変」を 3,「悪化」を 4 などのように表した群を表す変数を準備しなければならない（次項も参照のこと)。

従属変数と独立変数 回帰分析（第 2 章参照）の場合は，いくつかの（原因となる）変数をもとにして，（結果である）1 つの変数の値を予測する。例えば高血圧症の治療において，投薬量，食塩摂取量，体重変化をもとにして血圧変動量を予測できるかどうか，もし予測できるならばどの要因が最も予測値を左右するかを知りたいような場合である。このような場合には，予測に使用する変数（独立変数あるいは説明変数）と，予測される変数（従属変数）が共に必要である。

判別分析は，回帰分析と密接な関係にある。前述の判別分析の場合の群を表す変数は従属変数であるというようにも考えられる。従って，判別分析の場合には，群を予測するために使用される変数（説明変数）も用意しなければならない。

回帰分析と判別分析は密接な関係にあり，前者で使用される「従属変数」と後者で使用される「群分け変

数」を総称して，**外的基準**と呼ぶ場合もある。

時間変数　時間を表す変数は 2 通りある。

1 つは，いわゆる**時系列分析**の場合に用いられる概念である。例えば，ある薬物を経口投与し，1 時間おきに測定した血中濃度の動向をモデルにあてはめるような場合である。または，ある疾患の死亡率の年次推移データをもとにして，数年先の死亡率の動向を予測するような場合である。これらの場合はいずれも，時間変数を独立変数とした回帰分析である。

もう 1 つは，生存率解析（第 6 章参照）の場合に用いられる概念である。例えば，胃癌の切除手術をした患者について，生存期間を解析対象にするような場合である。このような場合には，時間変数は結果（治療効果）を表す変数として用いられる。時系列分析の場合と異なるのは，各患者ごとに開始時点・終了時点が異なることである。生存期間は前もって計算した値をデータファイル中に用意する必要がある。生存率の解析においては生存期間を従属変数（外的基準）として，それを左右する臨床所見を独立変数（説明変数）とした回帰分析とみなすこともできる。

エンドポイント変数　生存率の解析においては，最終的に全ての患者が死亡することは前提としない。あるケースは研究終了（データ解析開始）時点では生存している。このようなことから，各ケースが死亡しているか生存しているかの区別（研究終了時の患者の状態＝**エンドポイント**）を表す変数が必要になる。エンドポイント変数は前述の群分け変数の 1 種であるが，生存率解析においては特別な意味合いで利用される。

1.4　連続変数とカテゴリー変数

表 1.1 に示した諸分析手法では，分析に使用する変数の平均値を求めたり相関係数を求めたりするような過程が含まれる。つまり，これらの変数が連続変数であることを仮定している。

ところで，データとしては質問への回答が「はい」,「どちらでもない」,「いいえ」のように記録されたり，職業が「自営業」,「サラリーマン」,「主婦」のように記録されることもある。前者は順序尺度，後者は名義尺度で測定されるデータである。これらを総称して，**カテゴリーデータ**と呼ぼう。例えば前者のような順序尺度変数において，各選択肢に，2，1，0 のように数値を与えれば多変量解析を行えるように思うかもしれないが，それは誤りである。このような場合には，それぞれの回答に対して一つの仮想的な変数（**ダミー変数**）を考えるとよい。ダミー変数とは 0 か 1 か，いずれかの値しか取らないものである。

簡単にするために,「はい」,「どちらでもない」,「いいえ」の 3 つのカテゴリーを持つ変数 x を考え，各カテゴリーに 2，1，0 という数値を与えるものとする。次に，x が取る 3 種類の数値に対応して 3 個のダミー変数 d_1，d_2，d_3 を考える。もとの変数とダミー変数は表 1.3 のように対応させる。

表 1.3 で，$w_1 = 0$，$w_2 = 1$，$w_3 = 2$ とすると，3 つのダミー変数でもとの変数を表現できることがわかる。

$w_1 = 0$ ということは，最初のダミー変数は合成変数に寄与しないので，考えなくてもよいということになるので，表 1.4 のようになる。すなわちカテゴリーが 3 つのときは，2 つのダミー変数でもとの変数を表すことができるということである。

表 1.3　もとの変数とダミー変数の対応

x のとる値	d_1	d_2	d_3	合成変数（小計）
0	1	0	0	$f = w_1d_1 + w_2d_2 + w_3d_3 = w_11 + w_20 + w_30 = w_1$
1	0	1	0	$f = w_1d_1 + w_2d_2 + w_3d_3 = w_10 + w_21 + w_30 = w_2$
2	0	0	1	$f = w_1d_1 + w_2d_2 + w_3d_3 = w_10 + w_20 + w_31 = w_3$

このやり方では，1 個の変数を表すために 2 個のダミー変数が必要となり，一見無駄なように思えるが，以下のような利点がある。

表 1.4　もとの変数とダミー変数の対応

x	d_2	d_3	合成変数（小計）
0	0	0	$f = w_2d_2 + w_3d_3 = w_20 + w_30 = 0$
1	1	0	$f = w_2d_2 + w_3d_3 = w_21 + w_30 = w_2$
2	0	1	$f = w_2d_2 + w_3d_3 = w_20 + w_31 = w_3$

- ダミー変数は 2 種類の値のいずれかを取る名義尺度（順序尺度）変数（二値データともいう）と考えられるが，二値データの場合に限り，そのような変数は間隔尺度（比尺度）変数と見なすことができる。
- $f = w_2\,d_2 + w_3\,d_3$ という合成変数の重みは多変量解析では目的に添うように与えられる。すなわち，$w_2 = a$，$w_3 = b$ のように最適な重み付けができるということは x が 0，a，b という値を取るものとして扱うことになる。x が 0，1，2 という値をとると仮定するのは客観的ではないが，ダミー変数を用いれば客観的な重み付けができる。

ダミー変数を用いればあらゆる多変量解析は可能であるが，計算や解釈を容易にするために林知己夫により開発された**数量化理論**がある。代表的なものに，数量化 I 類，数量化 II 類，数量化 III 類があるが，それぞれ重回帰分析，判別分析，主成分分析と対応する。

第 2 章

回帰分析

2.1　重回帰分析

重回帰分析の目的は，いくつかの変数に基づいて，別の変数を予測することである。

単純にするために，一つの変数（**独立変数**）から別の変数（**従属変数**）を予測することを考える。独立変数を X，従属変数を Y，Y の予測値を $\hat{Y}$ としたとき，図 2.1 のような，各点の近くを通るような直線 $\hat{Y} = aX + b$ が考えられる。この直線式の定数は**最小二乗法**により求めることができる。

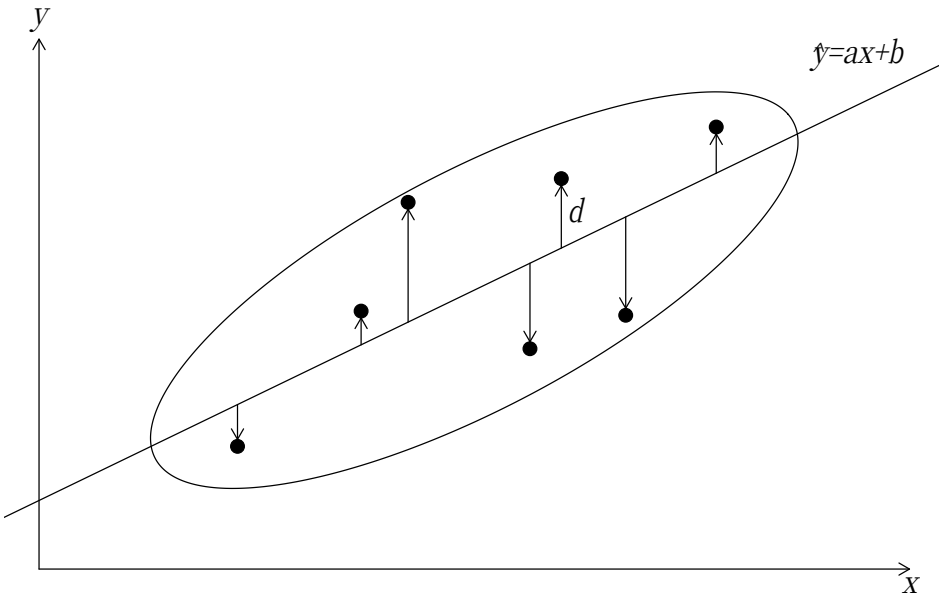

図 2.1　回帰直線とは何か

ケース数を n，変数 X と変数 Y の平均値をそれぞれ $\bar{X}$，$\bar{Y}$ として，

$$\begin{cases} a = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ b = \bar{Y} - a\bar{X} \end{cases} \tag{2.1}$$

となる。

複数個の独立変数がある場合には，回帰式は一般的に

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p \tag{2.2}$$

のようになる。(2.2) 式の右辺も合成変数（変数の重みつき合計）の形になっていることがわかる。

2.1.1 偏回帰係数の求め方

従属変数が Y，p 個の独立変数が X_i（$i = 1, 2, \ldots, p$）の場合の重回帰モデルは，β_0，β_1，…，β_p を未知母数，ϵ が平均 0，分散 σ^2 の正規分布確率変数とすると

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{2.3}$$

で表される。

(2.3) 式に含まれる未知母数を推定する（b_0，b_1，…，b_p とする）ことにより，従属変数の予測値 $\hat{Y}$ は，重回帰式 $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p$ により求められる（ϵ は予測誤差となる）。

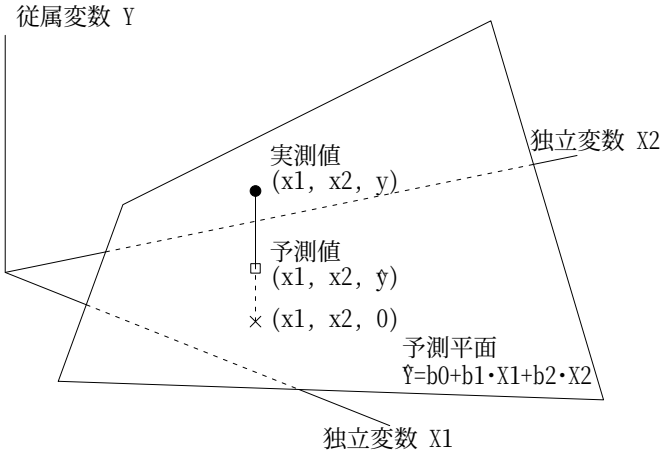

図 2.2　二つの独立変数による従属変数の予測

以下では図 2.2 のような独立変数が 2 個の場合を考えることにする。予測値 $\hat{Y}$ は予測平面 $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$ 上にある。実測値 Y との差（残差）$e_i = Y_i - \hat{Y}_i$ は正負の符号を持つので，その 2 乗和が最小になるように独立変数にかけられる重み b_i（**偏回帰係数**）および定数項 b_0 を定める。この手法を**最小二乗法**と呼び，得られる係数を**最小二乗推定値**と呼ぶ。

$$\begin{aligned} Q &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n} \left\{ Y_i - (b_0 + b_1 X_{i1} + b_2 X_{i2}) \right\}^2 \rightarrow \text{最小にする} \end{aligned} \tag{2.4}$$

（2.4）式を，b_0，b_1，b_2 で偏微分して 0 とおく。

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\displaystyle\sum_{i=1}^{n}\Big\{Y_i - (b_0 + b_1 X_{i1} + b_2 X_{i2})\Big\} = 0 \\ \dfrac{\partial Q}{\partial b_1} = -2\displaystyle\sum_{i=1}^{n} X_{i1}\Big\{Y_i - (b_0 + b_1 X_{i1} + b_2 X_{i2})\Big\} = 0 \\ \dfrac{\partial Q}{\partial b_2} = -2\displaystyle\sum_{i=1}^{n} X_{i2}\Big\{Y_i - (b_0 + b_1 X_{i1} + b_2 X_{i2})\Big\} = 0 \end{cases} \tag{2.5}$$

変数 Y，X_1，X_2 の平均値を $\bar{Y}$，$\bar{X}_1$，$\bar{X}_2$ としたとき，（2.5）式の最初の式から，

$$b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 \tag{2.6}$$

が得られる。

（2.5）式の 2 つ目以降の式に（2.6）式を代入し，さらに独立変数 X_i，X_j 間の変動・共変動

$$S_{ij} = \sum_{k=1}^{n}(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j) \tag{2.7}$$

および，独立変数 X_i と従属変数 Y の共変動

$$S_{iy} = \sum_{k=1}^{n}(X_{ki} - \bar{X}_i)(Y_k - \bar{Y}) \tag{2.8}$$

を代入して整理すると，

$$\begin{cases} b_1 S_{11} + b_2 S_{12} = S_{1y} \\ b_1 S_{21} + b_2 S_{22} = S_{2y} \end{cases} \tag{2.9}$$

という連立方程式（**正規方程式**と呼ぶ）が得られる。

一般的な表し方としては，独立変数間の変動・共変動行列を $\boldsymbol{S}$，独立変数と従属変数間の共変動ベクトルを $\boldsymbol{c}$，偏回帰係数ベクトルを $\boldsymbol{b}$ とすると，

$$\boldsymbol{S}\,\boldsymbol{b} = \boldsymbol{c} \tag{2.10}$$

となるので，$\boldsymbol{S}$ の逆行列を $\boldsymbol{S}^{-1}$ とすれば，偏回帰係数は（2.11）式で求められる。

$$\boldsymbol{b} = \boldsymbol{S}^{-1}\,\boldsymbol{c} \tag{2.11}$$

定数項 b_0 は（2.6）式から求められる。

重回帰分析の結果は，表 2.1 のように表される。分析には $X1$〜$X4$ の 4 つの独立変数が使われ，偏回帰係数の列を見ることにより，$31.9670 \times X1 + 48.0183 \times X2 - 29.536 \times X3 - 454.373 \times X4 + 594.562$ という予測式が得られたことが読みとれる。

表 2.1　重回帰分析結果の表示例

独立変数	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数
$X1$	31.9670	6.840	4.673	<0.001	0.620
$X2$	48.0183	12.322	3.897	<0.001	0.478
$X3$	−29.536	11.990	2.463	0.018	−0.240
$X4$	−454.373	171.514	2.649	0.011	−0.284
定数項	594.562	789.895	0.753	0.456	

2.1.2　標準化偏回帰係数

同じデータを使って分析しても，ある独立変数を 10 倍したときの分析結果と 1/10 にしたときの分析結果を比べると，予測値は同じになるが，偏回帰係数の大きさは違ったものになる。例えば，測定単位を cm から mm に変えると，変数のとる値は 10 倍になり，偏回帰係数は 1/10 の大きさになる。表 2.1 の例だと，$X1$ が 10 倍になると，予測式は，$3.19670 \times (10 \times X1) + 48.0183 \times X2 - 29.536 \times X3 - 454.373 \times X4 + 594.562$ となるわけである。

また，g で測定される変数の偏回帰係数と cm で測定される変数の偏回帰係数を比較しても意味がない。

つまり，偏回帰係数が大きいから予測値に大きな影響を与えている（予測するために重要である）という判断はできないのである。そこで，偏回帰係数の大きさが測定単位によって左右されないようにするためには，各変数を平均 0，分散 1 に標準化しておくとよい。このようなデータに基づいて前項に示した計算により得られるものを，**標準化偏回帰係数** b'_i と呼ぶ。標準化偏回帰係数は，ある独立変数が 1 標準偏差変動したときに，標準化された従属変数が何単位変動するかを表す。

なお，偏回帰係数と標準化偏回帰係数には，

$$b'_i = b_i \sqrt{S_{ii} / S_{yy}} \tag{2.12}$$

の関係があるので，実際には標準化したデータを使用して再分析しなくても標準化偏回帰係数を求めることができる。

また，コンピュータで計算するような場合には誤差の影響等を考慮し，(2.9) 式の正規方程式を変動共変動行列（ベクトル）ではなく相関係数行列に置き換えておけば，標準化偏回帰係数の方が先に求まる。その後で (2.12) 式から導ける (2.13) 式によって，偏回帰係数を求めることになる。

$$b_i = b'_i \sqrt{S_{yy} / S_{ii}} \tag{2.13}$$

標準化偏回帰係数が負の値である場合は，独立変数が大きくなると従属変数は小さくなることを表し，正の値である場合は独立変数が大きくなると従属変数も大きくなることを表す。従属変数への影響力の大きさは，標準化偏回帰係数の絶対値を考えなくてはならないのである。表 2.1 の例では，偏回帰係数の絶対値が最も大きいのは $X4$ であるが，標準化偏回帰係数の絶対値が最も大きいのは $X1$ であり，従属変数に最も大きな影響を与えるのは $X4$ ではなくて $X1$ である。

2.1.3　偏回帰係数の検定

偏回帰係数は，ある独立変数が従属変数にどのように影響を及ぼすかを表すものである。正の値を取る場合には独立変数が大きな値であれば従属変数も大きな値になり，逆に負の値を取る場合には，独立変数が大きな値であれば従属変数は小さな値になることを意味する。もし，偏回帰係数の値が 0 に近ければ[*1]，その独立変数が大きな値をとろうが小さな値をとろうが従属変数に何ら影響を及ぼさない。そこで，ある独立変数が従属変数に影響を及ぼすかどうかについて，偏回帰係数が 0 であるかどうかの検定が必要になる。

検定は独立変数ごとに以下のような手順で行われる。

帰無仮説 H_0:「$b_i = 0$（$i = 1, 2, \ldots, p$）」

対立仮説 H_1:「$b_i \neq 0$（$i = 1, 2, \ldots, p$）」

有意水準 α で両側検定を行う（片側検定も定義できる）。

偏回帰係数 b_i の標準誤差を $SE(b_i)$ とすると，s^{ii} を $\boldsymbol{S}^{-1}$ の要素として，表 2.2 に示す MS_e を用いて，

$$SE(b_i) = \sqrt{s^{ii} MS_e} \tag{2.14}$$

[*1] 正確に言うとすれば，標準化偏回帰係数が 0 に近ければということである。

$$t_0 = |b_i| / SE(b_i) \tag{2.15}$$

t_0 は，自由度が $n-p-1$ の t 分布に従う。有意確率を $P = \Pr\{|t| \geqq t_0\}$ とすると，

- $P > \alpha$ のとき，帰無仮説を採択する。「偏回帰係数は 0 である」
- $P \leqq \alpha$ のとき，帰無仮説を棄却する。「偏回帰係数は 0 でない」

表 2.1 の例では，「偏回帰係数」の列に示されている数値を「標準誤差」の列に示されている数値で割ったものが t 値の列であり，それに基づいて P 値の列が計算されている。この例ではいずれも 5% の有意水準のもとで帰無仮説は棄却される（偏回帰係数は 0 ではない）ことを表している。

2.1.4 定数項の検定

定数項が 0 であるかどうかは，たいていの場合は不要である。しかし，理論的に考えて独立変数の値が 0 であるときに予測値が 0 である（回帰超平面が原点を通る）ことがわかっている場合に，分析によって得られた定数項が 0 であるかどうかの検定が必要になる。

検定は以下のような手順で行われる。

帰無仮説 H_0: 「$b_0 = 0$」

対立仮説 H_1: 「$b_0 \neq 0$」

有意水準 α で両側検定を行う（片側検定も定義できる）。

定数項 b_0 の標準誤差は，

$$SE(b_0) = \sqrt{\left\{ \frac{1}{n} + \sum_{i=1}^{p} \sum_{j=1}^{p} \bar{X}_i \bar{X}_j s^{ij} \right\} MS_e} \tag{2.16}$$

となり，

$$t_0 = |b_0| / SE(b_0) \tag{2.17}$$

t_0 は，自由度が $n-p-1$ の t 分布に従う。有意確率を $P = \Pr\{|t| \geqq t_0\}$ とすると，

- $P > \alpha$ のとき，帰無仮説を採択する。「定数項は 0 である」
- $P \leqq \alpha$ のとき，帰無仮説を棄却する。「定数項は 0 でない」

表 2.1 の例では，「定数項」と記された行の P 値を見ると，5% の有意水準のもとで帰無仮説は採択される（定数項は 0 でないとはいえない）ことを表している。

2.1.5 偏回帰係数および定数項の信頼限界

$(1-\alpha)\,100\,\%$ 信頼限界を求める。

自由度が $\nu = n-p-1$ の t 分布において，上側確率が $\alpha/2$ となるパーセント点を $t_{(\alpha/2,\,\nu)}$，$SE(b_i)$ を（2.14），（2.16）式中のものとすると，信頼限界は（2.18）式で与えられる。

$$b_i \pm t_{(\alpha/2,\,\nu)} SE(b_i) \tag{2.18}$$

2.1.6 回帰の分散分析

2.1.3 項では，個々の独立変数が従属変数に影響を及ぼしていると言えるかどうかの検定であったが，本項は分析に用いた全ての独立変数で従属変数が予測できると言えるかどうかの検定である。この検定での帰無仮説は，得

