多変量解析 講義ノート

村田 昇

2020 年 8 月 26 日

早稲田大学 先進理工学部
電気・情報生命工学科

はじめに

多変量解析法は，複数の特徴量をもつデータを解析するために開発された方法の総称である．この講義では，その中でも基本的な方法に的を絞り，その考え方を修得することを目的とする．

データを縮約してその構造をより鮮明に捉えるために，多数の特徴量の線形結合によって新たな特徴量を構成する手法として回帰分析・主成分分析を，多数の変量の間に内在する関係を探り出し，それを手掛りにデータを分類する手法として判別分析・クラスタ分析を学ぶ．また有無といった本来量ではなく質を表わすような変数を数量化して，上記手法を適用する方法についても学ぶ．

なお，数理統計学と線形代数の基礎的な知識は既に学んでいることを前提とする．

目次

準備 1

1.1 多変量解析とは

1.1.1 目的

多変量データとは大まかに言えば多次元の観測データのことである．多次元データの中から数学的に，あるいはより限定して言えば確率的に記述される構造を抽出し，データの解析・予測・判別などを行う様々な手法を総称して**多変量解析** (multivariate analysis) という．

多変量解析の多くの手法は，多次元データの持つ本質的な性質を損なわずに，できる限り低い次元にデータを縮約した簡潔な記述を求める方法と捉えることもできる．特に人が直接的・直感的に理解できるような 1 次元，2 次元もしくは 3 次元空間への射影を通して，多次元のデータの持つ重要な性質を視覚的に把握するため (視覚化) に利用されることも多い．

高次元データを低次元に変換する方法は無数にあるが，古典的な多変量解析では基本的に線形変換を用いてこれを行う．線形変換を用いるのは，主に

- 最も良い変換を求める方法が解析的に計算しやすい
- 変換後の変数の確率法則について詳しく議論できる場合が多い

といった計算上・理論上の理由からである．また，線形変換には

- 非線形の対応関係を最も簡潔に近似したもの

としての側面もあり，より複雑な非線形性を含む状況を理解するための局所的な基本モデルとしての役割も重要である．

1.1.2 参考文献および資料

教科書は特に指定しないが，参考書としては以下を推薦する．

- 多変量解析入門, 永田靖・棟近雅彦，サイエンス社.
- 数理統計学，竹内啓，東洋経済社.
- 確率・統計入門，小針晛宏，岩波書店.
- 「逆」引き統計学，Gopal K. Kanji (池谷裕二，久我奈穂子訳)，講談社.

演習で用いる統計言語 R については以下のサイトに十分な情報がある．

- R 本家: http://www.r-project.org
- RjpWiki: http://www.okada.jp.org/RWiki/

また最近では多数の書籍が出版されているので，それらを手に取ってみるのも良いだろう．

1.2 データの取り扱い

1.2.1 用語

以下では多変量解析に特徴的な用語を説明する．
まず，解析の対象を表している多変量という言葉であるが，これは

変量 (variate) 対象の特徴を表す何らかの量

多変量 (multivariate) 多次元の変量

を表す．端的に言えば多次元観測データのことであり，解析においては多次元の確率変数として扱う．単一の確率変数においては，平均や分散といった統計量を用いてその確率変数の性質を捉えることができるが，多変量解析の場合，個々の変数の性質や特徴だけでなく，変数間の関係が解析の重要な対象となる．
変量は大きく分けると，以下のように質的なものと量的なものがある．

質的変数 区分だけ決められていて数値にならないもの
例: 犬・猫，血液型，成績，アンケートの回答など

量的変数 数値として自然に扱うことのできるもの
例: 身長，体重，電圧，電流，気温，株価など

質的変数は分野によっては (カテゴリ) ラベルやクラスと呼ばれることもある．量的変数は典型的には実験などで得られる数値データであり，和や積といった操作が自然に定義されるものを想定している．質的変数は適当な数値を割り当てて表現されることも多いが，与えられた数値をそのまま計算に用いることは必ずしも適切でないので，注意しなくてはならない．
更に質的変数は以下の 2 種類に分類される．

名義尺度 区別があるだけで順序付けできないもの
例: 犬・猫，血液型など

順序尺度 順序関係や大小関係が決められるもの
例: 成績 (A+,A,B,C,F)，アンケートの回答 (良い，ふつう，悪い) など

これらの違いは，与えられたデータのみからは区別できないことが多いので，必要に応じて変数の内容に踏み込んで考えることが重要である．
量的変数も以下の 2 種類に分類される．

比率尺度 原点が決められていて，値の比に意味があるもの
例: 電流，年齢など

間隔尺度 原点は自由にとれ，値の差に意味があるもの
例: 電圧，気温など

これらについても，数値からだけではどちらの属性かは区別できないので，変数の特性について理解しておく必要がある．

1.2.2 記法

多変量解析においては，多次元の確率変数ベクトルを一つの観測値 (データ) と考え，これが多数観測される状況を想定し解析を行う．

例えば，1 つのデータを p 次元のベクトル $\boldsymbol{x}$ とし，その各成分を $x_1, x_2, \ldots, x_p$ で表す．このとき 1 つの**観測値ベクトル** $\boldsymbol{x}$ は列ベクトル $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\mathsf{T}$ として扱う．一方，複数のデータを扱う場合には，各データを行ベクトルとして n 個の観測値を並べて，一つの行列

$$X = \begin{pmatrix} \boldsymbol{x}_1^\mathsf{T} \\ \boldsymbol{x}_2^\mathsf{T} \\ \vdots \\ \boldsymbol{x}_n^\mathsf{T} \end{pmatrix} = \overset{p \text{ 次元}}{\begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}} \; n \text{ 次元}$$

として表す．観測値を纏めた行列 (以後**観測値行列**と呼ぶ) は大文字で表し，観測ベクトルは小文字の太字で表す．

一般に，多変量解析においてデータを行列として並べるときには，上記のように行が観測データの番号，列が観測データの各要素に対応する標記が用いられる．

1.2.3 質的なデータの扱い

以降の章では，カテゴリラベルなどの目的変数を除き，確率変数は基本的に量的変数として扱う．質的変数の場合には数量化という操作により，量的変数と同様に扱うことが可能となる．例えば $A+, A, B, C$ で表される成績のような順序尺度の場合，

$$\begin{aligned} A+ &\to 95 \\ A &\to 85 \\ B &\to 75 \\ C &\to 65 \end{aligned}$$

という対応関係により，その順番に矛盾が生じないように数字を割り振ることがある．このとき以下のデータはこの規則によって

$$\begin{pmatrix} B \\ A \\ A \\ C \\ A+ \\ B \end{pmatrix} = \begin{pmatrix} 75 \\ 85 \\ 85 \\ 65 \\ 95 \\ 75 \end{pmatrix}$$

に変換されるが，その値は恣意的に決められたものであることに注意しなくてはならない．例えば $A+$ と A の能力差は，B と C の能力差と同じとは限らないからである．こうした恣意性を避ける最も単純な方法は以下のように質的変数がどのカテゴリに含まれ

るかを，0, 1 のベクトルとして表す方法である．

$$\begin{aligned} A+ &\to (1,0,0,0) \\ A &\to (0,1,0,0) \\ B &\to (0,0,1,0) \\ C &\to (0,0,0,1) \end{aligned}$$

これをダミー変数と呼ぶことがある．このとき上の例のデータは

$$\begin{pmatrix} B \\ A \\ A \\ C \\ A+ \\ B \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

と変換される．ただし，この行列は行方向の和が常に 1 となり冗長な表現になっているため，通常はどこか 1 列 (例えば第 1 列) を取り除いた 6×3 の行列によって質的変数を量的変数に変換することになる．