られる重回帰式は従属変数の予測に役に立たないということでもある。

帰無仮説 H_0:「分析に使用した独立変数で，従属変数は説明できない」

対立仮説 H_1:「分析に使用した独立変数で，従属変数は説明できる」

有意水準 α で両側検定を行う（片側検定は定義できない）。

従属変数の変動は回帰によって説明できる部分と，説明できない部分に分解される。

$$\begin{aligned} \sum_{i=1}^{n}(Y_i-\bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \\ S_t &= S_r+S_e \\ \text{全変動} &= \text{回帰で説明される変動} + \text{残差（回帰で説明できない変動）} \end{aligned} \tag{2.19}$$

S_t は従属変数の変動であるので，一番最初に求めることができる。次に求めるのは，S_r である。これは（2.19）式によって求めるよりは，独立変数と従属変数の共変動 S_{iy} が求められているので（2.20）式によった方が簡単である。

$$S_r=\sum_{i=1}^{p} b_i\, S_{iy} \tag{2.20}$$

最後に，S_e は $S_e = S_t - S_r$ の関係式から求める。

これらに基づいて，表 2.2 のような**分散分析表**の形にまとめる。

表 2.2　回帰の分散分析表

変動要因	平方和	自由度	平均平方	F 値
回帰	S_r	p	$MS_r = S_r / p$	MS_r / MS_e
残差	S_e	$n-p-1$	$MS_e = S_e / (n-p-1)$	
全体	S_t	$n-1$	$MS_t = S_t / (n-1)$	

F 値は自由度が（p, $n-p-1$）の F 分布に従う。有意確率を P とすると，

- $P > \alpha$ のとき，帰無仮説を採択する。「分析に使用した独立変数で，従属変数は説明できない」
- $P \leqq \alpha$ のとき，帰無仮説を棄却する。「分析に使用した独立変数で，従属変数は説明できる」

2.1.7　重相関係数と寄与率

従属変数の全変動のうち，回帰によって説明できる割合（寄与率）は重相関係数の 2 乗（R^2; **決定係数**）に等しい（$0 \leqq R^2 \leqq 1$ である）。重相関係数も重相関係数の 2 乗も 0 から 1 までの値をとる。重相関係数の 2 乗が 0 に近ければ重回帰式は従属変数の予測に役に立たないことを意味しており，1 に近ければ近いほど予測に役に立つことを意味する。重相関係数の検定は，前述の「回帰の分散分析」と同じ意味である。したがって，回帰の分散分析の結果として帰無仮説が棄却されれば，重相関係数も有意である。

$$R^2 = 1-\frac{S_e}{S_t}=\frac{S_r}{S_t} \tag{2.21}$$

予測に有効な変数でなくても，独立変数を増やしてゆけば寄与率はだんだんと 1 に近付く。そのため，寄与率が高くなったのが追加された独立変数の効果なのかどうかがわからなくなることがある。このため，**自由度調整済みの重相関係数の 2 乗**（R^{2*}）が定義される。

$$R^{2*} = 1-\frac{S_e/(n-p-1)}{S_t/(n-1)}=1-\frac{MS_e}{MS_t} \tag{2.22}$$

$R^2 \neq 1$ である限り，R^{2*} は R^2 よりも小さい。R^{2*} が増加する限り，追加された独立変数は有効であることを意味する。ある変数を独立変数として加えたとき，R^{2*} が前よりも減少したとすれば，その変数は重回帰式に組み込むのはふさわしくないということを意味する。

2.1.8 多重共線性

予測を行うための重回帰式を作るときに，独立変数の候補がたくさんあるときに，全部を使う必要はないかも知れないし逆に特定の条件を持つような変数を使うと問題が起きることもありうる。そのようなものの一つが多重共線性という概念である。

原則として，変数選択を行わない場合には，相関の高い変数を一緒にして重回帰式には含めないほうがよい。どの程度の相関なら一緒に含めて良いかは一概にはいえない。しかし例えばもし，独立変数の中に相互に従属なものが含まれていると（例えば変数 A，B とその合計値 $C = A + B$ が共に含まれていると）分析は失敗する。

場合によっては，各独立変数と従属変数との相関係数の符号と，偏回帰係数の符号が一致しない場合が生ずる。これは，「予測を行う」という観点から偏回帰係数が定められるので，重回帰式に含まれた変数相互間の関連で符号が決められるためである。このようなことが起きるのは，独立変数間に相関の高いものが混ざっていることが原因である（ある変数で予測しすぎた部分を別の変数で打消しているような場合がある）。しかし，このようなことは因果関係を考える上では不都合なので，符号が一致しない独立変数を除いた重回帰式を探索するとよいであろう。

重回帰分析において，以上のような不都合な状態が生じることを，「**多重共線性**がある」という[*2]。

独立変数間の相関係数行列の逆行列の要素を r^{ii} としたとき，

$$R = \sqrt{1 - 1 / r^{ii}} \tag{2.23}$$

は，独立変数 i を残りの独立変数で予測するときの重相関係数になっている。したがって，この数値が大きいものは独立変数としてふさわしくないことを表す。これと同じことであるが $1 / r^{ii}$ を**トレランス**，r^{ii} を**分散拡大要因 VIF** と呼ぶことがある。この場合には，トレランスが低い（分散拡大要因が大きい）独立変数は除く方がよいことを表す。どの程度の場合に多重共線性があるとするかは相対的なものであるが，トレランスが 0.1 以下（分散拡大要因が 10 以上）のときには，そのような変数を取り除いて再分析してみるべきであろう。

2.1.9 変数選択

重回帰分析を行うときには，多重共線性を避けるように注意が必要であるほか，多くの独立変数から有用な少数個の独立変数を精選して重回帰式を作ることが重要になることもある。理論的基盤があるときはそれに従って独立変数を選択することが可能であるが，探索的に行うためには「**総当たり法**」や，以下に示すような変数の選択法がある。

変数増加法 最初に最も予測に有効な独立変数（従属変数との相関係数が最も大きいもの）を重回帰式に取り入れる。次の段階では，残りの独立変数の中で最も予測に有効な独立変数を取り入れる。予測精度の改善が一定限度以上である間，この操作を繰り返す。

変数減少法 最初に全ての独立変数を含む重回帰式を作る。次に，その中から最も予測に有効でない独立変数を除去する。予測精度の低下が一定限度以内である間，この操作を繰り返す。

変数増減法 変数増加法では，いったん重回帰式に取り込まれた独立変数は除去されることはないが，後の段階になってそれまでに取り込まれた独立変数の重要性が低くなることがある。変数増減法は各段階で変数を追加した後で除去すべき独立変数がないかをチェックする。

変数減増法 最初に全ての独立変数を含む重回帰式を作る。その後に続く各段階では，まず既に重回帰式に取り入

[*2] 厳密な意味での多重共線性は，たとえば $X_i = uX_j + vX_k$ のように，独立変数が一次従属である（ある独立変数が，他の複数の独立変数の線形結合で表される）場合を指す。

れられている独立変数の中から最も予測に有効でない独立変数を除き，取り入れられていない独立変数の中に取り入れるべきものがないかをチェックする。

独立変数の追加・除去の基準としては，各変数の**偏 F 値**（(2.15) 式の右辺を二乗したもの）に基づく F_{in}，F_{out}，それを有意確率に換算した P_{in}，P_{out} がある。いずれも，各独立変数の偏回帰係数の有意性検定と関連しており，後者で例えば $P_{in} = P_{out} = 0.05$ を指定するということは，最終的な重回帰式に含まれる全ての独立変数の偏回帰係数が 0 であるという帰無仮説が有意水準 5 % で棄却されるということを意味する。

ステップワイズ変数選択によって独立変数の候補から自動的に重回帰式に取り入れる場合には，理論的に妥当な変数が必ずしも選択されないという不都合が生じる場合も多い。このような場合にはステップワイズ変数選択の結果を参考にして，変数選択を行わないで分析するのがよい。

2.1.10　回帰診断

予測値と標準化残差のプロット（残差分析）により，重回帰モデルの妥当性が検証できる。

図 2.3 のように，予測値の大小にかかわらず標準化残差が一様に散らばっていれば，重回帰モデルは妥当である。

図 2.4 のように，データの中に外れ値がある場合には，数個の点が飛離れた位置にプロットされる。

図 2.5 のように，予測値が大きく（小さく）なるにつれ標準化残差の大きさが変化するような場合には分散が不均一であり，重回帰モデルが妥当でないことを表している。

図 2.6 のように，標準化残差が曲線的な変動を見せるときは，独立変数と従属変数が曲線相関を示す。重回帰分析では，個々の説明変数は従属変数と直線相関関係にあることが仮定されている。個々の独立変数と従属変数の組合せで散布図を描き，直線相関から大幅にずれる独立変数は適当な変数変換をしてから用いた方がよい場合もある。

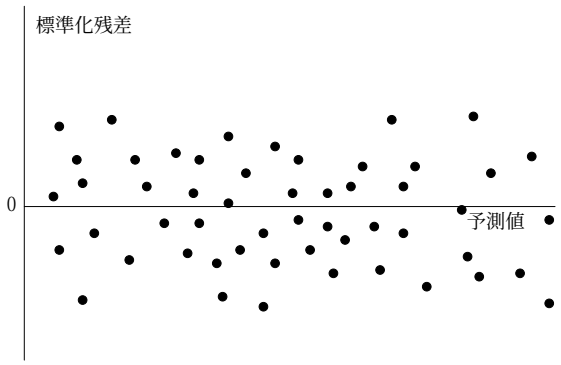

図 2.3　残差分析

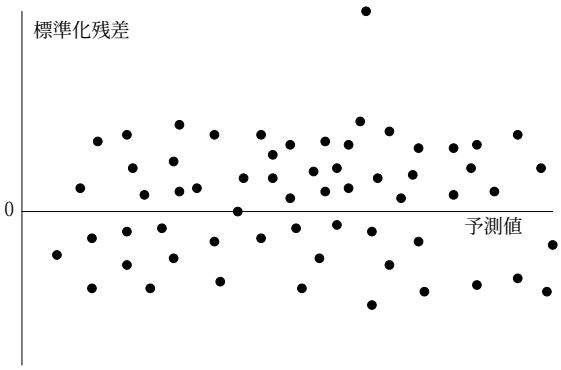

図 2.4　残差分析

2.2　多項式回帰分析

多項式回帰は，重回帰分析の特別な場合である。

p 個の独立変数 $X_1, X_2, \ldots, X_p$ を用いて，従属変数 Y を予測する重回帰式は以下のように表される。

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p \tag{2.24}$$

多項式回帰では，独立変数が 1 個（X とする）であり，X の p 次多項式による従属変数 Y の予測式が以下のように表される。

$$\hat{Y} = b_0 + b_1 X + b_2 X^2 + \cdots + b_p X^p \tag{2.25}$$

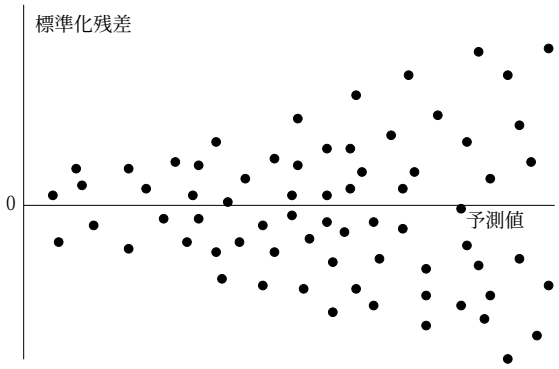

図 2.5　残差分析

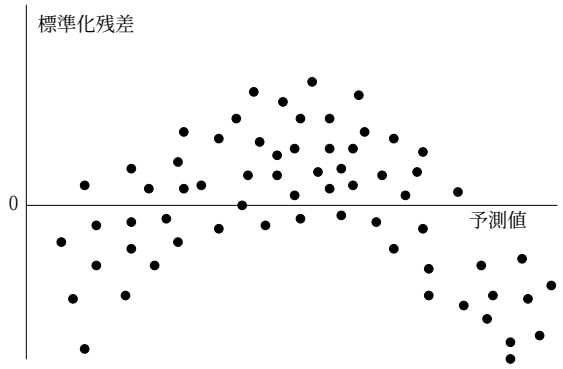

図 2.6　残差分析

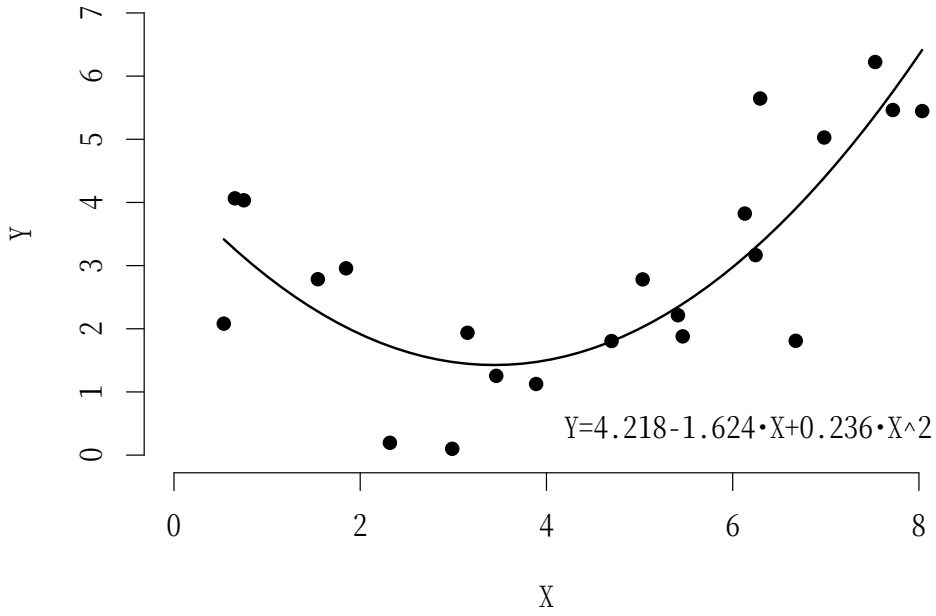

図 2.7　二次式へのあてはめ例

すなわち，重回帰式では p 個の変数が使われるが，多項式回帰では 1 個の独立変数の p 個のべき乗 X^i（$i = 1, 2, \ldots, p$）が使われるという違いがある。偏回帰係数の推定法については 2.1.1 項を参照のこと。

次数の高い多項式を使えば，データ中の独立変数の動く範囲内ではあてはまりはよくなる。しかし，その範囲以外では全く使用に堪えない予測式ができる。次数はあまり高くしないほうがよい。

データ点が n 組あるとき，$n-1$ 次式は完全に各点を通る。$n-1$ 次以上の多項式にはあてはめできない。なお，データ点が重なる場合には，これ以下の次数の多項式にしかあてはめできない場合もある。

なお，多項式回帰分析との関連から見れば，独立変数に対して任意の変数変換を行って重回帰モデルを構成してよいことがわかる。すなわち，逆数変換（$1/X_i$），対数変換（$\ln X_i$），平方根変換（$\sqrt{X_i}$）などの可能性がある（一つのモデルにこれらが混在していてもよい）。

2.3　漸近指数曲線へのあてはめと重回帰分析

独立変数の増加（減少）に伴って，次第に一定値に近づいていくような曲線を**漸近指数曲線**と呼ぶ。パラメータは 3 個であり，母回帰関数を，$f(\alpha, \beta, \gamma) = \alpha\beta^X + \gamma$ とすると，

1. $0 < \beta < 1$ かつ $\alpha < 0$ のとき，
   $X = 0$ のときに $\alpha + \gamma$ の値から始り，X が大きくなると増加し，漸近値 γ に近付く。
2. $0 < \beta < 1$ かつ $\alpha > 0$ のとき，
   $X = 0$ のときに $\alpha + \gamma$ の値から始り，X が大きくなると減少し，漸近値 γ に近付く。
3. $\beta > 1$ かつ $\alpha < 0$ のとき，
   $X = 0$ のときに $\alpha + \gamma$ の値から始り，X が大きくなると減少する。X が負の値をとるとき漸近値 γ に近付く。
4. $\beta > 1$ かつ $\alpha > 0$ のとき，
   $X = 0$ のときに $\alpha + \gamma$ の値から始り，X が大きくなると増加する。X が負の値をとるとき漸近値 γ に近付く。

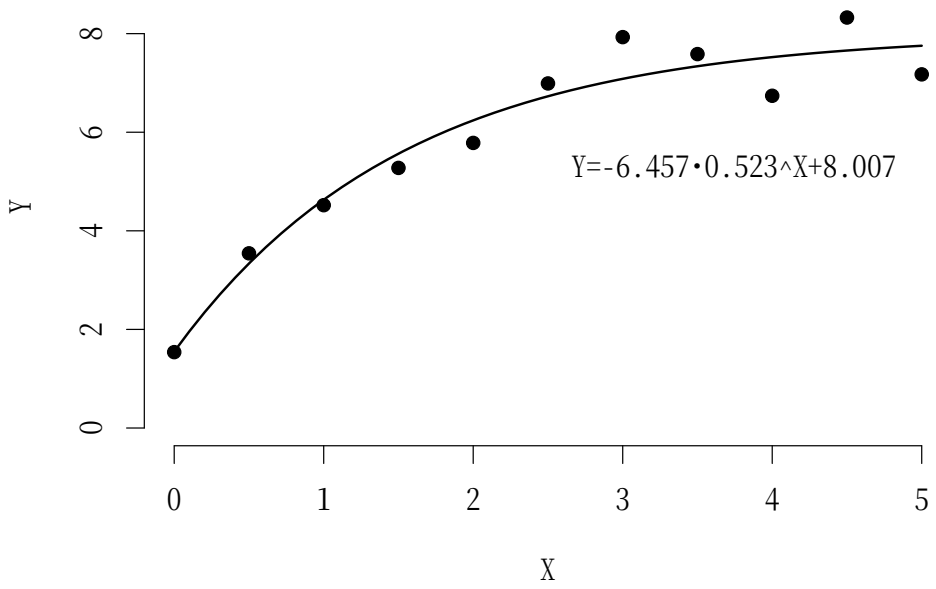

図 2.8　漸近指数曲線の概形

予測式は（2.26）式のように表される。

$$Y = a\,b^X + c \tag{2.26}$$

漸近指数関数は母数 β について非線形であるため，線形近似を行い，繰返し計算でパラメータを推定する。
u_1 を β の第 1 近似とすると，テーラー展開によって，

$$\alpha\beta^X + \gamma \quad \alpha\,u_1^X + \alpha\,(\beta - u_1)\,X\,u_1^{X-1} + \gamma \tag{2.27}$$

となる。ここで，$V = u_1^X$，$W = X\,u_1^{X-1}$ とおいて，$Y = a\,V + b\,W + c$ なる重回帰式を求める。

求めるパラメータの第 2 近似は，$\gamma = c$，$\alpha = a$，また，$b = a\,(u_2 - u_1)$ より，β の第 2 近似は $u_2 = u_1 + b\,/\,a$ となる。この繰返し計算を $b\,/\,a$ が十分小さくなるまで繰返す。

なお，データ数が 7 以上の場合には，データを 7 分割し以下のようにして β の第 1 近似を求める[*3]。

$$u_1 = \left(\frac{y_6 + y_5 + y_4 - y_2 - 2\,y_1}{y_5 + y_4 + y_3 - y_1 - 2\,y_0} \right)^{1/m} \tag{2.28}$$

y_i は 7 分割された各区分での従属変数の平均値，m は各区分のデータ数。

[*3] Snedecor, G. W. and Cochran, W. G.: Statistical Method, 6th edition. 1967.
訳書 畑村又好, 奥野忠一, 津村善郎: 統計的方法. 岩波書店, 1972.