1.3 対象とする手法

講義で取り扱うのは，回帰分析 (単回帰，重回帰)，主成分分析，判別分析 (線形，2 次)，クラスタ分析，多次元尺度構成法という多変量解析における最も基本的な分析方法である．いずれも行列・ベクトルを用いた線形演算によって効率良く分析を行うことができる．

例えば，多数の人の身長・体重・年齢・血圧・血糖値・心臓病の有無・100m 走のタイムなどのデータがあったとする．ここから

- 身長から体重を予想する
- 身長と体重から 100m のタイムを予想する
- 身長と体重から体の大きさを表す指標を作成する
- 心臓病の有無以外のデータから心臓の病気になりそうか予想する
- 似通った特徴を持つ人のグループを形成する

ための比較的簡単な規則を見付け出すことを考える．これらの目的は，変数を組み合わせて新たな量を作ったり，あるいはその量を比較したりすることによって達成されるが，それぞれ多変量解析における

- 回帰分析 (単回帰)
- 回帰分析 (重回帰)
- 主成分分析
- 判別分析

- クラスタ分析

という手法に対応する．

ただし，これらの手法を適用する場合，心臓病の有無といったカテゴリラベルである質的な変数を計算上どのように取り扱うかについては注意を要する．前述したダミー変数を用いた数量化という手法を用いると，質的変数をあたかも量的変数のように取り扱うことができるので，講義では量的変数を対象として上記の手法を扱うことにする．なお，質的変数において回帰分析に相当する方法を数量化 I 類，判別分析に相当する方法を数量化 II 類，主成分分析に相当する方法を数量化 III 類，多次元尺度構成法に相当する方法を数量化 IV 類という．

また，時間に余裕があれば確率変数の系列を扱うための時系列解析の基本的な考え方と，経済時系列などの実データを扱うための方法論であるテクニカル分析を取り上げる．

1.3.1 回帰分析

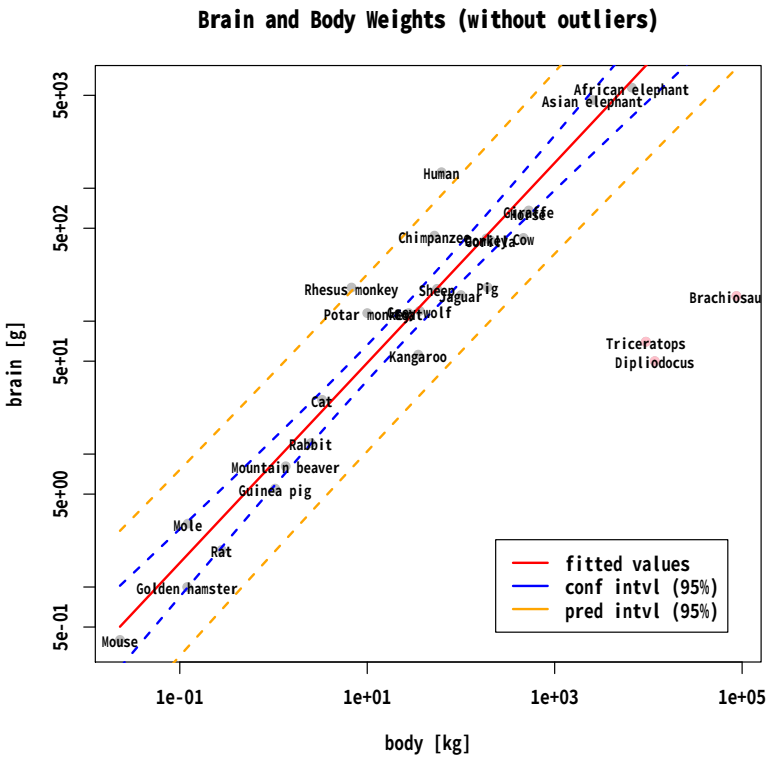


図 1.1: 実データによる回帰分析の例: 28 種の陸上動物の体重と脳重量の関係．特殊なデータ (外れ値) を除くことで，良好な回帰結果が得られている．(データは R の MASS パッケージの “Animals” を用いた．)

多変数の一部を説明変数，残りを目的変数とし，説明変数を入力とし目的変数を出力する回帰式を構成する．また，その回帰式の妥当性について議論する．

例えば

- 身長から体重は予測できるか？
  (**回帰式の推定**)
- ある身長の人がある体重だったとき，それは普通か？
  (**外れ値の判定**)
- 身長から体重を予測する式はどのくらい妥当か？
  (**寄与率の計算**)

といったことが解析の目的となる．

1.3.2 主成分分析

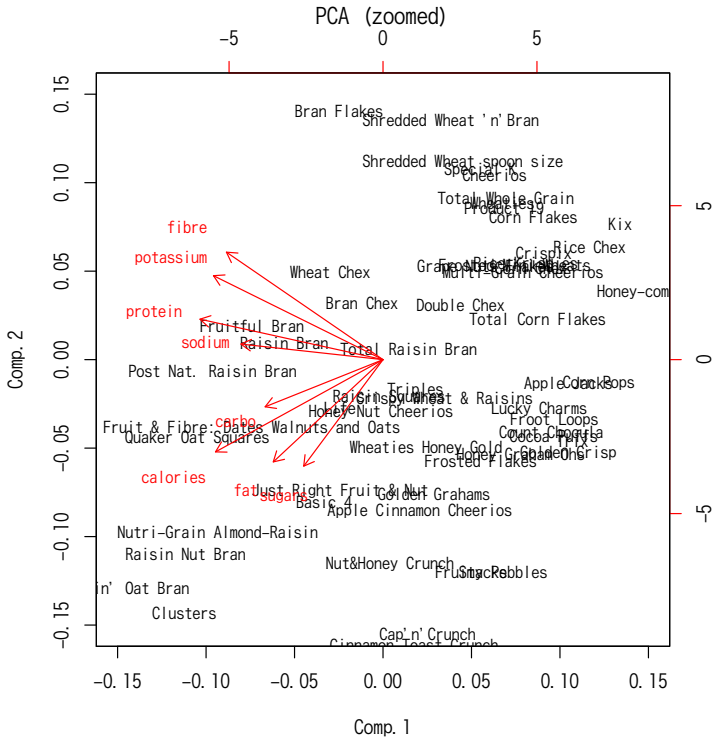


図 1.2: 実データによる主成分分析の例: 1993 年にアメリカにおいて販売されていた 65 種のシリアルができるだけ区別できるように 2 つの主成分方向に射影した図．含有成分が異なる様々なものが販売され，それぞれがどういった特徴を持つ商品であるかがわかる．(データは R の MASS パッケージの “UScereal” を用いた．)

多変数を低次元の部分空間に射影して近似表現するとき，最も情報を保持することができる部分空間を求める．また，適切な低次元部分空間の次数を議論する．

例えば

- 身長・体重から人の大きさを表す 1 つの量は作れないだろうか？
  (**主成分方向の決定**)

- スポーツテスト (50m 走，走り幅跳び，反復横跳び，背筋力など複数項目) の結果から，運動能力を表すための少数の指標は作れないだろうか？
  (**次元縮約**)
- 縮約された部分空間には元の情報はどのように保持され，どのような意味が与えられるのだろうか？
  (**主成分負荷の解釈**)
- 部分空間による近似の良さ，あるいは失われた情報はどのようにして評価すればよいだろうか？
  (**寄与率の計算**)

といったことが解析の目的となる．

1.3.3 判別分析

いくつかの量的変数とカテゴリ (クラス) を表す 1 つの質的変数からなるデータに対して，量的変数からカテゴリを予想する判別式を構成する．また，その判別式の妥当性について議論する．
例えば