2.4　ロジスティック曲線へのあてはめと重回帰分析

人口とか，テレビや携帯電話の普及台数などは，最初は緩やかに増加するがだんだん増加の速度が速くなり，十分大きくなると頭打ちになり増加速度はだんだんと鈍くなる。この様子は図 2.9 のようになるが，これは動物の成長（例えば身長とか体重）の過程によく似ているので，このような曲線を**成長曲線**と呼ぶ。

ロジスティック曲線も成長曲線の一つである。独立変数が等間隔でない場合や，より妥当なあてはめを行う場合には，非線形最小二乗あてはめを行うことになるが，ここでは以下のような条件の下でのあてはめを考える。

1. 独立変数は 1 から始まる整数値を使用する。
2. 従属変数は全て正の値でなければならない（0 も不可）。

データが飽和点に達していない部分のみ（指数的な増加部分だけ）の場合には，あてはめに失敗する場合がある。このような場合には非線形最小二乗法によるあてはめを行う。

ロジスティック曲線（2.29）式の両辺の逆数をとると，（2.30）式のようになる。

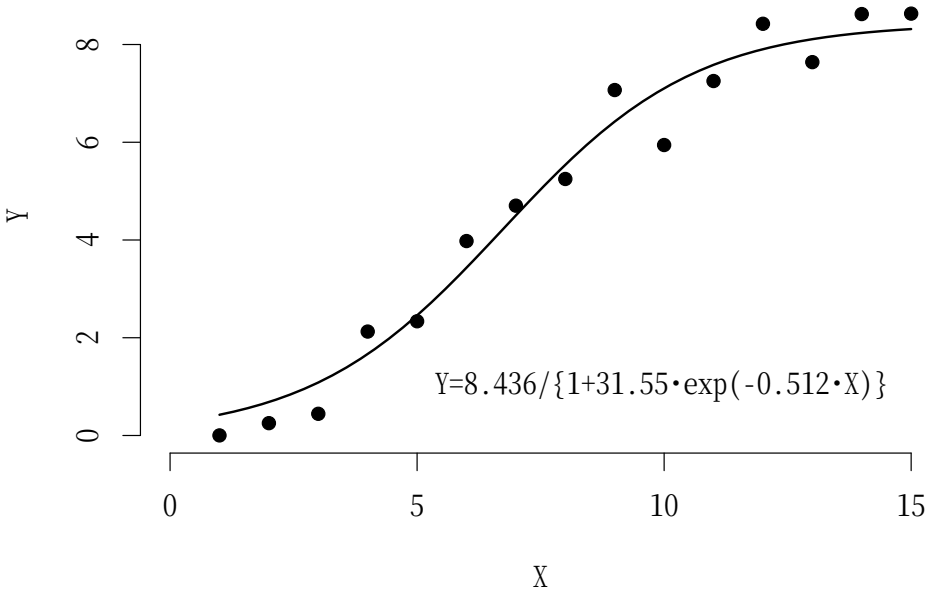

図 2.9　ロジスティック曲線の概形

$$y = \frac{a}{1 + b\ \exp(-c\,x)} \tag{2.29}$$

$$\frac{1}{y} = \frac{1}{a} + \frac{b}{a}\ \exp(-c\,x) \tag{2.30}$$

ここで，$Y = 1\,/\,y$，$A = 1\,/\,a$，$B = b\,/\,a$ とおくと（2.31）式のようになる。

$$Y = A + B\ \exp(-c\,x) \tag{2.31}$$

（2.31）式は，未知のパラメータ A，B については線形であるが，c については非線形である。そこで，以下のような逐次的に近似する手法をとる。

パラメータ c の近似値を c_1 とする（$c = c_1 + \delta$）。

$$\exp(-c\,x) \quad \exp(-c_1\,x) - \delta\,x\ \exp(-c_1\,x) \tag{2.32}$$

（2.31）式に代入して，

$$Y = A + B\left\{\exp(-c_1\,x) - \delta\,x\,\exp(-c_1\,x)\right\} \tag{2.33}$$

$X_1 = \exp(-c_1\,x)$，$X_2 = x\exp(-c_1\,x)$，$C = B\,\delta$ とおくと，

$$Y = A + B\,X_1 - C\,X_2 \tag{2.34}$$

（2.34）式は，2 個の独立変数（X_1，X_2）からなる重回帰式であるので，A，B，C を求めることができる。c の近似値 c_1 の改良値 c_2 は，$\delta = C\,/\,B$ であるから，$c_2 = c_1 + \delta$ と表される。

パラメータ c の修正量 δ が十分小さくなるまで（2.34）式の重回帰式を繰返して計算する。

（2.29）式の c の初期値は何らかの方法で事前に推定されていなければならない。n 個のデータを $m = [\,n/3\,]$ 個ずつに 3 区分し，以下の式で得られたものを初期値とする（[　] はガウス記号）。ただし，この方法では成長限界の正確な推定ができないことが往々にしてある。

$$c_1 = \frac{1}{m}\ \ln\left(\frac{S_1 - S_2}{S_2 - S_3}\right) \tag{2.35}$$

$$S_1 = \sum_{i=1}^{m}\ln(y_i), \qquad S_2 = \sum_{i=1}^{m}\ln(y_{m+i}), \qquad S_3 = \sum_{i=1}^{m}\ln(y_{2m+i}) \tag{2.36}$$

2.5　変数変換などにより重回帰分析に帰結できる回帰分析

漸近指数曲線やロジスティック回帰分析は，非線形回帰分析と呼ばれるもので，簡単な計算では解が求まらない。このため，コンピュータの利用が制限されている場合には，計算量の少ない重回帰分析に帰結させて解を求めようとするものである。実際的には，非線形最小二乗法によるコンピュータプログラム[*4]を用いれば，簡単に解を求めることができる。

本節では，重回帰分析に帰結できるその他のモデルについてまとめておく。以下のモデルの記述において，特に断りのない限り，ギリシア文字は未知母数を表すものとする（同じ記号が同じ数値を表すわけではない）。

2.5.1　累乗モデル

ϵ を平均 1，有限分散の連続分布をする乗法的確率変数として，

$$Y = \beta_0\,X_1^{\beta_1}\,X_2^{\beta_2}\,X_3^{\beta_3}\,\epsilon \tag{2.37}$$

で表されるモデルである。

（2.37）式は両辺の自然対数をとることにより，

$$\ln Y = \ln\beta_0 + \beta_1\,\ln X_1 + \beta_2\,\ln X_2 + \beta_3\,\ln X_3 + \ln\epsilon \tag{2.38}$$

となる。ここで，$\ln Y$，$\ln\beta_0$，$\ln X_1$ などを新たに Y，β_0，X などと書き換えると，（2.3）式で表されるモデルになるので，重回帰分析に帰結できる（ϵ ではなく，$\ln\epsilon$ が平均 0，分散 σ^2 の正規分布に従うことが仮定されることに注意）。

なお，（2.37）式の一番簡単なモデルは，

$$Y = \beta_0\,X^{\beta_1}\,\epsilon \tag{2.39}$$

であるが，これは指数曲線 $Y = aX^b$ を表すものである。

[*4] Excel ででも可能である。「ソルバー」を用いる。

2.5.2　指数モデル

累乗モデルと数式的に似ているが,

$$Y = \beta_0 \beta_1^{X_1} \cdots \beta_p^{X_p} \epsilon \tag{2.40}$$

は全く別のモデルである。

その一番簡単なモデルは,

$$Y = \beta_0 \beta_1^{X_1} \tag{2.41}$$

であるが，これはもう一つの指数曲線 $Y = a\, b^X$ である。

指数モデルは通常は定数 e の累乗として（2.42）式のように表されることが多い。

$$\begin{aligned} Y &= \exp(\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p)\, \epsilon \\ &= e^{\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p}\, \epsilon \\ &= e^{\alpha_0}\, e^{\alpha_1 X_1} \cdots e^{\alpha_p X_p}\, \epsilon \\ &= e^{\alpha_0}\, (e^{\alpha_1})^{X_1} \cdots (e^{\alpha_p})^{X_p}\, \epsilon \\ &= \beta_0 \beta_1^{X_1} \cdots \beta_p^{X_p} \epsilon \end{aligned} \tag{2.42}$$

（2.42）式の両辺の自然対数をとることにより，（2.43）式の形になるので，重回帰分析に帰結できる。

$$\begin{aligned} \ln Y &= \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p + \ln \epsilon \\ &= \log \beta_0 + \log \beta_1 X_1 + \cdots + \log \beta_p X_p + \ln \epsilon \end{aligned} \tag{2.43}$$

2.5.3　逆数モデル

$$Y = \frac{1}{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon} \tag{2.44}$$

両辺の逆数をとることにより，（2.45）式の形になるので，重回帰分析に帰結できる。

$$\frac{1}{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{2.45}$$

2.5.4　多重ロジスティックモデル

多重ロジスティックモデルは，2.4 のロジスティック曲線の一般形である。

$$\begin{aligned} Y &= \frac{1}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon)} \\ &= \frac{1}{1 + \exp(\beta_0)\, \exp(\beta_1 X_1 + \cdots + \beta_p X_p + \epsilon)} \\ &= \frac{1}{1 + \gamma\, \exp(\beta_1 X_1 + \cdots + \beta_p X_p + \epsilon)} \end{aligned} \tag{2.46}$$

両辺の逆数をとり，1 を引いてから自然対数をとることにより，（2.47）式の形になるので，重回帰分析に帰結できる。

$$\ln\left(\frac{1}{Y} - 1\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{2.47}$$

2.6　重回帰分析におけるダミー変数の利用

2.6.1　ダミー変数を用いた季節変動の扱い

景気や商品の需要分析を行うとき，3 か月を単位とする四半期データが用いられることがある。季節ごとの変動を表現するためにダミー変数が使われることがある。

例えばビールの消費量は気温と高い相関を持つが，気温そのもののデータではなく季節を表すダミー変数を考える。これは一見すると情報を 100% 利用していないともいえるが，逆に気温以外の要因（年中行事など生活に密着した様々な要因）を考慮しているともいえよう。

さて，1.4 節で述べたように，4 つの季節を表すためには，ある季節を基準として，残りの 3 つの季節の相対差を表現するために 3 個のダミー変数を用いればよい。例えば，第 I 四半期（春）を基準としたとき，第 II 〜 IV 四半期（夏，秋，冬）を表すダミー変数を d_2，d_3，d_4 とする。また，ビールの需要は時間によっても変化しているので時間の要素も重回帰式に取り入れよう。ビールの需要 Y_t は（2.48）式のように表されるであろう。

$$Y_t = b_0 + b_1 X_t + b_2 d_2 + b_3 d_3 + b_4 d_4 \tag{2.48}$$

各季節ごとの予測式は結局，表 2.3 のように表現される。

表 2.3　各四半期とダミー変数の対応

	d_2	d_3	d_4	予測式
第 I 四半期	0	0	0	$Y_t = b_0 + b_1 X_t$
第 II 四半期	1	0	0	$Y_t = b_0 + b_1 X_t + b_2$
第 III 四半期	0	1	0	$Y_t = b_0 + b_1 X_t + b_3$
第 IV 四半期	0	0	1	$Y_t = b_0 + b_1 X_t + b_4$

予測式から解釈できることは，以下のようになるであろう。

- 時間要因に関する重みは b_i である。すなわち，時間 X_t が 1 増加すると，ビールの需要は b_1 増加する。
- 第 II 四半期は第 I 四半期に比べてビールの需要は b_2 だけ大きい。
- 第 III 四半期は第 I 四半期に比べてビールの需要は b_3 だけ大きい。
- 第 IV 四半期は第 I 四半期に比べてビールの需要は b_4 だけ大きい（たぶん b_4 は負の値であろう）。

2.6.2　数量化 I 類 — ダミー変数を用いた重回帰分析

数量化 I 類とは，**カテゴリーデータ**を説明変数として，連続変数である従属変数を**予測**する手法である。

ダミー変数を用いる重回帰分析と等価な解析手法である。

説明変数 X_i（$i = 1, 2, \ldots, p$）が，それぞれ m_i 個の選択肢を持つ（このような変数を特に**アイテム変数**と呼ぼう）。各選択肢が選ばれたら 1，選ばれなかったら 0 をとるような $\sum m_i$ 個の変数 C_{ij}（$i = 1,\ 2,\ \ldots,\ p$; $j = 1,\ 2,\ \ldots,\ m_i$）を定義する。ここで，各カテゴリーに特定の数値 a_{ij}（$i = 1,\ 2,\ \ldots,\ p$; $j = 1,\ 2,\ \ldots,\ m_i$）を割当て，$\hat{Y} = \sum\sum a_{ij} C_{ij}$ で従属変数 Y を予測しようと考える。

表 2.4 に示した例においてみてみると，例えば 1 番目のケースの予測値として $a_{11} + a_{22} + a_{32}$ を使用するわけである。

各カテゴリーにどのような数値を与えたらよいかは，C_{ij} を独立変数として以下のような重回帰式を求めることに帰着できる。

表 2.4　カテゴリー変数を説明変数として従属変数を予測する

従属変数（連続変数）	説明変数（カテゴリー変数）								
	X_1			X_2				X_3	
Y	C_{11}	C_{12}	C_{13}	C_{21}	C_{22}	C_{23}	C_{24}	C_{31}	C_{32}
31.3	1	0	0	0	1	0	0	0	1
25.1	0	1	0	1	0	0	0	1	0
34.7	1	0	0	0	0	1	0	1	0
29.6	0	0	1	0	0	0	1	0	1
⋮									
カテゴリーに与えられる数値	a_{11}	a_{12}	a_{13}	a_{21}	a_{22}	a_{23}	a_{24}	a_{31}	a_{32}

$$
\begin{aligned}
\hat{Y} &= a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13} \\
&+ a_{21}C_{21} + a_{22}C_{22} + a_{23}C_{23} + a_{24}C_{24} \\
&+ a_{31}C_{31} + a_{32}C_{32}
\end{aligned}
$$

ただし，各説明変数において情報が冗長であるので，2 番目以降の各説明変数から 1 個ずつカテゴリーを消去した重回帰分析を行う（例えば，C_{11} と C_{12} が 0 なら C_{13} が 1 であることはただちにわかる）。

$$
\begin{aligned}
\hat{Y} &= a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13} \\
&+ a_{22}C_{22} + a_{23}C_{23} + a_{24}C_{24} \\
&+ a_{32}C_{32}
\end{aligned}
$$

なお，以上で求めた各カテゴリーに与える数値は，各説明変数ごとに平均値がゼロになるように正規化されて利用される。

2.6.3　ダミー変数を用いた重回帰分析と一元配置分散分析

一元配置分散分析は 3 群以上の平均値の差を検定する手法である。

表 2.5 のようなデータについて検定を行った結果は，表 2.6 のような分散分析表として得られる。

表 2.5　一元配置分散分析を行うデータ

	第 1 群	第 2 群	第 3 群
	5.1	5.5	8.8
	4.6	5.3	7.0
	5.6	5.9	5.7
	4.6	7.3	6.7
	4.4	6.4	5.4
	6.4	6.1	8.6
平均値	5.117	6.083	7.033

表 2.5 のデータにおいて，群を識別するために 2 つのダミー変数を用いて，測定値を予測する重回帰分析を行ってみる。分析に使用するデータは表 2.7 のように表す（2.6.1，2.6.2 を参照）。

重回帰分析の結果は，表 2.8，表 2.9 のようになる。

表 2.6　一元配置分散分析の分散分析表

変動要因	平方和	自由度	平均平方	F 値	P 値
級間	11.0211	2	5.511	5.288	0.018
級内	15.6300	15	1.042		
全体	26.6511	17	1.568		

表 2.7　重回帰分析に使用するために準備するデータ

ダミー変数 1	ダミー変数 2	従属変数
0	0	5.1
0	0	4.6
0	0	5.6
0	0	4.6
0	0	4.4
0	0	6.4
1	0	5.5
1	0	5.3
1	0	5.9
1	0	7.3
1	0	6.4
1	0	6.1
0	1	8.8
0	1	7.0
0	1	5.7
0	1	6.7
0	1	5.4
0	1	8.6

表 2.9 は表 2.6 と数値が全く同じになることがわかる。すなわち，回帰による変動（平方和）は「群の情報を表す 2 つのダミー変数で説明される変動」，残差は「ダミー変数で説明しきれない誤差」という意味になる。

また，この分析例は数量化 I 類の例にもなっている。すなわち，表 2.8 は従属変数とした測定値が，

$$\text{測定値} = 0.967 \times \text{ダミー変数 1} + 1.917 \times \text{ダミー変数 2} + 5.117 \tag{2.49}$$

という重回帰式で予測できることを表しているが，この式において 2 つのダミー変数に（0,0），（1,0），（0,1）という値の組を代入したときの測定値の予測値がそれぞれの群の測定値の平均値になることがわかる。

表 2.8　重回帰分析の結果

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数
ダミー変数 1	0.967	0.589	1.640	0.122	0.374
ダミー変数 2	1.917	0.589	3.252	0.005	0.743
定数項	5.117	0.417	12.278	0.000	

表 2.9　重回帰分析における分散分析表

変動要因	平方和	自由度	平均平方	F 値	P 値
回帰	11.021	2	5.511	5.288	0.018
残差	15.630	15	1.042		
全体	26.651	17			

演習問題

問題 2.1

表 2.10 のような，10 ケース，3 変数のデータおいて，変数 X_1，X_2 を用いて Y を予測する重回帰式を求めなさい。

表 2.10　重回帰分析用のデータ

X_1	X_2	Y
1.2	1.9	0.9
1.6	2.7	1.3
3.5	3.7	2.0
4.0	3.1	1.8
5.6	3.5	2.2
5.7	7.5	3.5
6.7	1.2	1.9
7.5	3.7	2.7
8.5	0.6	2.1
9.7	5.1	3.6

問題 2.2

表 2.11 に示すようなデータがある。3 個の独立変数（アイテム変数：X_1, X_2, X_3）は 3 個のカテゴリーを持つ変数であり，それぞれ 1 から 3 までの整数値でコード化されている。従属変数（Y）は連続変数である。数量化 I 類およびダミー変数を用いる重回帰分析を適用し，両者の結果を比較しなさい。

表 2.11　アイテム変数データ例

ケース	X_1	X_2	X_3	Y
1	1	2	2	6837
2	3	2	2	7397
3	1	2	2	7195
4	1	1	1	6710
5	2	3	2	6670
6	1	3	3	6279
7	2	2	2	6601
8	1	1	1	4929
9	3	2	2	5471
10	1	1	2	6164
11	1	1	1	5095
12	1	1	1	4766
13	2	1	1	6525
14	1	1	1	5087
15	1	3	2	6060

第 3 章

判別分析

3.1　線形判別分析

判別分析の目的は，いくつかの変数に基づいて，各データがどの群に所属するかを判定することである。

単純にするために，データが 2 つの群に分けられており，それぞれ 2 個の変数 x_1，x_2 の値が観察されているとする。

x_1 あるいは x_2 においてデータの分布を描くと，図 3.1 のように 2 群が重なる部分が大きいことがわかる。ここで図に示したような座標軸 f を考えると，各データがこの座標軸上でとる値は，

$$f = a\,x_1 + b\,x_2 \tag{3.1}$$

のように合成変数の形になることがわかる。座標軸 f 上でのデータの分布を描くと，各群の重なる部分が小さくなることがわかる。これは，座標軸 f 上で，ある値より大きい値であるか小さい値であるかによって，そのデータがいずれの群に属するかを判定できることを意味する。

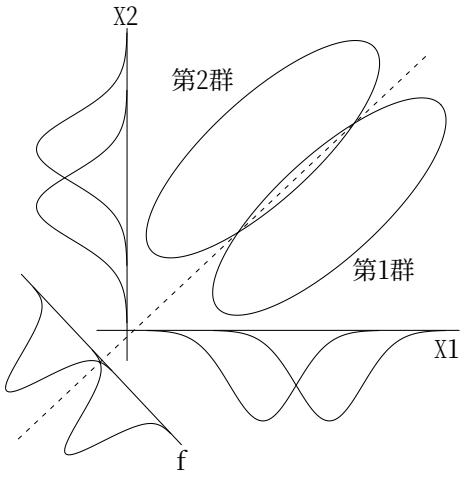

図 3.1　判別分析とは何か

3.1.1　相関比を最大にすることによる判別係数の求め方

群の数を k，各群のケース数を n_1，n_2，…，n_k とする。p 個の変数を X_1，X_2，…，X_p として，任意の重み係数 a_1, a_2, …, a_p を用いて作られる合成変量を Z とする。

$$Z = a_1\,X_1 + a_2\,X_2 + \cdots + a_p\,X_p \tag{3.2}$$

第 j 群，第 i ケースの合成変量を Z_{ij}（$j = 1,\ 2,\ \ldots,\ k$；$i = 1,\ 2,\ \ldots,\ n_j$）とする。

$$Z_{ij} = a_1 X_{1ij} + a_2 X_{2ij} + \cdots + a_p X_{pij} \tag{3.3}$$

全体の平均値を $\bar{Z}$，第 j 群における平均値を $\bar{Z}_j$ とすれば，Z の平方和 S_t は**群内平方和** S_w と**群間平方和** S_b に分解できる。

$$S_t = S_w + S_b \tag{3.4}$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Z_{ij} - \bar{Z})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Z_{ij} - \bar{Z}_j)^2 + \sum_{j=1}^{k} n_j (\bar{Z}_j - \bar{Z})^2 \tag{3.5}$$