- 心臓病を患っている人とそうでない人の身長，体重，血圧から，両者のグループの境界を求めることができるだろうか？
  (*1* **次判別式**，*2* **次判別式**)
- ある人の身長，体重，血圧が判ったとき，その人が将来心臓病になるかどうか予想できるだろうか？
  (**判別式による予測**)
- 判別の精度，あるいは判別の間違いによる危険度はどのようにして評価すればよいだろうか？
  (**誤判別率の検証**)

といったことが解析の目的となる．

1.3.4 クラスタ分析と多次元尺度構成法

3 次元を越える高次元データは，そのままでは図として表現することができず，データ間の関係を視覚的に捉えることは難しい．また，距離や類似度などデータ間の関係は規定されているが，座標系が与えられていないデータも同様に視覚的に捉えることは難しい．このような，そのままでは見ることのできない多次元のデータ間の関係を，樹形図 (クラスタ分析) や 2 次元地図 (多次元尺度構成法) の形で視覚化して解析の助けとしようというのが，クラスタ分析と多次元尺度構成法である．
例えば

- 含まれている成分によって類似度が定義された食品をどのように分類すればよいだろうか？
  - 類似のものから順次統合していき，統合の様子を樹形図で表現した系統樹を描くことによってグループ化を行う
    (**クラスタ分析**)

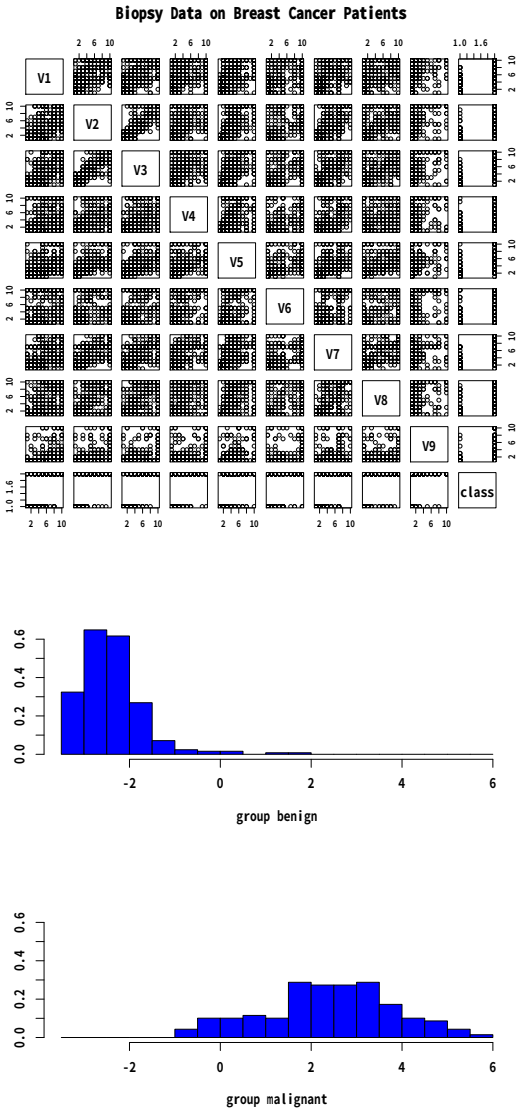


図 1.3: 実データによる判別分析の例: 9 種類の生検結果と乳癌の有無を調査したデータ (上図) に基づき，乳癌の有無を判定するための判別式を生検結果の線形式で構成したもの (下図)．判別式の値の分布を見ると，比較的良好な判別ルールが獲得されていることがわかる．(データは R の MASS パッケージの “biopsy” を用いた．)