Z により各群がよく判別できるということは，相関比 $\eta^2 = S_b / S_t$ が大きいということに対応するので（あるいは，S_b / S_w の比が大きいことに対応すると考えてもよい），相関比が最大になるように重み係数 $a_1,\ a_2,\ \ldots,\ a_p$ を決定すればよい。これを，**判別係数**と呼ぶ。

l 群の m 番目のケースの i 番目の変数の測定値を X_{iml}，l 群の i 番目の変数の平均値を $\bar{X}_{il}$，i 番目の変数の全体の平均値を $\bar{X}_i$ としたとき，観察値の群内および群間平方和・積和行列の要素を，

$$W_{ij} = \sum_{l=1}^{k}\sum_{m=1}^{n_l}(X_{iml} - \bar{X}_{il})(X_{jml} - \bar{X}_{jl}), \qquad \boldsymbol{W} = (W_{ij}) \tag{3.6}$$

$$B_{ij} = \sum_{l=1}^{k} n_l (\bar{X}_{il} - \bar{X}_i)(\bar{X}_{jl} - \bar{X}_j), \qquad \boldsymbol{B} = (B_{ij}) \tag{3.7}$$

とすると，

$$\theta = \frac{S_b}{S_w} = \frac{\displaystyle\sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j B_{ij}}{\displaystyle\sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j W_{ij}} \rightarrow \text{最大化} \tag{3.8}$$

（3.8）式を，a_i で偏微分して 0 とおき行列を用いて表すと，

$$(\boldsymbol{B} - \theta \boldsymbol{W})\,\boldsymbol{a} = 0 \tag{3.9}$$

となる。これが自明の解以外を持つためには係数行列式が（3.10）式の条件を満たさなければならない。

$$|\boldsymbol{B} - \theta \boldsymbol{W}| = 0 \tag{3.10}$$

$\boldsymbol{W}$ の逆行列を $\boldsymbol{W}^{-1}$ とすると，

$$|\boldsymbol{W}^{-1}\boldsymbol{B} - \theta \boldsymbol{I}| = 0 \tag{3.11}$$

となり，θ は行列 $\boldsymbol{W}^{-1}\boldsymbol{B}$ の**固有値**であることがわかる。（3.11）式を満たす固有値は複数個（$m = \min(p,\ k-1)$ 個）存在するが，（3.8）式を最大にするのはその内の最大の固有値である。また，係数ベクトル $\boldsymbol{a}$ はその固有値に対応する**固有ベクトル**である。

2 番目以降に大きい固有値とその固有ベクトルも判別関数を構成することができる。それぞれの判別関数の寄与率は（3.12）式のようになり，m より少ない判別関数で十分な判別を行うことができる。すなわち，**次元の減少を伴う判別**であるといわれる。

$$\text{寄与率} = \lambda_i \Big/ \sum_{j=1}^{m} \lambda_j \tag{3.12}$$

3.1.2 マハラノビスの距離に基づく判別係数の求め方

k 個の群の母集団の平均を $\boldsymbol{\mu}_j = (\mu_{1j}, \mu_{2j}, \ldots, \mu_{pj})'$ （$j = 1, 2, \ldots, k$），観測値を $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ とする。

各群の分散共分散行列を $\boldsymbol{\Sigma}_j$，その逆行列を $\boldsymbol{\Sigma}_j^{-1}$ とするき，（3.13）式による各群までの**マハラノビス距離**を計算し，最も近い群に属すると判定する。

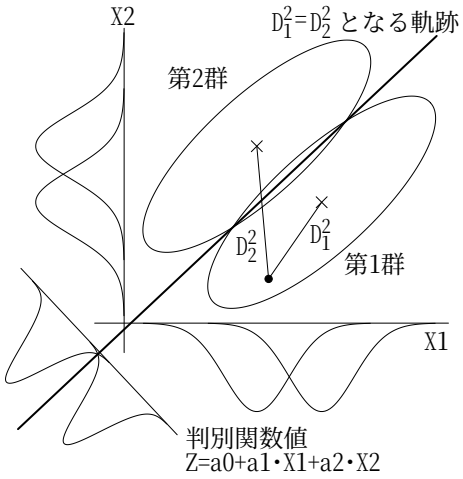

図 3.2 マハラノビス距離による判別

$$d_j^2 = (\boldsymbol{X} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_j) \tag{3.13}$$

もし，各群の分散共分散行列が等しい，すなわち $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ が仮定できれば（3.14）式のようになる。

$$\begin{aligned} d_j^2 &= (\boldsymbol{X} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_j) \\ &= \boldsymbol{X}' \boldsymbol{\Sigma}^{-1} \boldsymbol{X} - 2\, \boldsymbol{X}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j + \boldsymbol{\mu}_j' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j \end{aligned} \tag{3.14}$$

第 1 項は各群に共通，第 3 項は各群ごとに異なる定数（これを c_j とする）である。各ケースごとに異なるのは第 2 項のみであるから，（3.15）式の計算を行えばよい。

$$2\, \boldsymbol{X}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j = a_{1j} X_1 + a_{2j} X_2 + \ldots + a_{pj} X_p \tag{3.15}$$

係数 $a_{1j}, a_{2j}, \ldots, a_{pj}$ は $\boldsymbol{\Sigma}^{-1}$ の要素を σ^{ij} とすれば（3.16）式で求められる。

$$a_{ij} = 2\,(\sigma^{i1} \mu_{1j} + \sigma^{i2} \mu_{2j} + \ldots + \sigma^{ip} \mu_{pj}), \qquad i = 1, 2, \ldots, p \tag{3.16}$$

（3.14）式の第 1 項は群に関係ないため無視できるので，（3.17）式の数値が最も小さい群に属すると判定すればよい。（3.17）式は，**分類関数**と呼ばれる。

$$f_j = a_{1j} X_1 + a_{2j} X_2 + \ldots + a_{pj} X_p + c_j \tag{3.17}$$

また，マハラノビス距離の大小を比較する代わりにあらゆる 2 群の組合わせに対して，（3.18）式で表される ${}_kC_2$ 個の**判別関数**を定義しておくこともできる。例えば，第 1 群と第 2 群の判別関数は，

$$Z_{1\cdot 2} = d_1^2 - d_2^2 = f_1 - f_2$$

$$= \quad (a_{11} - a_{12})\, X_1 + (a_{21} - a_{22})\, X_2 + \cdots + (a_{p1} - a_{p2})\, X_p + (c_1 - c_2) \tag{3.18}$$

（3.18）式の判別関数は群の数が 2 群のときは 1 個の判別関数を計算すればよいので便利であるが，群の数が 3 群以上のときには，やや煩わしい。

判別方法としてはこの他に，（3.14）式の d_j^2 が自由度 p の χ^2 分布に従うことを利用する方法がある。各群の中心からのマハラノビス距離から $P_j = \Pr\{\chi^2 \geqq d_j^2\}$ を求め，P_j の最も大きい群に所属すると判別する。これは，d_j^2 の最も小さい群に所属すると判別することと同じであり，わざわざ各群へ所属する確率を求める必要はなさそうに思うかもしれないが，どの群にも所属しないケースの可能性を考えるとこのような方法をとる必要性がわかるであろう。

なお，各群の分散共分散行列が等しくない場合は，判別境界は，図 3.2 のような直線ではなく，図 3.3 のような二次曲線になるので，（3.13）式による判別は**二次の判別関数**と呼ばれる。

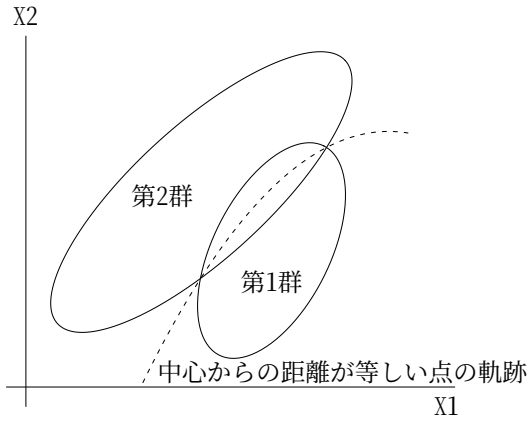

図 3.3　二次の判別関数

2 群の判別の場合，計算は若干簡単になる。各群の変動・共変動行列を $\boldsymbol{S}^{(1)}$，$\boldsymbol{S}^{(2)}$，2 群をプールしたときの分散・共分散行列を $\boldsymbol{V}$ とすると，それぞれの要素は以下のように定義される。

$$\begin{cases} S_{ij}^{(1)} = \displaystyle\sum_{m=1}^{n_1} (X_{im1} - \bar{X}_{i1})\,(X_{jm1} - \bar{X}_{j1}) \\ S_{ij}^{(2)} = \displaystyle\sum_{m=1}^{n_2} (X_{im2} - \bar{X}_{i2})\,(X_{jm2} - \bar{X}_{j2}) \\ V_{ij} = \dfrac{S_{ij}^{(1)} + S_{ij}^{(2)}}{n_1 + n_2 - 2} \end{cases} \tag{3.19}$$

また，各変数の平均値の差のベクトル $\boldsymbol{d}$ を，

$$\boldsymbol{d} = (\bar{X}_{11} - \bar{X}_{12},\, \bar{X}_{21} - \bar{X}_{22},\, \ldots,\, \bar{X}_{p1} - \bar{X}_{p2})' \tag{3.20}$$

とすると，$\boldsymbol{a}$ は以下の連立方程式を解けば得られる。

$$\begin{cases} a_1 V_{11} + a_2 V_{12} + \cdots + a_p V_{1p} = d_1 \\ a_1 V_{21} + a_2 V_{22} + \cdots + a_p V_{2p} = d_2 \\ \qquad \vdots \\ a_1 V_{p1} + a_2 V_{p2} + \cdots + a_p V_{pp} = d_p \end{cases} \tag{3.21}$$

行列で表せば，$\boldsymbol{V}\,\boldsymbol{a} = \boldsymbol{d}$ ゆえ，$\boldsymbol{V}$ の逆行列を $\boldsymbol{V}^{-1}$ とすれば，$\boldsymbol{a} = \boldsymbol{V}^{-1}\,\boldsymbol{d}$ である。

3.1.3 変数選択

多くの説明変数から有用な少数個の説明変数を精選して判別式を作ることが重要になることもある。理論的基盤があるときはそれに従って説明変数を選択することが可能であるが，そうでないときには以下に示すような変数の選択法がある。

変数増加法 最初に最も判別に有効な説明変数（外的基準変数との相関係数が最も大きいもの）を判別式に取り入れる。次の段階では，残りの説明変数の中で最も判別に有効な説明変数を取り入れる。判別精度の改善が一定限度以上である間，この操作を繰り返す。

変数減少法 最初に全ての説明変数を含む判別式を作る。次に，その中から最も判別に有効でない説明変数を除去する。判別精度の低下が一定限度以内である間，この操作を繰り返す。

変数増減法 変数増加法では，いったん判別式に取り込まれた説明変数は除去されることはないが，後の段階になってそれまでに取り込まれた説明変数の重要性が低くなることがある。変数増減法は各段階で変数を追加した後で除去すべき説明変数がないかをチェックする。

変数減増法 最初に全ての説明変数を含む判別式を作る。その後に続く各段階では，まず既に判別式に取り入れられている説明変数の中から最も判別に有効でない説明変数を除き，取り入れられていない説明変数の中に取り入れるべきものがないかをチェックする。

説明変数の追加・除去の基準としては，各変数の**偏 F 値**に基づく F_{in}，F_{out}，それを有意確率に換算した P_{in}，P_{out} がある。いずれも，各説明変数の判別係数の有意性検定と関連しており，後者で例えば $P_{in} = P_{out} = 0.05$ を指定するということは，最終的な判別式に含まれる全ての説明変数の判別係数が 0 であるという帰無仮説が有意水準 5 % で棄却されるということを意味する。

3.2 分析結果の出力例

フィッシャーのアヤメのデータにおいて，3 つの種を 4 つの変数で判別した結果を例示する。

表 3.1 は分類関数についての結果であり，それぞれの群ごとに各説明変数に対する重みが示されている。偏 F 値とそれに対する P 値が，それぞれの説明変数が判別にどのように寄与するかを表している。

表 3.1 分類関数の出力

	第 1 群	第 2 群	第 3 群	偏 F 値	P 値
萼片の長さ	−47.088	−31.396	−24.892	4.721	0.010
萼片の幅	−47.176	−14.145	−7.371	21.936	< 0.001
花弁の長さ	32.861	−10.423	−25.533	35.590	< 0.001
花弁の幅	34.797	−12.868	−42.158	24.904	< 0.001
定数項	170.420	143.510	206.540		

表 3.2 は判別関数に付いての結果であり，あらゆる 2 群の組み合わせごとに，その 2 群を判別するための各説明変数に対する重み（判別係数）が示されている。説明変数の相対的重要性を見るためには，標準化判別係数を見なければならない。

表 3.3 は個々のケースについて，各群の重心までの二乗距離とそのケースが当該群に所属する確率が表示される。2 群の判別の場合には判別値が表示されることもある。

表 3.4 は，実際の群と判別された群についての集計表の例である。

表 3.2　判別関数の出力

	判別係数 1 vs. 2	標準化 判別係数	判別係数 1 vs. 3	標準化 判別係数	判別係数 2 vs. 3	標準化 判別係数
萼片の長さ	7.846	6.475	11.098	9.159	3.252	2.684
萼片の幅	16.515	7.174	19.903	8.646	3.387	1.471
花弁の長さ	−21.642	−38.077	−29.197	−51.37	−7.555	−13.292
花弁の幅	−23.833	−18.105	−38.478	−29.231	−14.645	−11.126
定数項	−13.456		18.06		31.516	

表 3.3　個別の判別結果

ケース	実際の群		判別結果	二乗距離 1	確率 1	二乗距離 2	確率 2	判別値
:	:		:	:	:	:	:	:
17	1		1	2.473	0.480	13.070	0.004	5.299
18	1		1	4.308	0.230	5.174	0.160	0.433
19	1	誤判別	2	2.717	0.437	0.863	0.834	−0.927
20	1		1	3.356	0.340	4.543	0.209	0.594
21	1		1	8.234	0.041	9.182	0.027	0.474
22	1		1	1.350	0.717	12.605	0.006	5.627
23	1		1	1.816	0.611	6.558	0.087	2.371
24	1		1	3.943	0.268	14.862	0.002	5.460
25	1		1	2.490	0.477	9.178	0.027	3.344
26	2		2	14.64	0.002	3.255	0.354	−5.693
:	:		:	:	:	:	:	:

表 3.4　判別状況

	判別された群			
実際の群	群 1	群 2	群 3	合計
群 1	50	0	0	50
群 2	0	48	2	50
群 3	0	1	49	50

正判別率: 98%

3.3　正準判別分析

次元の減少を伴う判別分析である。3.1.1 の「相関比を最大にする判別」参照。

判別する群が 3 つ以上ある場合には，一般に，第 1 正準変量だけでは十分に判別が行えない。このような場合には，第 1 正準変量に直交する（無相関な）合成変量（第 2 正準変量）を作る。このような合成変量（重み係数）は，$r \leqq \min(k-1,\ p)$ 種類存在する。

フィッシャーのアヤメのデータに適用した結果は図 3.4 のようになる。

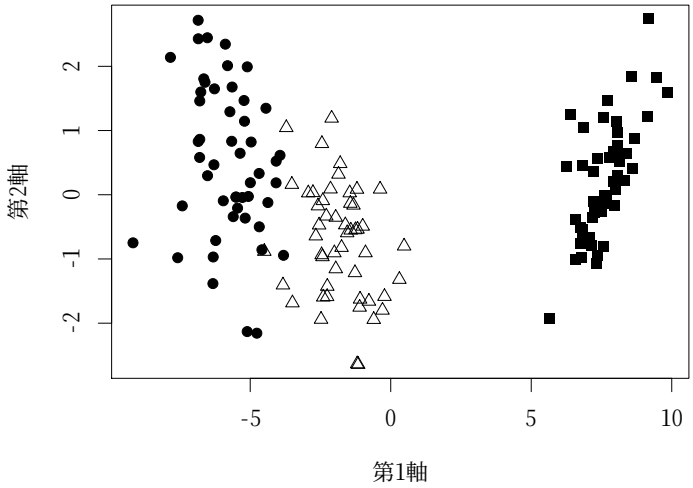

図 3.4　正準判別分析による判別図

3.4　二次の判別関数

群ごとの分散共分散行列が等しいと仮定できない場合に，採用される。3.1.2 の「マハラノビスの距離に基づく判別」参照。

3.5　数量化 II 類 — ダミー変数を用いた判別分析

数量化 II 類とは，**カテゴリーデータ**を説明変数として群を**判別**する手法である。

ダミー変数を用いる判別分析と等価な解析手法である。

説明変数 X_i（$i = 1, 2, \ldots, p$）が，それぞれ m_i 個の選択肢を持つ（このような変数を特に**アイテム変数**と呼ぼう）。各選択肢が選ばれたら 1，選ばれなかったら 0 をとるような $\sum m_i$ 個の変数 C_{ij}（$i = 1, 2, \ldots, p$; $j = 1, 2, \ldots, m_i$）を定義する。ここで，各カテゴリーに特定の数値 a_{ij}（$i = 1, 2, \ldots, p$; $j = 1, 2, \ldots, m_i$）を割当て，$S = \sum\sum a_{ij} C_{ij}$ というサンプルスコア（判別値）を計算して各ケースがどの群に属するかを判別しようと考える。

表 3.5 に示した例においてみてみると，例えば 1 番目のケースの判別をするために $a_{11} + a_{22} + a_{32}$ を使用するわけである。

各カテゴリーにどのような数値を与えたらよいかは，C_{ij} を独立変数として以下のような**判別式**を求めることに帰着できる。

$$\begin{aligned} S &= a_{11} C_{11} + a_{12} C_{12} + a_{13} C_{13} \\ &+ a_{21} C_{21} + a_{22} C_{22} + a_{23} C_{23} + a_{24} C_{24} \\ &+ a_{31} C_{31} + a_{32} C_{32} \end{aligned}$$

ただし，各説明変数において情報が冗長であるので，各説明変数から 1 個ずつカテゴリーを消去した判別分析を行う（例えば，C_{11} と C_{12} が 0 なら C_{13} が 1 であることはただちにわかる）。

$$\begin{aligned} S &= a_{12} C_{12} + a_{13} C_{13} \\ &+ a_{22} C_{22} + a_{23} C_{23} + a_{24} C_{24} \\ &+ a_{32} C_{32} \end{aligned}$$

なお，以上で求めた各カテゴリーに与える数値は，各説明変数ごとに平均値がゼロになるように正規化されて利用される。

表 3.5　カテゴリー変数で群の判別を行う

従属変数（群変数）			説明変数（カテゴリー変数）								
			X_1			X_2				X_3	
Y_1	Y_2	Y_3	C_{11}	C_{12}	C_{13}	C_{21}	C_{22}	C_{23}	C_{24}	C_{31}	C_{32}
1	0	0	1	0	0	0	1	0	0	0	1
0	1	0	0	1	0	1	0	0	0	1	0
1	0	0	1	0	0	0	0	1	0	1	0
0	0	1	0	0	1	0	0	0	1	0	1
⋮											
カテゴリーに与えられる数値			a_{11}	a_{12}	a_{13}	a_{21}	a_{22}	a_{23}	a_{24}	a_{31}	a_{32}

演習問題

問題 3.1

2 群で平均値が全く同じである変数は判別に役立つだろうか。

問題 3.2

2 群のそれぞれが 6 ケースからなる，2 変数のデータがある（表 3.6）。2 群の判別関数を求めなさい。

表 3.6　判別分析用のデータ

	第 1 群		第 2 群	
ケース	X_1	X_2	X_1	X_2
1	5	10	10	8
2	0	7	7	7
3	4	7	9	5
4	8	6	5	3
5	2	5	9	2
6	2	4	5	2

問題 3.3

表 3.7 に示すようなデータがある。3 個の独立変数（アイテム変数： X_1, X_2, X_3）は 3 個のカテゴリーを持つ変数であり，それぞれ 1 から 3 までの整数値でコード化されている。群変数（Y）は 1 または 2 の整数値を持つ。数量化 II 類およびダミー変数を用いる判別分析を適用し，両者の結果を比較しなさい。