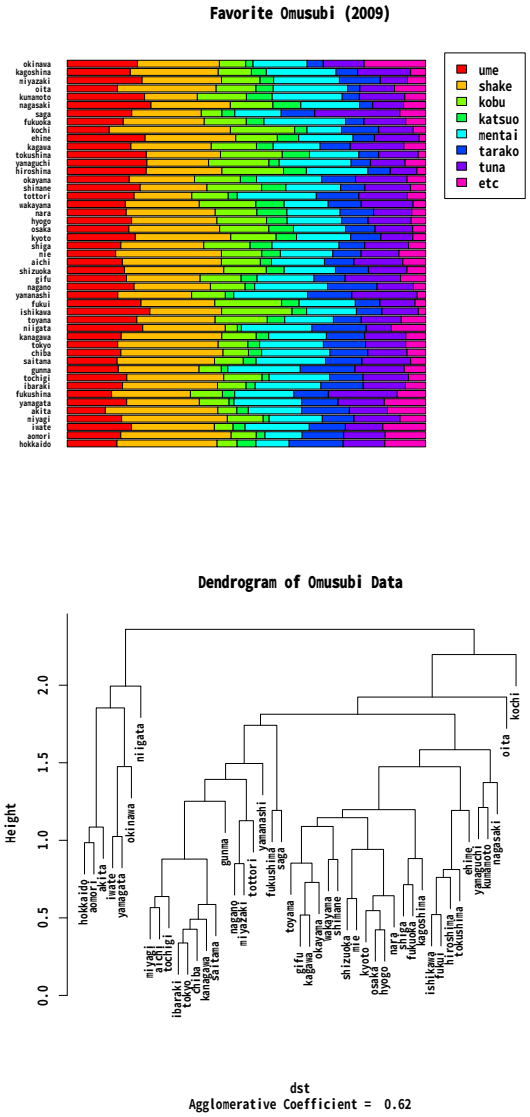


図 1.4: 実データによるクラスタ分析の例: 2009 年に行われた「好きなおむすびの具に関するアンケート」を用い各県での好きな具の分布 (上図) を算出し，分布の類似性に基づき各県の好みの近さを表す樹形図を求めたもの (下図)．解析においては地理的な関係を用いていないにも関らず，近傍には地域性が現れていることがわかる．(データはインターネットで公開されているものを用いた．)

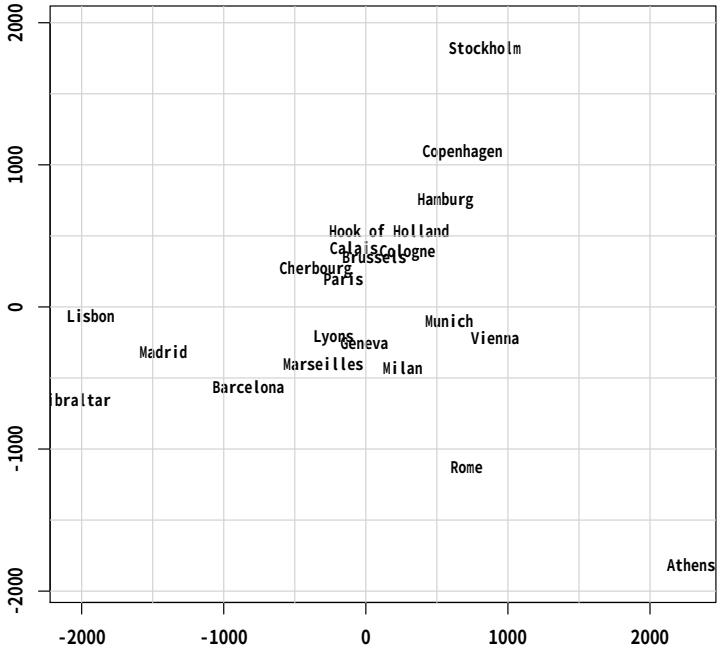


図 1.5: 実データによる多次元尺度構成法の例: ヨーロッパの主要都市間を結ぶ道路の長さから，多次元尺度構成法を用いて都市の物理的な配置を再構成したもの．道路は直線ではないが，都市間の直線距離を十分良く反映しているため，ほぼ正確な位置関係が再現されていることがわかる．(データは標準で R に含まれている “eurodist” を用いた．)

– 類似度をうまく表現する適当な座標軸を作り，近隣関係をできるだけ保持するように個々の食品を配置 (布置) する
(**多次元尺度構成法**)

といったことが解析の目的となる．

1.3.5 時系列解析

時系列とは時間とともに変化する確率変数の系列を表すが，変数同士の間に時間を跨いでの関係があるため，特別な取り扱いが必要な場合がある．講義では，時系列の正規性，独立性，定常性といった確率的な特徴を検定する方法や，自己回帰モデル (AR モデル) や移動平均モデル (MA モデル) といった基本的な時系列モデルについて説明する．

1.4 確率の復習

1.4.1 確率分布

確率分布とは，注目する事象 (集合) に対して，それが起きる確率 (区間 $[0,1]$ の実数) を返す関数

$$P(\text{事象}) = \text{確率値}$$

である．

実用上重要なのは離散分布か，あるいは**確率密度関数**を持つ絶対連続な分布である．絶対連続な分布においては，事象 A(見本空間の部分集合) が起きる確率は確率分布 P の確率密度関数 p の積分

$$P(A) = \int_A p(x)dx$$

で表される．特に，事象 A が十分小さな集合のときには，A に含まれる適当な点を x とし，A の大きさ (考える空間により体積や面積に相当) を $|A|$ と書くことにすれば，事象 A の起きる確率を

$$P(A) = p(x) \cdot |A|$$

で近似することができる．

1.4.2 正規分布

平均 μ，分散 σ^2 となる 1 次元正規分布の密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

である．特に $\mu = 0, \sigma = 1$ のとき**標準正規分布**という．

平均 $\boldsymbol{\mu}$，分散共分散行列 Σ となる p 次元正規分布の密度関数は

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T} \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

である．

正規分布の特徴は

- さまざまな誤差の集積は正規分布 (中心極限定理) となる
- 同じ分散 (ばらつき) を持つ分布の中で最も情報量 (エントロピー) が大きい

ことである．このため，分布に関する知識がないときには正規分布だと考えておくと安全なことが多い．また多変量解析の手法は，誤差の分布に関して正規性を仮定して導かれていることが多い．

1.4.3 正規分布から派生する分布

正規分布に従う独立な確率変数から計算される標本平均や不偏分散は，特別な分布に従うことが知られている．正規分布から派生する分布として重要なものには χ^2-分布，t-分布，F-分布がある．

χ^2-分布

$X_1, X_2, \ldots, X_d$ を標準正規分布 $\mathcal{N}(0,1)$ に従う独立な確率変数とする．このとき

$$Z = X_1^2 + X_2^2 + \cdots + X_d^2$$

の従う分布を自由度 d の χ^2-分布 ($\chi^2(d)$ と書く) と呼ぶ．自由度 d の χ^2-分布の密度関数は

$$f(x) = \frac{1}{\sqrt{2}^d \Gamma(d/2)} x^{d/2-1} e^{-x/2}$$

ただし，Γ はガンマ関数

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

である．χ^2-分布が現れる重要な例は，不偏分散である．$X_1, X_2, \ldots, X_n$ を正規分布 $\mathcal{N}(\mu, \sigma^2)$ に従う独立な確率変数とする．標本平均