表 3.7　アイテム変数データ

ケース	X_1	X_2	X_3	Y
1	1	2	2	2
2	3	2	2	2
3	1	2	2	2
4	1	1	1	2
5	2	3	2	2
6	1	3	3	2
7	2	2	2	2
8	1	1	1	1
9	3	2	2	1
10	1	1	2	2
11	1	1	1	1
12	1	1	1	1
13	2	1	1	2
14	1	1	1	1
15	1	3	2	2

問題 3.4

判別分析は重回帰分析と密接な関連がある。二群の判別の場合，図 3.5 のように従属変数を定義して重回帰分析を行うことにより，判別分析と同じ結果が得られる。適当なデータで確かめてみなさい。

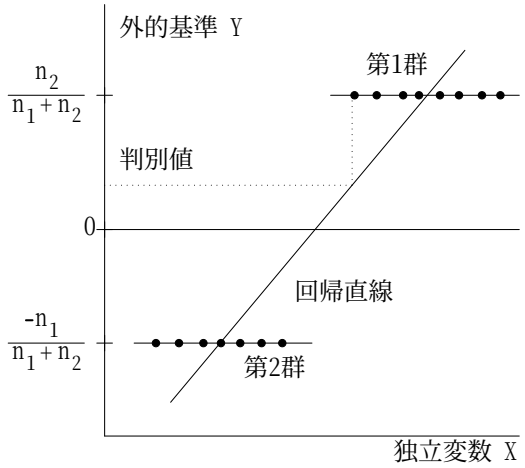

図 3.5　判別分析と重回帰分析の関連

第 4 章

主成分分析

主成分分析は，多変量データの持つ情報を，少数個の総合特性値（合成変数）に要約する手法である（**情報の縮約**）。

変数が持っている**情報量**には，以下のような性質がある。

- 1 個の変数が持つ情報量は，その変数の**分散の大きさ**である。
  100 点満点のテストである能力を測定しようとしたとき，あまりにも易しすぎたり難しすぎたりするテストでは分散が小さいので，各個人の能力段階をとらえるには役立たない。最も有効なテストは平均点が 50 点で，0 点から 100 点まで均等にばらつく[*1]ようなテストである。
- 異なった単位で測定された場合[*2]は，単に分散の大小では比較できない。このような場合は，平均値が 0，分散が 1 になるように標準化するとよい。
- 複数個の変数の合計値の分散は，それぞれの変数の分散の和であるから[*3]，合計値の持つ情報量は大きくなる。

標準化された 2 変数 x_1，x_2 を考えたとき，図 4.1 のように座標軸 x_1，x_2 で表されるものを，座標軸 f_1，f_2 で 表すことを考える。これは座標軸の回転であり，回転角 $\theta = 45^\circ$ としたとき，

$$\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{4.1}$$

すなわち，

$$\begin{cases} f_1 = x_1\cos\theta + x_2\sin\theta = (x_1 + x_2)/\sqrt{2} \\ f_2 = -x_1\sin\theta + x_2\cos\theta = (-x_1 + x_2)/\sqrt{2} \end{cases} \tag{4.2}$$

f_1 は最も分散（情報量）の大きい軸，f_2 は f_1 と直交して，次に分散の大きい軸である。f_1 は x_1，x_2 と最も相関が高い。f_2 は f_1 とは全く別の基準である。つまり，f_1 と f_2 が直交するということは，f_1 と f_2 は無相関であることを意味する。もし，f_1 の分散が f_2 の分散に比べて大きければ，f_1 だけで評価することができる。すなわち，元の 2 変数を「同時に」考慮する代わりに，1 個の合成変数「だけ」を考えればよいことになる。

変数が 3 個以上の場合も同様に考えることができる。元の変数が p 個ある場合も，$m < p$ であるような少数個の合成変数を考えればよい。ここで重要なのは，考慮すべき変数の個数が少なくなることだけではなく，各合成変数間の相関が 0 であることから，「個々の合成変数を独立に評価してよい」ということである[*4]。

[*1] しかし，実際には中心極限定理が働くので正規分布に近い分布が予想される。

[*2] 身長を mm，cm，m で測ったとき，また，身長と体重の情報量の比較はできない。

[*3] 単純な和ではない。

[*4] 逆にいえば，変数間に相関がある限り，各変数を個別に評価することは「できない」ということである。

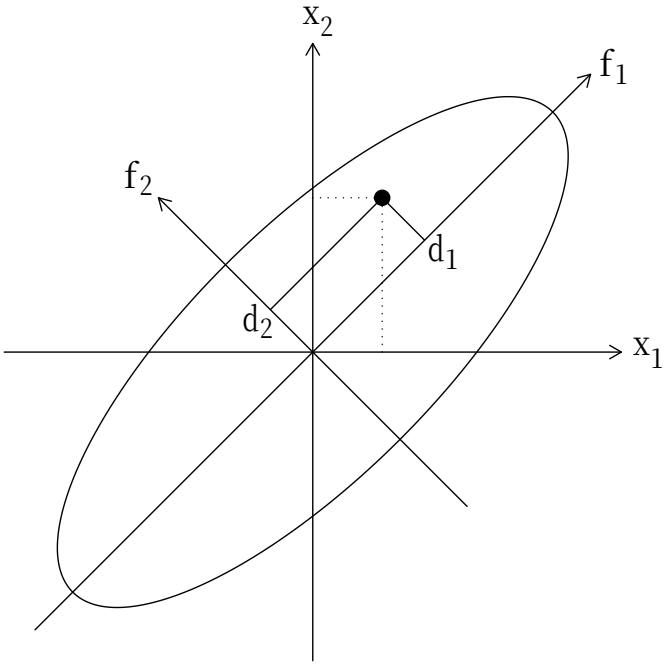

図 4.1　主成分分析とは何か

4.1　主成分の求め方

p 個の変数を $X_1, X_2, \ldots, X_p$，これらの重み付け合成変数を $Z_1, Z_2, \ldots, Z_m$ とする（$m \leqq p$）。

$$\begin{cases} Z_1 = L_{11} X_1 + L_{12} X_2 + \cdots + L_{1p} X_p \\ \quad \vdots \\ Z_i = L_{i1} X_1 + L_{i2} X_2 + \cdots + L_{ip} X_p \\ \quad \vdots \\ Z_m = L_{m1} X_1 + L_{m2} X_2 + \cdots + L_{mp} X_p \\ \quad \text{ただし，} L_{i1}^2 + L_{i2}^2 + \cdots + L_{ip}^2 = 1, \quad (i = 1, 2, \ldots, m) \end{cases} \tag{4.3}$$

このような m 個の合成変数において，以下のような性質を持つものを考える。

- 各合成変数の相関が 0 である。
- 合成変数の分散 $Var(Z_i)$ は，$Var(Z_1) \geqq Var(Z_2) \geqq \cdots \geqq Var(Z_m)$ である。

p 個の変数の相関係数行列の**固有値**を $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_i \geqq \cdots \geqq \lambda_m \geqq \cdots \geqq \lambda_p \geqq 0$ としたとき，λ_i に対応する**固有ベクトル**を重みとした合成変数が Z_i に対応し，Z_i の分散が λ_i に等しくなる（固有ベクトルは互に直交する，すなわち，互に相関が 0 である）。

$Z_1, Z_2, \ldots, Z_m$ は**主成分**と呼ばれ，そのうちで最も分散の大きい Z_1 は第 1 主成分，次に分散の大きい Z_2 は第 2 主成分，以下順に第 m 主成分と呼ばれる。各主成分と，もとの各変数の間の相関係数は**因子負荷量**と呼ばれる。因子負荷量は，第 i 主成分の重み $L_{i1}, L_{i2}, \ldots, L_{ip}$ に，対応する固有値の平方根をかけたものである。すなわち，Z_i と変数 X_1 の相関係数は $a_{1i} = L_{i1}\sqrt{\lambda_i}$，$Z_i$ と変数 X_2 の 相関係数は $a_{2i} = L_{i2}\sqrt{\lambda_i}$ などとなる。

主成分分析の結果を表 4.1 のように表す。

寄与率の欄は，各変数が m 個の主成分でどれくらい説明されるかを表す（$0 \leqq$ 寄与率 $\leqq 1$）。

もう一つの寄与率は，各主成分がもとの情報をどれくらい説明しているかを表すもので，相関係数から出発した主成分分析の場合には，p 個の変数の持つ情報量の合計は p なので，例えば第 1 主成分の寄与率は $\sum_{j=1}^{p} a_{j1}^2 / p$ で

表 4.1　主成分分析の結果の表現

	第 1 主成分	第 2 主成分	⋯	第 m 主成分	寄与率
X_1	a_{11}	a_{12}	⋯	a_{1m}	$\sum a_{1k}^2$
X_2	a_{21}	a_{22}	⋯	a_{2m}	$\sum a_{2k}^2$
⋮	⋮	⋮	⋯	⋮	⋮
X_p	a_{p1}	a_{p2}	⋯	a_{pm}	$\sum a_{pk}^2$
固有値	$\sum a_{j1}^2$	$\sum a_{j2}^2$	⋯	$\sum a_{jm}^2$	
寄与率	$\sum a_{j1}^2 / p$	$\sum a_{j2}^2 / p$	⋯	$\sum a_{jm}^2 / p$	

ある。

実際の分析結果は表 4.2 のようになる。第 4 主成分まで求めているので，各変数の寄与率（最右端の列）は 1 になっている。通常は，固有値が 1 以上の主成分まで求めるのが普通である。第一主成分は花弁の長さと幅，萼片の長さの主成分であり，第二主成分は萼片特にその幅についての主成分であることがわかる。第二主成分までで，全体の 95.8% の情報を集約していることが分かる。

表 4.2　主成分分析の結果出力例

	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分	寄与率
萼片の長さ	0.890	−0.361	0.276	0.038	1.000
萼片の幅	−0.460	−0.883	−0.094	−0.018	1.000
花弁の長さ	0.992	−0.023	−0.054	−0.115	1.000
花弁の幅	0.965	−0.064	−0.243	0.075	1.000
固有値	2.918	0.914	0.147	0.021	
寄与率（%）	73.0	22.9	3.7	0.5	
累積寄与率	73.0	95.8	99.5	100.0	

4.2　主成分軸の回転（直交回転）

主成分の解釈を容易にするために主成分軸の回転を行うことができる。図 4.2 は 6 変数についての主成分分析の結果であるが，第 1 主成分の因子負荷量がいずれも大きく，違いは第 2 主成分の因子負荷量の正負で表現されている。軸を 45 度回転させると図 4.3 のようになりそれぞれの主成分において因子負荷量の大きい変数のみを考えればよいことになる。

回転前および回転後の因子負荷量行列を $\boldsymbol{A}$，$\boldsymbol{B}$（$p \times m$ 行列），回転行列を $\boldsymbol{T}$（$m \times m$ 行列）とすると，両者の関係は，

$$\boldsymbol{B} = \boldsymbol{A}\,\boldsymbol{T} \tag{4.4}$$

である。回転後の因子負荷量行列 $\boldsymbol{B}$ において，共通性

$$h_j^2 = \sum_{k=1}^{m} b_{jk}^2 \tag{4.5}$$

で因子負荷量を基準化したものを

$$q_{ij} = b_{ij} / h_j \tag{4.6}$$

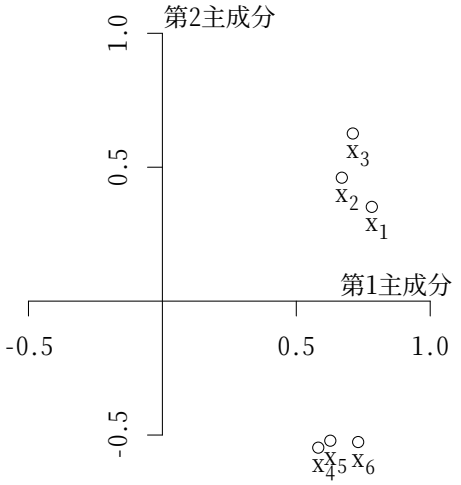

図 4.2　回転前の因子負荷量

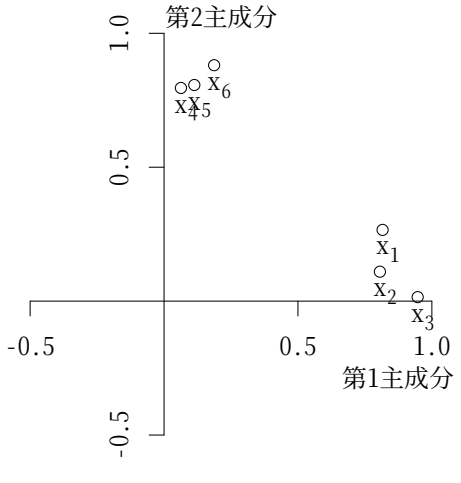

図 4.3　回転後の因子負荷量

とする。全要素について，**オーソマックス基準**

$$c = \sum_{k=1}^{m}\sum_{j=1}^{p} q_{jk}^4 - \frac{w}{p}\sum_{k=1}^{m}\left(\sum_{j=1}^{p} q_{jk}^2\right)^2 \tag{4.7}$$

を最大にすることができる。w のとる値によって特性の異なる因子負荷量行列が得られる。

$$w = \begin{cases} 1, & \text{バリマックス回転} \\ 0.5, & \text{バイコーティマックス回転} \\ 0, & \text{コーティマックス回転} \\ m/2, & \text{エクィマックス回転} \end{cases} \tag{4.8}$$

c を最大化するような回転行列 $\boldsymbol{T}$ は反復的に求める。すなわち，m 個の主成分から 2 個（k, k'）を選び，その平面内で c を最大化する回転行列 $\boldsymbol{T}_{kk'}$ を求める。

主成分 k と k' の平面内で角度 θ だけ回転させるときの回転行列は，対角成分が 1，他の成分が 0 の $m \times m$ 単位正方行列の 4 つの要素を（4.9）式としたものである。

$$\begin{cases} t_{kk} = \cos\theta \\ t_{kk'} = -\sin\theta \\ t_{k'k} = \sin\theta \\ t_{k'k'} = \cos\theta \end{cases} \tag{4.9}$$

ただし，θ は（4.10）式を満たすものとする。

$$\begin{cases} \tan 4\theta = \dfrac{D - 2AB/p}{C - (A^2 - B^2)/p} \\ (D - 2AB/p)\sin 4\theta > 0 \end{cases} \tag{4.10}$$

ここで，$r_j = a_{jk} / h_j$，$s_j = a_{jk'} / h_j$ として，

$$\begin{cases} A = \sum_{j=1}^{p} (r_j^2 - s_j^2) \\ B = 2 \sum_{j=1}^{p} r_j \, s_j \\ C = \sum_{j=1}^{p} (r_j^4 + s_j^4 - 6 \, r_j^2 \, s_j^2) \\ D = 4 \sum_{j=1}^{p} r_j \, s_j \, (r_j^2 - s_j^2) \end{cases} \tag{4.11}$$

m 個の主成分から 2 個を選ぶ $m(m-1)/2$ 通りについて同様に回転を行う。このようなサイクルを繰り返すと c は一定の最大値に近づき収束する。施された回転は，各回転行列の積である。

4.3　主成分得点係数の求め方

主成分得点係数行列 $\boldsymbol{W}$ は，回転前または回転後の因子負荷量行列を $\boldsymbol{A}$ としたとき，（4.12）式である。

$$\boldsymbol{W} = \boldsymbol{A} \, (\boldsymbol{A}' \, \boldsymbol{A})^{-1} \tag{4.12}$$

なお，回転を行わないときは $\boldsymbol{A}' \, \boldsymbol{A}$ は各主成分の固有値を要素とする対角行列であり，回転を行う場合は $\boldsymbol{A}' \, \boldsymbol{A}$ の逆行列を求める。この際，逆行列が求まらない場合には，主成分得点も求めることができない。

4.4　主成分得点の求め方

全データが変数ごとに平均値 0，分散 1 になるように標準化された行列を $\boldsymbol{Z}$ とすると，**主成分得点行列** $\hat{\boldsymbol{F}}$ は（4.13）式である。

$$\hat{\boldsymbol{F}} = \boldsymbol{Z} \, \boldsymbol{W} \tag{4.13}$$

フィッシャーのアヤメのデータを分析した結果における，主成分得点の分布は図 4.4 のようになる。第 1 主成分が特徴の違いのほとんどを表現していることがわかる。

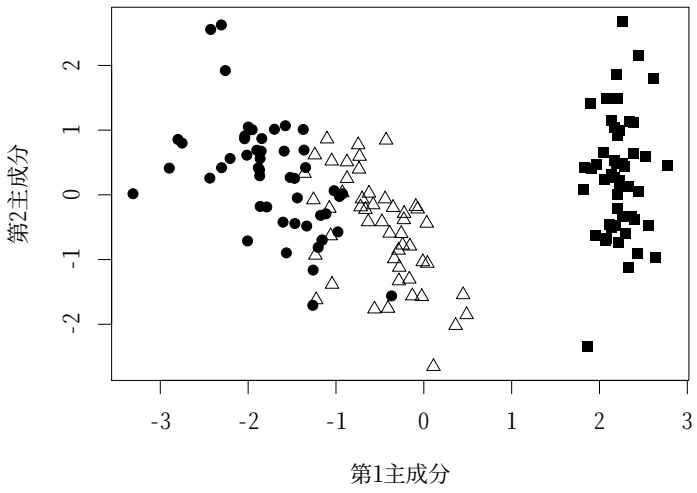

図 4.4　主成分得点の分布

4.5　合成変数

多変量解析では，複数個の変数の重み付き合計値を用いる。解析の目的により各種の重み付けを行う。p 個の変数 x_1，x_2，$\cdots$，x_p が，重み w_1，w_2，$\cdots$，w_p で重み付けされた f を，合成変数と呼ぶ。

$$f = w_1\, x_1 + w_2\, x_2 + \cdots + w_p\, x_p \tag{4.14}$$

ここで，x_i はそれぞれ標準化されており，それぞれの変数が n ケースについて測定されているとき，$n \times p$ の大きさのデータ行列を $\boldsymbol{X}$ で表すことにする。

$$\underset{n\times p}{\boldsymbol{X}} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & x_{ij} & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{4.15}$$

変数 x_j の平均値 $E(x_j)$ と分散 $V(x_j)$ はそれぞれ定義により，

$$E(x_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0 \tag{4.16}$$

$$V(x_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1 \tag{4.17}$$

となる。

また，変数 x_j と x_k の共分散 Cov_{jk} と相関係数 r_{jk} は，

$$Cov_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\, x_{ik} \tag{4.18}$$

$$r_{jk} = \frac{Cov(x_j,\ x_k)}{\sqrt{V(x_j)\ V(x_k)}} = Cov(x_j,\ x_k) \tag{4.19}$$

となる。相関係数行列 $\boldsymbol{R}$ は，データ行列 $\boldsymbol{X}$ を用いて，

$$\underset{p\times p}{\boldsymbol{R}} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & r_{ij} & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix} = \frac{1}{n}\boldsymbol{X'X} \tag{4.20}$$

のように表される（$\boldsymbol{X'}$ は転置行列を表す）。

重みベクトルを $\boldsymbol{w'} = (w_1,\ w_2,\ \cdots,\ w_p)$ として，

$$\boldsymbol{f} = \boldsymbol{Xw} \tag{4.21}$$

で表すことにする。

合成変数 f の平均値と分散は，

$$\begin{aligned} E(f) &= \frac{1}{n}\sum_{i=1}^{n} f_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} w_j\, x_{ij}\right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{p} w_j \left(\frac{1}{n} \sum_{i=1}^{n} x_{ij} \right) \\
&= 0
\end{aligned} \tag{4.22}
$$

$$
\begin{aligned}
V(f) &= \frac{1}{n} \sum_{i=1}^{n} f_i^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} w_j \, x_{ij} \right)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} w_j \, x_{ij} \right) \left(\sum_{k=1}^{p} w_k \, x_{ik} \right) \\
&= \sum_{j=1}^{p} \sum_{k=1}^{p} w_j \, w_k \left(\frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik} \right) \\
&= \sum_{j=1}^{p} \sum_{k=1}^{p} w_j \, w_k \, r_{jk} \\
&= \boldsymbol{w}' \boldsymbol{R} \boldsymbol{w}
\end{aligned} \tag{4.23}
$$

となる。

標準化された合成変数ベクトル $\boldsymbol{g}$ は,

$$
\boldsymbol{g} = \frac{\boldsymbol{f}}{\sqrt{V(f)}} = \frac{\boldsymbol{X}\boldsymbol{w}}{\sqrt{\boldsymbol{w}' \boldsymbol{R} \boldsymbol{w}}} \tag{4.24}
$$

となり，合成変数 g と，もとの変数 x_j との相関係数は,

$$
a_j = r(g, \, x_j) = r(f, \, x_j) = \frac{1}{n} \sum_{i=1}^{n} g_i \, x_{ij} \tag{4.25}
$$

となるので，全ての変数との相関係数ベクトルは,

$$
\begin{aligned}
\boldsymbol{a} &= \frac{1}{n} \boldsymbol{X}' \boldsymbol{g} \\
&= \frac{1}{n} \boldsymbol{X}' \frac{\boldsymbol{X}\boldsymbol{w}}{\sqrt{\boldsymbol{w}' \boldsymbol{R} \boldsymbol{w}}} \\
&= \frac{\boldsymbol{R}\boldsymbol{w}}{\sqrt{\boldsymbol{w}' \boldsymbol{R} \boldsymbol{w}}}
\end{aligned} \tag{4.26}
$$

ここで，データ行列から標準化された合成変数を直接求めることのできる重みベクトル（**標準重みベクトル**）を $\boldsymbol{w}_s$ とする。すなわち,

$$
\boldsymbol{g} = \boldsymbol{X}\boldsymbol{w}_s \tag{4.27}
$$

とすると，(4.24) 式から，$\boldsymbol{w}$ が既知であるときは,

$$
\boldsymbol{w}_s = \frac{\boldsymbol{w}}{\sqrt{\boldsymbol{w}' \boldsymbol{R} \boldsymbol{w}}} \tag{4.28}
$$

となる。

また，$\boldsymbol{a}$ が既知のときは，(4.27) 式の両辺に左から $\frac{1}{n}\boldsymbol{X}'$ をかけて,

$$
\boldsymbol{g} = \boldsymbol{X}\boldsymbol{w}_s
$$

$$\begin{aligned} \frac{1}{n}\boldsymbol{X}'\boldsymbol{g} &= \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{w}_s \\ \boldsymbol{a} &= \boldsymbol{R}\boldsymbol{w}_s \end{aligned} \tag{4.29}$$

さらに，両辺に $\boldsymbol{R}$ の逆行列 $\boldsymbol{R}^{-1}$ を 左からかけることにより，

$$\boldsymbol{R}^{-1}\boldsymbol{a} = \boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{w}_s = \boldsymbol{w}_s \tag{4.30}$$

が得られる。

相関係数行列 $\boldsymbol{R}$ の固有値を λ，固有ベクトルを $\boldsymbol{u}$ とすると，（4.31）式に示すような性質がある。

$$\boldsymbol{R}\boldsymbol{u} = \lambda\boldsymbol{u}, \qquad \lambda = \boldsymbol{u}'\boldsymbol{R}\boldsymbol{u}, \qquad \boldsymbol{u}'\,\boldsymbol{u} = 1 \tag{4.31}$$

これらと前述の関連式を組み合わすと，主成分分析の場合には以下のような関連があることを示せる。

$$V(f) \equiv \lambda \qquad \text{固有値} \tag{4.32}$$

$$\boldsymbol{w} \equiv \boldsymbol{u} \qquad \text{固有ベクトル} \tag{4.33}$$

$$\boldsymbol{a} = \frac{\boldsymbol{R}\boldsymbol{w}}{\sqrt{\boldsymbol{w}'\boldsymbol{R}\boldsymbol{w}}} = \frac{\lambda\boldsymbol{w}}{\sqrt{\lambda}} = \sqrt{\lambda}\boldsymbol{w} \qquad \text{因子負荷量} \tag{4.34}$$

$$\boldsymbol{w}_s = \frac{\boldsymbol{w}}{\sqrt{\lambda}} \qquad \text{主成分得点係数} \tag{4.35}$$

演習問題

問題 4.1

表 4.3 のような，10 ケース，3 変数データ行列において，変数 X_1，X_2，X_3 に対する重みを 0.5，−0.6，1.1 としたときに得られる合成変数 $\boldsymbol{f}$ において，（4.14）式〜（4.30）式を確認しなさい。

表 4.3　標準化されていない 3 変数データ行列

No.	X_1	X_2	X_3
1	9	10	10
2	1	2	5
3	8	3	6
4	6	8	1
5	9	4	4
6	7	10	1
7	10	4	6
8	7	7	3
9	3	10	4
10	4	1	4
$E(X_i)$	6.4	5.9	4.4
$V(X_i)$	7.64	11.09	6.24

第 5 章

因子分析

因子分析は，多変量データから潜在的ないくつかの**共通因子**を推定する手法である。

5.1　因子の求め方

例えば，「知能テスト」が把握する「知能」は，「空間把握能力」，「計算能力」，「言語解釈能力」などに分けられそうである。p 種類の「知能テスト」がどのような知能を測定できるかを例示すると 表 5.1 および図 5.1 のようになるであろう。

表 5.1　多因子モデル

知能テスト	空間認識	計算	言語解釈	
A	○	◎	△	◎ 非常によく把握できる
B	×	△	◎	○ よく把握できる
C	◎	◎	×	△ あまり把握できない
⋮				× ほとんど把握できない

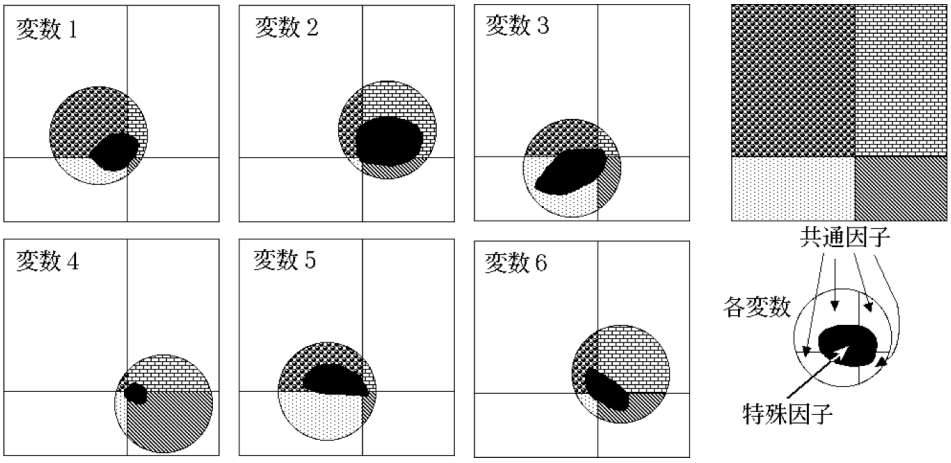

図 5.1　因子分析における共通因子・特殊因子の概念図

p 種類の知能テストが測定する m 種類の能力を F_1, F_2, …, F_m, 知能テストの得点を X_1, X_2, …, X_p としたと

き，これらの得点は以下のように表せるであろう。

$$\begin{cases} X_1 = a_{11} F_1 + a_{12} F_2 + \cdots + a_{1m} F_m + E_1 \\ \quad \vdots \\ X_i = a_{i1} F_1 + a_{i2} F_2 + \cdots + a_{im} F_m + E_i \\ \quad \vdots \\ X_p = a_{p1} F_1 + a_{p2} F_2 + \cdots + a_{pm} F_m + E_p \end{cases} \tag{5.1}$$

$F_1, F_2, \ldots, F_m$ は各知能テストが共通して把握できるある特性であり，**共通因子**と呼ばれるものである。各特性が得点にどの程度反映されるかを表すのが a_{ij}（$i = 1, 2, \ldots, p$; $j = 1, 2, \ldots, m$）であり，共通因子と各知能テストの得点の間の相関係数に相当し**因子負荷量**と呼ばれる。$E_1, E_2, \ldots, E_p$ は**特殊因子**（**独自因子**）と呼ばれ，各知能テストによってのみ把握される特性である。

因子分析の結果を表 5.2 のように表す。

表 5.2　因子分析結果の表

	第 1 因子	第 2 因子	$\cdots$	第 m 因子	共通性
X_1	a_{11}	a_{12}	$\cdots$	a_{1m}	$\sum a_{1k}^2$
X_2	a_{21}	a_{22}	$\cdots$	a_{2m}	$\sum a_{2k}^2$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
X_p	a_{p1}	a_{p2}	$\cdots$	a_{pm}	$\sum a_{pk}^2$
因子負荷量 2 乗和	$\sum a_{j1}^2$	$\sum a_{j2}^2$	$\cdots$	$\sum a_{jm}^2$	
寄与率	$\sum a_{j1}^2 / p$	$\sum a_{j2}^2 / p$	$\cdots$	$\sum a_{jm}^2 / p$	

共通性の欄は，各変数が m 個の共通因子でどれくらい説明されるかを表す（$0 \leqq$ 共通性 $\leqq 1$）。共通性の推定は，**重相関係数の 2 乗値（SMC）**などを初期値として**主因子解**を求め，得られた因子負荷量から改善された共通性の推定値を求めるという手順を，収束するまで繰返す。

各因子がもとの情報をどれくらい説明しているかを表す因子の**寄与率**は，各因子の因子負荷量の 2 乗和を変数の数で割ったものである。例えば，第 1 因子の寄与率は $\displaystyle\sum_{j=1}^{p} a_{j1}^2 / p$ である。

自分の性格についてのアンケート調査を因子分析した結果は表 5.3 のように表される。特に第 1 因子において，全ての変数の因子負荷量が高めになっており，因子の特徴がわかりにくくなっている。

表 5.3　性格の因子分析（因子の回転をしない結果）

	第 1 因子	第 2 因子	共通性
外向性	0.523	0.503	0.526
社交性	0.456	0.488	0.446
積極性	0.487	0.686	0.708
知性	0.672	−0.189	0.487
信頼性	0.612	−0.237	0.431
素直さ	0.813	−0.381	0.806
因子負荷量の二乗和	2.206	1.199	
寄与率（%）	36.8	20.0	
累積寄与率	36.8	56.7	

5.2　因子軸の直交回転

因子の解釈を容易にするために**因子軸の回転**を行うことができる。**直交回転**では，因子軸は直交する（因子間の相関は 0 である）。回転前および回転後の因子負荷量行列を $\boldsymbol{A}$，$\boldsymbol{B}$（$p \times m$ 行列），回転行列を $\boldsymbol{T}$（$m \times m$ 行列）とすると，両者の関係は，

$$\boldsymbol{B} = \boldsymbol{A}\,\boldsymbol{T} \tag{5.2}$$

である。回転後の因子負荷量行列 $\boldsymbol{B}$ において，共通性

$$h_j^2 = \sum_{k=1}^{m} b_{jk}^2 \tag{5.3}$$

で因子負荷量を基準化したものを

$$q_{ij} = b_{ij} \,/\, h_j \tag{5.4}$$

とする。全要素について，**オーソマックス基準**

$$c = \sum_{k=1}^{m}\sum_{j=1}^{p} q_{jk}^4 - \frac{w}{p}\sum_{k=1}^{m}\left(\sum_{j=1}^{p} q_{jk}^2\right)^2 \tag{5.5}$$

を最大にすることができる。w のとる値によって特性の異なる因子負荷量行列が得られる。

$$w = \begin{cases} 1, & \text{バリマックス回転} \\ 0.5, & \text{バイコーティマックス回転} \\ 0, & \text{コーティマックス回転} \\ m/2, & \text{エクィマックス回転} \end{cases} \tag{5.6}$$

c を最大化するような回転行列 $\boldsymbol{T}$ は主成分分析の場合と同様にして反復的に求める（4.2 項参照）。

前出の性格の因子分析において，因子軸をバリマックス回転させた結果は表 5.4 のようになる。表 5.3 と比べると第 1，第 2 因子共に，因子負荷量の大きさにメリハリがきいて解釈しやすくなっていることがわかる。すなわち，第 1 因子は，知性，信頼性，素直さの因子であり，第 2 因子は外向性，社交性，積極性の因子であると解釈できる。

また，因子軸の回転を行っても因子負荷量の二乗和の合計と，共通性の値は回転前の結果と変わりがないこともわかる。

表 5.4　性格の因子分析（バリマックス回転後の結果）

	第 1 因子	第 2 因子	共通性
外向性	0.184	0.702	0.526
社交性	0.135	0.654	0.446
積極性	0.058	0.840	0.708
知性	0.672	0.189	0.487
信頼性	0.646	0.117	0.431
素直さ	0.892	0.099	0.806
因子負荷量の二乗和	1.720	1.684	
寄与率（%）	28.7	28.1	
累積寄与率	28.7	56.7	

5.3　因子軸の斜交回転

斜交回転では因子軸は直交せず，因子間に相関がある。

最大化すべき基準としては，(5.7) 式の**オブリミン基準**

$$c = \sum_{k=1}^{m}\sum_{l=1}^{m}\left[\sum_{j=1}^{p} q_{jk}\, q_{jl} - \frac{w}{p}\left(\sum_{j=1}^{p} q_{jk}^2\right)\left(\sum_{j=1}^{p} q_{jl}^2\right)\right] \tag{5.7}$$

が使われる。w のとる値によって特性の異なる因子負荷量行列が得られる。

$$w = \begin{cases} 1, & \text{コバリミン回転} \\ 0.5, & \text{バイコーティミン回転} \\ 0, & \text{コーティミン回転} \end{cases} \tag{5.8}$$

c を最大化するような回転行列 $\boldsymbol{T}$ は直交回転のときと同じようにして反復的に求める（4.2 参照）。

従来は，主因子法により因子負荷量を求めて，その結果を直交回転であるバリマックス回転することが普通に行われていた。しかし，因子軸が直交することを仮定する必然性はなく，斜交解は直交解を含むので[*1]，最近では，最尤法により因子負荷量をもとめ，その後斜交回転であるプロマックス回転を行うことが多い。結果は表 5.5 のように表示される。因子負荷量の二乗和や寄与率・累積寄与率は意味を持たなくなるので記述する必要がない。

表 5.5　性格の因子分析（プロマックス回転後の結果）

	第 1 因子	第 2 因子	共通性
外向性	0.056	0.703	0.526
社交性	0.014	0.662	0.446
積極性	−0.104	0.874	0.708
知性	0.673	0.061	0.487
信頼性	0.660	−0.010	0.431
素直さ	0.924	−0.079	0.806

5.4　因子得点係数の求め方

因子得点係数行列 $\boldsymbol{W}$ は，相関係数行列を $\boldsymbol{R}$，回転前または回転後の因子負荷量行列を $\boldsymbol{A}$ としたとき，(5.9) 式で表される。

$$\boldsymbol{W} = \boldsymbol{R}^{-1}\,\boldsymbol{A} \tag{5.9}$$

この際，相関係数行列の逆行列が求まらない場合には，因子得点も求めることができない。

5.5　因子得点の求め方

全データが変数ごとに平均値 0，分散 1 になるように標準化された行列を $\boldsymbol{Z}$ とすると，因子得点行列 $\hat{\boldsymbol{F}}$ は (5.10) 式で表される。

$$\hat{\boldsymbol{F}} = \boldsymbol{Z}\,\boldsymbol{W} \tag{5.10}$$

[*1] 斜交解を求めて因子軸が直交に近ければ直交解でも良かったんだと言うことが確認できるだけのこと。

5.6 相関係数行列の吟味

5.6.1 反イメージ相関係数行列

因子分析が有効であるためには変数間の相関係数（の絶対値）がある程度大きくなければならない。小さな相関係数が多い場合には共通因子があるとは仮定しにくいからである。変数間の相関関係の強さは偏相関係数で判定することができる。変数が共通因子を持つ場合には偏相関係数は小さくなるはずである。偏相関係数の符号を逆転したものは**反イメージ相関係数**と呼ばれ，この値がゼロに近いときは因子分析が有効であることを示すが，そうでない場合には得られたデータに対して因子分析を適用するのは不適切であることを意味する。

5.6.2 Kaiser–Meyer–Olkin のサンプリング適切性基準

（5.11）式で計算される，観察された相関係数と偏相関係数の比は，因子分析を用いることの適切性を判定するもので，**サンプリング適切性基準**と呼ばれる。

もし，全ての変数間の偏相関係数の二乗和が相関係数の二乗和に比べて小さいときは KMO の値は 1 に近くなる。KMO の値が小さいということは，2 変数間の相関関係を他の変数によって説明することができにくいということを意味するので，因子分析を適用することが不適切であることを示す。

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 \quad - \quad \sum\sum_{i \neq j} a_{ij}^2} \tag{5.11}$$

Kaiser は表 5.6 のような判定基準を提案している[*2]。

表 5.6 サンプリング適切性基準

KMO	判定	
0.9 以上	marvelous	素晴らしい
0.8 以上	meritorious	価値がある
0.7 以上	middling	まずまず
0.6 以上	mediocre	並み
0.5 以上	miserable	惨め
0.5 未満	unacceptable	ふさわしくない

また，個々の変数についても（5.12）式によってサンプリング適切性 MSA_i が評価できる。因子分析が有効であるためには個々の変数に対する MSA_i が十分に大きいことが必要である。MSA_i の小さい変数は因子分析から除去する必要もあろう。

$$MSA_i = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 \quad - \quad \sum_{i \neq j} a_{ij}^2} \tag{5.12}$$

[*2] Kaiser, H. F.: An index of factorial simplicity. *Psychometrika*, **39**, 31-36, 1974.

第 6 章

その他の多変量解析手法

6.1 数量化 III 類

数量化 III 類は質的変数に基づき，ケースおよび変数の似通ったものをまとめる手法である。
分析に使用する変数が間隔尺度以上の場合の主成分分析に相当する。

6.1.1 アイテムデータとカテゴリーデータ

分析に用いるデータには 2 種類ある。

アイテムデータ： 変数 I_i（$i = 1, 2, \ldots, p$）が，それぞれ m_i 個の選択肢を持つ。各ケースは，表 6.1 に示すように，変数 I_i の値として 1, 2, ..., m_i の値を持つ。このデータを，1 個のアイテム変数 I_i を m_i 個のカテゴリー変数（C_{ij}; i=1, 2, ..., p ; j=1, 2, ..., m_i）に対応させ，アイテム変数のとる値が j のとき $C_{ij} = 1$ で，それ以外のとき 0 をとるように，$\sum m_i$ 個のカテゴリーデータに展開する（表 6.2）。各ケースあたりの「反応あり」のカテゴリー数は常に p 個である。

表 6.1 アイテムデータの例

ケース番号	I_1	I_2	I_3
1	1	2	2
2	2	1	1
3	1	3	1
4	3	4	2
5	3	1	2

カテゴリーデータ： 変数 C_i（$i = 1, 2, \ldots, p$）が，それぞれ「反応あり」の場合に 1,「反応なし」の場合に 0 の 値をとる。表 6.3 に例を示したように，各ケースあたりの「反応あり」のカテゴリー数は一定ではない。

アイテムデータはカテゴリーデータに変換できるが，両者は微妙な違いがある。食べ物の好き嫌いを分析することを考えてみよう。調査票の作り方には 2 種類ある。

食べ物を列挙しておき，好きなものに○を付けさせる調査によって得られるのは，カテゴリーデータである。すなわち，n 個のカテゴリーそれぞれに○がついていたかいなかったかであり，各被験者ごとに○の個数は異なる。

もう一つの方法は，食べ物について，「好き」,「嫌い」の二つの選択肢のいずれかに○を付けさせる方法である。この調査結果をデータファイル化する際には，アイテムデータとして入力することも，カテゴリーデータとして入力することも可能である。アイテムデータとして入力する場合は，この調査結果を，好きを 1，嫌いを 2 として入力する。n 種類の食物に対して選択肢が 2 個ずつあるので，計 $2n$ 個のカテゴリーデータに展開される。各被験者あたり○は n 個あることになる。カテゴリーデータとして入力する場合は，前述の場合にならって，n 個のカテゴ

表 6.2　アイテムデータのカテゴリーデータへの展開

	I_1			I_2				I_3	
ケース番号	C_{11}	C_{12}	C_{13}	C_{21}	C_{22}	C_{23}	C_{24}	C_{31}	C_{32}
1	1	0	0	0	1	0	0	0	1
2	0	1	0	1	0	0	0	1	0
3	1	0	0	0	0	1	0	1	0
4	0	0	1	0	0	0	1	0	1
5	0	0	1	1	0	0	0	0	1

リーデータとして，好きと答えた場合に 1，嫌いと答えた場合には 0 として入力する。各被験者あたり○の数は異なることになる。

分析結果（とその解釈）も両者では異なったものになる。いずれの取り扱い方がよいかは一概にはいえないが，カテゴリーデータとして扱った場合に，嫌いという選択の情報を失うことになることや，もし後者の調査法で各項目の選択肢が 3 個以上であった場合との関連からいえば，アイテムデータとして取り扱った方がよいように思われる。両方の分析を行って比較してみるのが一番よい方法かもしれない。

6.1.2　考え方

アイテムデータも，分析に利用される時点ではカテゴリーデータに展開されるので，以下では表 6.3 の例を用いて説明する。

表 6.3　カテゴリーデータの例

ケース番号	C_1	C_2	C_3	C_4	C_5
1	1	0	1	1	0
2	0	1	1	0	1
3	1	0	0	1	0
4	0	1	1	0	0
5	1	0	1	1	0

数量化 III 類では，表 6.3 の行と列を入れ替え，互いに似ているケースとカテゴリーが隣り合わせになるように配置しなおすことを目標とする。例えば，表 6.4 はそのようなものの 1 例である。これを見ると，対角線上に 1 が集まっていることがわかる。

表 6.4 は，各ケースとカテゴリーを等間隔に配置する場合であるが，さらに，各ケースとカテゴリーにケースとカテゴリーは等間隔に配置される必要性はない。そこで，ケースに割当てられる数値を Y_i（$i = 1, 2, \ldots, n$），カテゴリーに割当てられる数値を X_j（$j = 1, 2, \ldots, p$）とする。全ケースの反応のあるカテゴリーについて Y と X

表 6.4　カテゴリーデータの行と列の入れ替え例

ケース番号	C_1	C_4	C_3	C_2	C_5
3	1	1	0	0	0
1	1	1	1	0	0
5	1	1	1	0	0
4	0	0	1	1	0
2	0	0	1	1	1

の組合せを作る。すなわち，表 6.3 の場合には，ケース 1 については（Y_1, X_1），（Y_1, X_3），（Y_1, X_4）の 3 組，ケース 3 については（Y_3, X_1），（Y_3, X_4）の 2 組などとなる。全体でこのような X， Y の組が N 個あったとき，X と Y の相関が最も高くなるようにすれば N 個の点は互に似ているもの同士が近くに配置されることになる。

なお，ケースとカテゴリーの相関が高くなるような数値の与えかたは何通りか考えられるので，最も相関が高くなる場合，次に相関が高くなる場合…という具合に，何通りかの解が存在する。それぞれの解は互に他と直交する（相関がない）ように選ばれるので，1 通りの解で解釈が十分できない場合には，いくつかの解を組合せて解釈するとよい。

6.2 数量化 IV 類

数量化 IV 類は，**類似度行列**を元にして，対象の**二次元配置**を求める。任意の基準による類似度行列を使用できるが，類似度行列は対称行列でなくてもよい。類似の方法として 6.5 節の主座標分析がある。

n 個の対象間の**類似度**を表す類似度行列を $\boldsymbol{S} = (S_{ij})$ とする。$S_{ii} = 0$，対象 i と対象 j が似ていないほど負の，絶対値の大きな値をとるとする。類似度行列は対称行列でなくてもよい。数量化 IV 類は，各対象に数値を割当て，対象間の**ユークリッド距離**が類似度の高い対象間では小さく，類似度の低い対象間では大きくなるようにすることを目的とする。

まず，1 次元の数量化を考える。対象 i に c_i，対象 j に c_j という数値を割当てると，対象 i と対象 j の間の**ユークリッド平方距離**は $\Delta_{ij}^2 = (c_i - c_j)^2$ となる。対象間のユークリッド距離が，類似度の高い対象間では小さく，類似度の低い対象間では大きくなるようにするので，（6.1）式で定義される Q が最大になるように c_i，c_j を定めればよいわけである。

$$Q = -\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}(c_i - c_j)^2 \tag{6.1}$$

1 次元だけでは類似度を十分説明できない場合には，対象 i に（c_{i1}，c_{i2}，…，c_{im}），対象 j に（c_{j1}，c_{j2}，…，c_{jm}）のように数値を割り当てる。このとき，対象 i と対象 j の間の m 次元空間でのユークリッド平方距離 Δ_{ij}^2 は（6.2）式のようになるので，これを（6.1）式に用いる。

$$\Delta_{ij}^2 = \sum_{k=1}^{m}(c_{ik} - c_{jk})^2 \tag{6.2}$$

（6.1）式は固有値問題を解くことになる。

6.3 クラスター分析

似通った個体あるいは変数のグループ化を行うための分析手法である。クラスター分析の結果は，図 6.1 のような**デンドログラム（樹状図）**として表現される。

個体が似通っているかどうかの判定基準としてはいくつかあるが，取り扱いが容易な**ユークリッド距離**を用いる[*1]。

n 個の個体について，p 個の変数 $X_{i1}, X_{i2}, \ldots, X_{ip}$（$i = 1, 2, \ldots, n$）があるとする。初期状態として，$n$ 個の**クラスター**があるとする（各クラスターは 1 個体ずつを含むと考える）。

[*1] 個体のクラスター分析を行う場合には，解析に用いるデータを正規化する場合としない場合では結果がかなり異なることがある。解析に使用する変数が異なった単位で表されているときには，正規化した方がよいかもしれない。しかし，ある変数が決定的な性質を持つ場合には，正規化することは他の変数と同格に取り扱ってしまうことになるので正規化しない方がよいかもしれない。

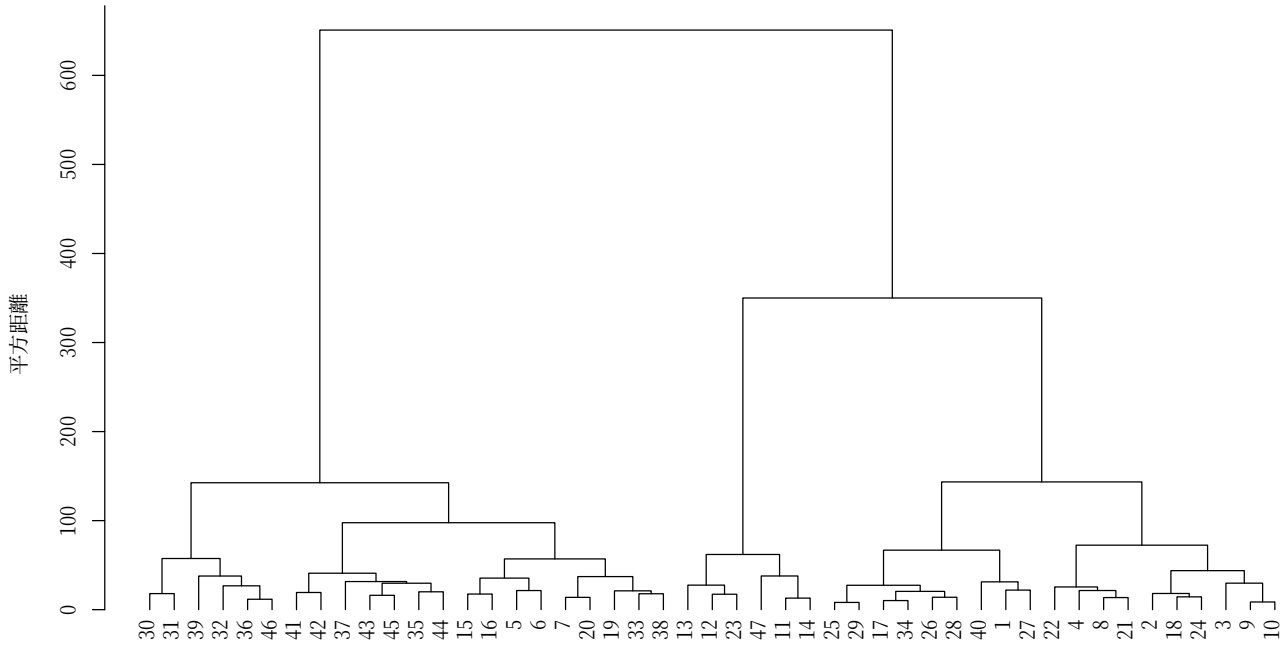

図 6.1　デンドログラム（樹状図）

- 第 1 段階

  クラスター間の**ユークリッド平方距離** d_{ij}^2 を計算する。

$$d_{ij}^2 = \sum_{k=1}^{p} (X_{ik} - X_{jk})^2, \qquad (i, j = 1, 2, \ldots, n) \tag{6.3}$$

- 第 2 段階

  ユークリッド平方距離の最も近いクラスターを併合して，1 つのクラスターとする。

  クラスター a とクラスター b が併合されてクラスター c が作られるとする。

  d_{ab}, d_{xa}, d_{xb} を，クラスター a とクラスター b が併合される前の各クラスター間の距離としたとき，併合後のクラスター c とクラスター x（$x \neq a$, $x \neq b$）との距離は（6.4），（6.5）式で表される。

$$d_{xc} = \alpha_a d_{xa} + \alpha_b d_{xb} + \beta d_{ab} + \gamma |d_{xa} - d_{xb}| \tag{6.4}$$

$$d_{xc}^2 = \alpha_a d_{xa}^2 + \alpha_b d_{xb}^2 + \beta d_{ab}^2 + \gamma |d_{xa}^2 - d_{xb}^2| \tag{6.5}$$

  α_a, α_b, β, γ は表 6.5 に示すような定数

- 第 3 段階

  2 個のクラスターが 1 個のクラスターにまとめられたので，総クラスター数が 1 個減る。クラスター数が 1 になるまで第 2 段階を繰返す。

（6.4）式または（6.5）式で併合後のユークリッド距離を計算するときの定数 α_a, α_b, β, γ をどのように選ぶかによって，表 6.5 に示す 7 種類のクラスター分析が行える。

各手法の分類感度は，クラスターの融合によって空間が拡散される場合に高く，濃縮される場合に低くなる。各手法の特徴は以下の通りである。

表 6.5 クラスター分析の各種法で距離の再定義において使用されるパラメータ

	α_a	α_b	β	γ	使用される式
最短距離法	0.5	0.5	0	−0.5	(6.4)
最長距離法	0.5	0.5	0	0.5	(6.4)
メディアン法	0.5	0.5	−0.25	0	(6.4)
重心法	n_a / n_c	n_b / n_c	$-(n_a n_b) / n_c^2$	0	(6.5)
群平均法	n_a / n_c	n_b / n_c	0	0	(6.5)
可変法	$(1-\beta^*)/2$	$(1-\beta^*)/2$	β^*	0	(6.5)
ウォード法	$\dfrac{n_x+n_a}{n_x+n_c}$	$\dfrac{n_x+n_b}{n_x+n_c}$	$-\dfrac{n_x}{n_x+n_c}$	0	(6.5)

β^* は 1 未満の任意の値

手法	特徴
ウォード法	最も明確なクラスターを作る（分類感度が高い）。
最近隣法	分類感度は低く，鎖状のクラスターを作る傾向がある。
最遠隣法	空間の拡散が起こり，分類感度は高い。
メディアン法	最近隣法と最遠隣法の折衷法である。
メディアン法，重心法	クラスター間の距離の逆転が生じる場合がある。
可変法	パラメータ（β）の選択によって空間の濃縮・拡散を制御できるので，バラエティーに富んだ結果を生み出す。β としては 1 未満の値を指定する。β の値が 1 に近いほど空間の濃縮が起こる（分類感度が低くなる）。負の値をとれば，空間の拡散が起こる（分類感度が高くなる）。一般に，−0.25 〜 0 の範囲の値を与えるのがよいといわれている。

変数のクラスター分析を行う場合には，変数 i と変数 j の相関係数を r_{ij} としたとき，2 変数間の距離が（6.6）式で表されることになるので，個体のクラスター分析と同じように取り扱うことができる。

$$d_{ij}^2 = 2\,(1 - r_{ij}), \qquad (i, j = 1, 2, \ldots, p) \tag{6.6}$$

6.4 クロンバックの α 信頼性係数

アンケート調査などで，対象とする領域のある特性を測定するために複数の質問項目への答えの合計値（特に**尺度得点**と呼ばれる）を使うことがある。

クロンバックの α 信頼性係数は，尺度に含まれる個々の質問項目が**内的整合性**を持つかどうか（目的とする特性を測定する質問項目群であるか）を判定するために用いられる。

k 個の変数の合計点 Y の不偏分散を S_Y^2，変数 i の不偏分散を S_i^2，としたとき，α 信頼性係数は（6.7）式で求められる。

$$\alpha = \frac{k}{k-1}\left(1 - \sum_{j=1}^{k} S_j^2 / S_Y^2\right) \tag{6.7}$$

クロンバックの α 信頼性係数は，通常，0.8 以上でなければ妥当な尺度とはみなせない。尺度内に目的とする特性を測定するとはいえない質問項目が含まれていると，クロンバックの α 信頼性係数の値が小さくなる。そこで，このような不適当な質問項目を特定するために以下のような補助的な分析を行うとよい。

α'　当該アイテムを除いたときのクロンバックの α 信頼性係数
この数値が大きい場合は，当該アイテムは尺度を構成するアイテムとしては不適切であることを意味する。

r'　当該アイテムを除いた合計と当該アイテムとの相関係数
この数値が小さい場合は，当該アイテムは尺度を構成するアイテムとしては不適切であることを意味する。

R^2　当該アイテムとそれ以外のアイテムの重相関係数の二乗
これは，当該アイテムを従属変数とし，それ以外のアイテムを独立変数としたときの重回帰分析で得られるものである。従ってこの数値が小さいアイテムは，他のアイテムと共通部分が小さく，尺度を構成するアイテムとしては不適切であることを意味する。

6.5　主座標分析

類似度行列を元にして，対象の**二次元配置**を求める。任意の基準による類似度行列を使用できるが，類似度行列は対称行列でなければならない[*2]。

n 個の対象間の**類似度**を表す類似度行列を $\boldsymbol{S} = (S_{ij})$ とする。$S_{ii} = 0$，$S_{ij} = S_{ji}$，対象 i と対象 j が似ていないほど負の，絶対値の大きな値をとるとする。$\boldsymbol{S}$ の固有値が $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n \geqq 0$，第 k 固有値に対応する固有ベクトルを $\boldsymbol{c}_k = (c_{1k}, c_{2k}, \ldots, c_{nk})'$，$\| \boldsymbol{c}_k \|^2 = \lambda_k$ とする。対象 i の座標を $(c_{i1}, c_{i2}, \ldots, c_{in})$，対象 j の座標を $(c_{j1}, c_{j2}, \ldots, c_{jn})$ としたとき，対象 i と対象 j の間の**ユークリッド平方距離** Δ_{ij}^2 は以下のようになる。

$$\Delta_{ij}^2 = \sum_{k=1}^{n} (c_{ik} - c_{jk})^2 = S_{ii}^2 + S_{jj}^2 - 2S_{ij} = -2S_{ij} \tag{6.8}$$

もし，類似度行列 $\boldsymbol{S}$ の固有値の m 番目以降の固有値が小さいならば，n 次元空間に対象を配置するかわりに m 次元空間に配置しても対象間の位置関係に大きな誤差はないであろう。

$$\Delta_{ij}^2 \quad \sum_{k=1}^{m} (c_{ik} - c_{jk})^2 \tag{6.9}$$

主座標分析では特に $m = 2$ すなわち二次元平面上に対象を布置することが行われる。この場合に，(6.10) 式により，二次元までの近似の効率を判定できる。

$$\text{寄与率} = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \cdots + \lambda_n} \tag{6.10}$$

6.6　多重ロジスティックモデル

予後を**予測**するための**多重ロジスティックモデル**を適用する。

基準変数が，ある事象があったかなかったかのような **0/1 型のデータ**の場合に，重回帰式を求めると，予測値は負の値や 1 より大きい値をとるので不適当である。このような場合には (6.11) 式のようなロジスティックモデルが適用できる。ある事象が発生する確率を P としたとき，$P/(1-P)$ は**オッズ比**，その対数をとった $\log\{P/(1-P)\}$ は**ロジット**または**対数オッズ**と呼ばれる。ロジットが独立変数の線形結合式で表せるとするのがロジスティックモデルである。

$$\log\left(\frac{P}{1-P}\right) = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p = \lambda \tag{6.11}$$

これを変形すると，(6.12) 式のロジスティック関数が得られる。P は 0 ～ 1 の範囲の値をとる。

$$P = \frac{1}{1 + \exp(-\lambda)} = \frac{1}{1 + \exp\{-(b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p)\}} \tag{6.12}$$

[*2] 同様の目的を達成する手法として 6.1 節の**数量化 IV 類**がある。数量化 IV 類では，分析に用いる類似度行列は対称行列でなくともよい。

$b_0, b_1, b_2, \ldots, b_p$ は最尤法によって求めることができる。最尤法で係数を求める場合には初期値が必要であるが，Truett–Cornfield による判別係数を初期値とすることで，たいていの場合はうまく行く。

付録 A

スカラー，ベクトル，行列について

A.1　スカラー

スカラーとは一つの数値（変数）で表されるものである。例えば，「A 君の身長は 178.3 cm である」とか，「温度が 20.4℃ である」というような場合，178.3 とか 20.4 はスカラーである。

A.2　ベクトル

ベクトルとは複数個の数値の組で表されるものである。例えば，体格を考えるとき，「身長，体重，胸囲，坐高」の 4 つの測定値をひとまとめにして取り扱うと便利である。ベクトルはスカラーの集合である。ベクトルを構成するスカラーを「ベクトルの要素」という。

● ベクトルは，行ベクトルと列ベクトルという二つの表し方がある。

(a_1, a_2, a_3, a_4)　　行ベクトル　　$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}$　　列ベクトル

一般に，列ベクトルを基準にすることが多いので，$\boldsymbol{a} = (a_1, a_2, a_3, a_4)'$ のように表記される。$'$ は転置（行と列を入れ替えること）を表す。

● 二つのベクトルが等しい $\boldsymbol{a} = \boldsymbol{b}$ ということは，対応する要素がすべて等しい $a_1 = b_1, a_2 = b_2, \ldots, a_p = b_p$ ということである。

● スカラーとベクトルの積は，各要素の定数倍を要素とするベクトルである。

$$k\,\boldsymbol{a} = k(a_1, a_2, \ldots, a_p) = (k\,a_1, k\,a_2, \ldots, k\,a_p)$$

なお，$k\,\boldsymbol{a} = \boldsymbol{a}\,k$ である。

● $k_1 = k_2 = 0$ でない限り

$$k_1\,\boldsymbol{a} + k_2\,\boldsymbol{b} = \boldsymbol{0}$$

となるようなスカラ k_1，k_2 が存在しないならば，ベクトル $\boldsymbol{a}$ と $\boldsymbol{b}$ は**線形独立**であるという（$\boldsymbol{0}$ は要素が全て 0 であるようなベクトル）。

● 要素数が同じベクトル同士の和（差）は以下のようにして定義される。結果はベクトルになる。

$$\boldsymbol{a} \pm \boldsymbol{b} = (a_1, a_2, \ldots, a_p) \pm (b_1, b_2, \ldots, b_p) = (a_1 \pm b_1, a_2 \pm b_2, \ldots, a_p \pm b_p)$$

● 要素数が同じベクトル同士の**内積**は以下のようにして定義される。結果はスカラーになる。

$$\boldsymbol{a}'\,\boldsymbol{b} = (a_1, a_2, \ldots, a_p) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = a_1 b_1 + a_2 b_2 + \ldots + a_p b_p$$

つまり，対応する要素の積和であるので，

$$\boldsymbol{a}'\,\boldsymbol{b} = \boldsymbol{b}'\,\boldsymbol{a}$$

である。

また，任意のベクトルの自分自身との内積は，要素の二乗和である。

$$\boldsymbol{a}'\,\boldsymbol{a} = a_1^2 + a_2^2 + \ldots + a_p^2 = \sum_{i=1}^{p} a_i^2$$

● ベクトルの長さ（**ノルム**）$\|\boldsymbol{a}\|$ はベクトルの内積 $\boldsymbol{a}'\,\boldsymbol{a}$ の平方根である。

$$\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}'\,\boldsymbol{a}}$$

● 二つのベクトルがなす角を θ とすると，

$$\cos\theta = \frac{\boldsymbol{a}'\,\boldsymbol{b}}{\|\boldsymbol{a}\|\,\|\boldsymbol{b}\|}$$

二つのベクトルが**直交**するというのは，$\cos\theta = 0$ となるときであり，そのときにだけ内積が 0 になる。逆にいえば，「内積が 0 になるときに，二つのベクトルが直交する」といえる。

なお，直交ベクトルで各ベクトルの長さが 1 であるとき，**正規直交**であるという。

● 要素数がそれぞれ m，n 個のベクトル同士の**外積**は以下のようにして定義される。結果は $m \times n$ 行列になる。

$$\boldsymbol{a}\,\boldsymbol{b}' = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} (b_1, b_2, b_3) = \begin{pmatrix} a_1 b_1 & a_1 b_2 & a_1 b_3 \\ a_2 b_1 & a_2 b_2 & a_2 b_3 \\ a_3 b_1 & a_3 b_2 & a_3 b_3 \\ a_4 b_1 & a_4 b_2 & a_4 b_3 \end{pmatrix}$$

A.3 行列

複数人について体格を表すベクトルがあるとき，「ベクトルのベクトル」を考えると便利である。行列は $\boldsymbol{X}$ のように太い英大文字で表す。以下に示す行列 $\boldsymbol{X}$ は，「3 行 4 列の行列」という。「4 要素を持つ列ベクトルが 3 個ある」と考えてもよいし，「3 要素を持つ列ベクトルが 4 個ある」と考えてもよい。ベクトルは行列の一種である。

$$\boldsymbol{X} = \begin{pmatrix} (x_{11}, x_{12}, x_{13}, x_{14}) \\ (x_{21}, x_{22}, x_{23}, x_{24}) \\ (x_{31}, x_{32}, x_{33}, x_{34}) \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \end{pmatrix}$$

行数と列数が同じである行列を**正方行列**と呼ぶ。

● 元の行列 $\boldsymbol{A}$ の行と列を入れ替えた行列を転置行列 $\boldsymbol{A}'$ という。

$$\boldsymbol{A} = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \quad , \quad \boldsymbol{A}' = \begin{pmatrix} a & d \\ b & e \\ c & f \end{pmatrix}$$

● 対角要素が 0 以外の値で，それ以外の要素が 0 である正方行列を**対角行列**と呼ぶ。

$$\boldsymbol{T} = \begin{pmatrix} t_{11} & 0 & 0 \\ 0 & t_{22} & 0 \\ 0 & 0 & t_{33} \end{pmatrix}$$

● 対角要素が 1 で，それ以外の要素が 0 である正方行列を**単位行列**と呼ぶ。

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

● 対角要素よりも下の要素が 0 である正方行列を**上三角行列**と呼ぶ。

$$\boldsymbol{I} = \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}$$

● 対角要素よりも上の要素が 0 である正方行列を**下三角行列**と呼ぶ。

$$\boldsymbol{I} = \begin{pmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{pmatrix}$$

● 対角要素を中心として対応する要素が等しい正方行列を**対称行列**と呼ぶ。あるいは，転置しても同じ行列といっても良い（$\boldsymbol{A} = \boldsymbol{A}'$）。

$$\boldsymbol{I} = \begin{pmatrix} a & b & d \\ b & c & e \\ d & e & f \end{pmatrix}$$

● 二つの行列が等しい $\boldsymbol{A} = \boldsymbol{B}$ ということは，対応する要素がすべて等しい $a_{ij} = b_{ij}$ ということである。

● 行列同士の和（差）は，以下のように定義される。結果は行列になる。

$$\begin{aligned} \boldsymbol{X} \pm \boldsymbol{Y} &= \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \end{pmatrix} \pm \begin{pmatrix} y_{11} & y_{12} & y_{13} & y_{14} \\ y_{21} & y_{22} & y_{23} & y_{24} \\ y_{31} & y_{32} & y_{33} & y_{34} \end{pmatrix} \\ &= \begin{pmatrix} x_{11} \pm y_{11} & x_{12} \pm y_{12} & x_{13} \pm y_{13} & x_{14} \pm y_{14} \\ x_{21} \pm y_{21} & x_{22} \pm y_{22} & x_{23} \pm y_{23} & x_{24} \pm y_{24} \\ x_{31} \pm y_{31} & x_{32} \pm y_{32} & x_{33} \pm y_{33} & x_{34} \pm y_{34} \end{pmatrix} \end{aligned}$$

● スカラーと行列の積は，以下のように定義される。結果は行列になる。

$$k\,\boldsymbol{X} = k \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \end{pmatrix} = \begin{pmatrix} k\,x_{11} & k\,x_{12} & k\,x_{13} & k\,x_{14} \\ k\,x_{21} & k\,x_{22} & k\,x_{23} & k\,x_{24} \\ k\,x_{31} & k\,x_{32} & k\,x_{33} & k\,x_{34} \end{pmatrix}$$

なお，$k\,\boldsymbol{X} = \boldsymbol{X}\,k$ である。

● $a \times b$ 行列と $b \times c$ 行列をこの順で掛けるときのみ行列の積が定義できる。積は以下のように定義され，結果は $a \times c$ 行列になる。

$$\boldsymbol{X}\,\boldsymbol{Y} = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{pmatrix} \times \begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ y_{31} & y_{32} \end{pmatrix}$$

$$= \begin{pmatrix} x_{11}\,y_{11} + x_{12}\,y_{21} + x_{13}\,y_{31} & x_{11}\,y_{12} + x_{12}\,y_{22} + x_{13}\,y_{32} \\ x_{21}\,y_{11} + x_{22}\,y_{21} + x_{23}\,y_{31} & x_{21}\,y_{12} + x_{22}\,y_{22} + x_{23}\,y_{32} \\ x_{31}\,y_{11} + x_{32}\,y_{21} + x_{33}\,y_{31} & x_{31}\,y_{12} + x_{32}\,y_{22} + x_{33}\,y_{32} \\ x_{41}\,y_{11} + x_{42}\,y_{21} + x_{43}\,y_{31} & x_{41}\,y_{12} + x_{42}\,y_{22} + x_{43}\,y_{32} \end{pmatrix}$$

● $\boldsymbol{A}$ が $n \times n$ 対称行列，$\boldsymbol{X}$ が $n \times 1$ 列ベクトルとしたとき，$\boldsymbol{X}'\,\boldsymbol{A}\,\boldsymbol{X}$ は，**二次形式**と呼ばれる。

$$\boldsymbol{X}'\,\boldsymbol{A}\,\boldsymbol{X} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

例：

$$\boldsymbol{X}'\,\boldsymbol{A}\,\boldsymbol{X} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 2\,x_1^2 + 2x_1\,x_2 + 3\,x_2^2$$

● 正方行列 $\boldsymbol{A}$ には**行列式**というスカラーが計算される。行列式を $|\boldsymbol{A}|$ と表す。

1. 2×2 行列

$$\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{の行列式は } |\boldsymbol{A}| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = a\,d - b\,c$$

である。

2. 3×3 行列

$$\boldsymbol{A} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix}$$

の行列式は $|\boldsymbol{A}| = a\,e\,k + b\,f\,g + c\,d\,h - a\,h\,f - d\,b\,k - g\,e\,c$ である。

3. 大きな行列の行列式は 2×2 の小行列式を求める問題に帰着できるが計算がやっかいなので，コンピュータプログラムに依るのがよい。Excel にも mdeterm という関数がある。

● 正方行列の行列式が 0 であるときには**特異**であるという。行列式が 0 でないときにはその行列は**正則**であるという。

● 行列の**階数**とは，その行列中の線形独立なベクトルの数に等しい。

● 行列 $\boldsymbol{A}$ に行列 $\boldsymbol{B}$ を掛けて結果が単位行列になるとき，$\boldsymbol{B}$ を，行列 $\boldsymbol{A}$ の**逆行列** $\boldsymbol{B} = \boldsymbol{A}^{-1}$ という。特異行列には逆行列はない。

1. 1×1 行列の場合（これはスカラーである）(a) の逆行列は ($1/a$) である。
2. 2×2 行列の場合は以下のようになる。

$$\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad , \qquad \boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|} \begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$$

3. 3×3 行列の場合は以下のようになる。

$$\boldsymbol{A}=\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix} \quad , \quad \boldsymbol{A}^{-1}=\frac{1}{|\boldsymbol{A}|}\begin{pmatrix} ek-fh & -bk+ch & bf-ce \\ -dk+fg & ak-cg & -af+cd \\ dh-eg & -ah+bg & ae-bd \end{pmatrix}$$

数値例：

$$\boldsymbol{A}=\begin{pmatrix} 2 & 2 & 3 \\ 3 & 3 & 2 \\ 3 & 1 & 5 \end{pmatrix} \quad , \quad \boldsymbol{A}^{-1}=\begin{pmatrix} -1.3 & 0.7 & 0.5 \\ 0.9 & -0.1 & -0.5 \\ 0.6 & -0.4 & 0.0 \end{pmatrix}$$

4. 4×4 行列以上の場合には一般公式によって逆行列を求めるのはやっかいであるため，コンピュータによる。例えば Excel には minverse という関数がある。
なお，対角行列の逆行列は対角成分がもとの対角成分の逆数になるだけである。

$$\boldsymbol{T}=\begin{pmatrix} t_{11} & 0 & 0 \\ 0 & t_{22} & 0 \\ 0 & 0 & t_{33} \end{pmatrix} \quad , \quad \boldsymbol{T}^{-1}=\begin{pmatrix} 1/t_{11} & 0 & 0 \\ 0 & 1/t_{22} & 0 \\ 0 & 0 & 1/t_{33} \end{pmatrix}$$

● $p\times p$ 正方行列 $\boldsymbol{A}$ において，対角要素の合計をトレースと呼ぶ。

$$tr(\boldsymbol{A})=\sum_{i=1}^{p} a_{ii}$$

● 行列 $\boldsymbol{A}$ を実対称行列，列ベクトルを $\boldsymbol{x}$，λ を定数値としたとき，

$$\boldsymbol{A}\,\boldsymbol{x}=\lambda\,\boldsymbol{I}\,\boldsymbol{x}$$

の関係が成り立つとき，λ は**固有値**，$\boldsymbol{x}$ はその固有値に対応する**固有ベクトル**と呼ばれる。これらを求めることは**固有値問題**といわれる。上の式を書き直すと

$$(\boldsymbol{A}\ -\lambda\,\boldsymbol{I})\,\boldsymbol{x}=0$$

となり，これは

$$\begin{pmatrix} a_{11}-\lambda & a_{12} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda \end{pmatrix}\times\begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \end{pmatrix}=\ 0$$

の形をした線形連立方程式を解くことを表す。

この連立方程式が自明な解以外を持つのは $(\boldsymbol{A}\ -\lambda\,\boldsymbol{I})$ が**特異**な場合，すなわち，$|\boldsymbol{A}\ -\lambda\,\boldsymbol{I}|=0$ のときに限られる。

例：

$$\boldsymbol{A}=\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$$

のとき，

$$|\boldsymbol{A}\ -\lambda\,\boldsymbol{I}|=\begin{vmatrix} 2-\lambda & 1 \\ 1 & 3-\lambda \end{vmatrix}=\lambda^2-5\,\lambda+5=0$$

となるので，$\lambda=3.618034, 1.381966$ を得る。大きい方から順に，第 1 固有値（λ_1），第 2 固有値（λ_2），…と呼ぶ。

それぞれの固有値に対応して固有ベクトルが得られるが，これはもとの方程式に λ の値を代入することにより求める。

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\,\boldsymbol{X} = \begin{pmatrix} 2-\lambda & 1 \\ 1 & 3-\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

まず $\lambda = 3.618034$ とすることにより，次の連立方程式が得られる。

$$\begin{cases} -1.618034\,x_1 + x_2 = 0 \\ x_1 - 0.618034\,x_2 = 0 \end{cases}$$

を解いて，$x_1 = 0.618034\,x_2$ が得られるが，この関係を満たす x_1 と x_2 は無数にある。そこで，普通は $\boldsymbol{x}'\boldsymbol{x} = \|\boldsymbol{x}\| = 1$ とする（ノルムを 1 にする。このようなベクトルを**正規固有ベクトル**という）。

$x_1^2 + x_2^2 = 0.618034^2\,x_2^2 + x_2^2 = 1.381966\,x_2^2 = 1$ より，$x_2 = 0.85065$，さらに，$x_1 = 0.52573$ となる。つまり，$\lambda_1 = 3.618034$ に対応する固有ベクトルは $\boldsymbol{x}' = (0.52573, 0.85065)$ である。

同様にして，$\lambda_2 = 1.381966$ に対応する固有ベクトルは $\boldsymbol{x}' = (0.85065, -0.52573)$ である。

なお，固有ベクトルは互いに直交する。

$\boldsymbol{x}'\boldsymbol{x} = (0.52573, 0.85065)\,(0.85065, -0.52573)' = 0$ である。

3×3 以上の対称行列の固有値・固有ベクトルを求めるのはやっかいであるので，コンピュータプログラムで求めるのがよい。Excel には該当の関数はないが，Mathematica では可能である。

パワー法というアルゴリズムに従えば，手数はかかるが Excel ででも求めることはできる。

付録 B

演習問題の解

B.1　第 2 章の演習問題

例解 2.1

各変数において，平均偏差を計算すると，表 B.1 のようになる。

表 B.1　平均偏差および予測値

ケース番号	X_1	X_2	Y	$X_1 - \bar{X}_1$	$X_2 - \bar{X}_2$	$Y - \bar{Y}$	$\hat{Y}$
1	1.2	1.9	0.9	–4.2	–1.4	–1.3	0.939
2	1.6	2.7	1.3	–3.8	–0.6	–0.9	1.250
3	3.5	3.7	2.0	–1.9	0.4	–0.2	1.926
4	4.0	3.1	1.8	–1.4	–0.2	–0.4	1.856
5	5.6	3.5	2.2	0.2	0.2	0.0	2.298
6	5.7	7.5	3.5	0.3	4.2	1.3	3.465
7	6.7	1.2	1.9	1.3	–2.1	–0.3	1.864
8	7.5	3.7	2.7	2.1	0.4	0.5	2.744
9	8.5	0.6	2.1	3.1	–2.7	–0.1	2.060
10	9.7	5.1	3.6	4.3	1.8	1.4	3.596
平均値	5.4	3.3	2.2				

正規方程式は，

$$\begin{cases} 71.98\, b_1 + 6.46\, b_2 = 16.58 \\ 6.46\, b_1 + 35.30\, b_2 = 11.44 \end{cases} \tag{B.1}$$

となり，これを解いて，$b_1 = 0.20462$, $b_2 = 0.28663$ となる。また，$b_0 = 2.2 - 0.20462 \cdot 5.4 - 0.28663 \cdot 3.3 = 0.14918$ となる。すなわち，求める重回帰式は $\hat{Y} = 0.20462\, X_1 + 0.28863\, X_2 + 0.14918$ である。

標準化偏回帰係数は $S_{yy} = 6.7$ であるから，$b'_1 = 0.20462\ \sqrt{(71.98/6.7)} = 0.67067$，$b'_2 = 0.28663\ \sqrt{(35.3/6.7)} = 0.65793$ である。

偏回帰係数が 0 であるという帰無仮説の検定に用いる t 値は，それぞれ 27.05056，26.53645 となり（自由度 7 の t 分布に従う），帰無仮説は棄却される。また，定数項が 0 であるという帰無仮説の検定の t 値は 2.73685 で，同じく帰無仮説は棄却される（有意水準 5%）。

分散分析は表 B.2 のようになる。

重相関係数の 2 乗は，$R^2 = 1 - 0.02834/6.70000 = 0.99577$ となる。

表 B.2　分散分析表

要因	平方和	自由度	平均平方	F 値	P 値
回帰	6.67164	2	3.33582	823.47652	0.00000
残差	0.02836	7	0.00405		
全体	6.70000	9	0.74444		

例解 2.2

以下の pdf ファイルを参照せよ。

`http://aoki2.si.gunma-u.ac.jp/LaTeX/sreg-qt1.pdf`

B.2　第 3 章の演習問題

例解 3.1

図 B.1 のような例を考えてみると，平均値が全く同じ変数であっても 2 群の判別に役立つことがわかる。

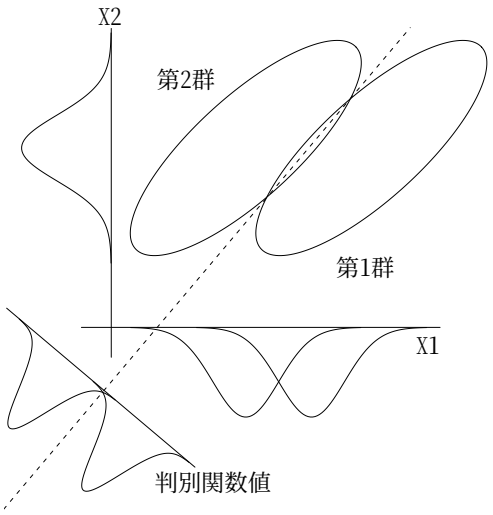

図 B.1　平均値が全く同じ変数が判別に役立つ場合の模式図

例解 3.2

各変数の平均値の差のベクトルは,

$$\boldsymbol{d} = (3.5 - 7.5,\ 6.5 - 4.5)' = (-4, 2)'$$

各群の変動・共変動行列は,

$$\boldsymbol{S}^{(1)} = \begin{pmatrix} 39.5 & 7.5 \\ 7.5 & 21.5 \end{pmatrix}, \qquad \boldsymbol{S}^{(2)} = \begin{pmatrix} 23.5 & 14.5 \\ 14.5 & 33.5 \end{pmatrix}$$

となり，プールした分散・共分散行列 $\boldsymbol{V}$ は,

$$\boldsymbol{V} = \begin{pmatrix} 6.3 & 2.2 \\ 2.2 & 5.5 \end{pmatrix}$$

であるので，解くべき連立方程式は,

$$\begin{cases} 6.3\,a_1 + 2.2\,a_2 = -4 \\ 2.2\,a_1 + 5.5\,a_2 = 2 \end{cases}$$

これを解いて，$a_1 = -0.88561$，$a_2 = 0.71788$ となるので，判別式 DF_0 は,

$$DF_0 = -0.88561\,X_1 + 0.71788\,X_2 + c$$

となる。

c は任意に決めてよいが, 各変数の「平均値の平均値」を判別式に代入して得られる数値の符号を変えたものにすると, DF の正負で群を判別できるので便利である。今の場合は, 平均値の平均値は, 5.5, 5.5 なので, $DF_0 = -0.92251$ となり，$c = 0.92251$ とする。すなわち最終的な判別関数は，$DF = -0.88561\,X_1 + 0.71788\,X_2 + 0.92251$。

判別結果は表 B.3 のようになる。

表 B.3　判別値および判別結果

群	ケース	X_1	X_2	判別値	判別群
1	1	5	10	3.67326	1
1	2	0	7	5.94767	1
1	3	4	7	2.40523	1
1	4	8	6	–1.85508	2
1	5	2	5	2.74069	1
1	6	2	4	2.02281	1
2	1	10	8	–2.19054	2
2	2	7	7	–0.25159	2
2	3	9	5	–3.45857	2
2	4	5	3	–1.35190	2
2	5	9	2	–5.61221	2
2	6	5	2	–2.06978	2

例解 3.3

以下の pdf ファイルを参照せよ。

http://aoki2.si.gunma-u.ac.jp/LaTeX/sdis-qt2.pdf

B.3　第 4 章の演習問題

例解 4.1

表 B.4，B.5，B.6 のようになる。

表 B.4　標準化されたデータ行列と合成変数

No.	x_1	x_2	x_3	$\boldsymbol{f}$	$\boldsymbol{g}$
1	0.94065	1.23117	2.24179	2.19759	1.49349
2	–1.95365	–1.17111	0.24019	–0.00995	–0.00676
3	0.57886	–0.87083	0.64051	1.51649	1.03061
4	–0.14471	0.63060	–1.36109	–1.94792	–1.32381
5	0.94065	–0.57054	–0.16013	0.63651	0.43257
6	0.21707	1.23117	–1.36109	–2.12736	–1.44577
7	1.30243	–0.57054	0.64051	1.69811	1.15404
8	0.21707	0.33031	–0.56045	–0.70615	–0.47990
9	–1.23008	1.23117	–0.16013	–1.52988	–1.03971
10	–0.86829	–1.47140	–0.16013	0.27255	0.18523
$E(x_i)$	0.00000	0.00000	0.00000	0.00000	0.00000
$V(x_i)$	1.00000	1.00000	1.00000	2.16515	1.00000

表 B.5　相関係数行列

	x_1	x_2	x_3
x_1	1.00000	0.16730	0.28097
x_2	0.16730	1.00000	–0.10338
x_3	0.28097	–0.10338	1.00000

表 B.6　相関係数ベクトルと標準重みベクトル

	$\boldsymbol{a}$	$\boldsymbol{w}_s$
x_1	0.48163	0.33980
x_2	–0.42820	–0.40776
x_3	0.88519	0.74756

索引

英字

あ

か

さ

た

な

は

ま

や

ら

わ