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

からの差の平方和の分散に対する比

$$S = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$$

を考えると，S は自由度 $d = n-1$ の χ^2-分布に従うこと

$$S \sim \chi^2(n-1)$$

が示される．なお，平均の推定のために標本平均を使っているので，偏差の平方和の自由度は (標本数 -1) となることに注意する．この結果を用いると，不偏分散

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

の分布は χ^2-分布を拡大縮小した分布であり，その平均値は真の分散 σ^2 となり，不偏性を持つことがわかる．

t-分布

X_1 を標準正規分布 $\mathcal{N}(0,1)$，X_2 を自由度 d の χ^2-分布に従う独立な確率変数とする．このとき

$$Z = \frac{X_1}{\sqrt{X_2/d}}$$

の従う分布を自由度 d の t-分布 ($\mathcal{T}(d)$ と書く) と呼ぶ．自由度 d の t-分布の密度関数は

$$f(x) = \frac{\Gamma((d+1)/2)}{\sqrt{d\pi}\Gamma(d/2)} (1 + x^2/d)^{-(d+1)/2}$$

である. t-分布が現われる重要な例は不偏分散で正規化した標本平均の真の平均からの偏差である. $X_1, X_2, \ldots, X_n$ を正規分布 $\mathcal{N}(\mu, \sigma^2)$ に従う独立な確率変数とする. 標本平均

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

と，不偏分散

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

を用いて，変数

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

を考えると，T は自由度 $d = n - 1$ の t-分布に従うこと

$$T \sim \mathcal{T}(n-1)$$

が示される.

F-分布

X_1 を自由度 d_1 の χ^2-分布 X_2 を自由度 d_2 の χ^2-分布に従う独立な確率変数とする. このとき

$$Z = \frac{X_1/d_1}{X_2/d_2}$$

の従う分布を自由度 d_1, d_2 の F-分布 ($\mathcal{F}(d_1, d_2)$ と書く) と呼ぶ. 自由度 d_1, d_2 の F-分布の密度関数は

$$f(x) = \frac{1}{B(d_1/2, d_2/2)}\left(\frac{d_1 x}{d_1 x + d_2}\right)^{d_1/2}\left(1 - \frac{d_1 x}{d_1 x + d_2}\right)^{d_2/2} x^{-1}$$

ただし，B はベータ関数

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$$

である. F-分布が現われる重要な例は，2 つの標本分散の比の分布である. $X_1, X_2, \ldots, X_m$ は正規分布 $\mathcal{N}(\mu, \sigma^2)$ に，$Y_1, Y_2, \ldots, Y_n$ は正規分布 $\mathcal{N}(\nu, \sigma^2)$ に従う独立な確率変数とする. 2 つは平均は異なるが同じ分散を持つことに注意する. このとき，それぞれの不偏分散の比

$$F = \frac{\sum_{i=1}^{m}(X_i - \bar{X})^2}{m-1} \Big/ \frac{\sum_{j=1}^{n}(Y_j - \bar{Y})^2}{n-1}$$

を考えると，F は自由度 $d_1 = m - 1, d_2 = n - 1$ の F-分布に従うこと

$$F \sim \mathcal{F}(m-1, n-1)$$

が示される．これを用いて 2 つの母集団が同じ分散をもつかどうか調べることができる．

χ^2-分布は主に分散の区間推定や検定に用いられる．これ以外に適合度検定や独立性の検定など様々な場面でも用いられる．t-分布は，特に推定における信頼区間を構成するときに威力を発揮する．F-分布は，上記のような単純な分散の比較以外に，残差の分散を比較して統計モデルの良否を判断する分散分析と呼ばれる分野などで活躍する．

1.4.4 最尤法

母数 (パラメタ) で記述された確率分布の母数を推定する方法として良く用いられるのが**最尤法**である．確率分布のモデルが与えられると，観測データの起こる確率 (密度) を考えることができるが，最尤法は観測データが最も起こりやすい母数を推定値として採用する方法である．このとき，観測データの起こる確率を母数の関数と考えたものを**尤度関数**と呼ぶ．密度関数が滑らかで，データが十分に多いときには最も推定精度の良い方法の 1 つとして知られている．

1.4.5 統計量

基本的な統計量として平均と分散を考えることが多いが，**平均**は

$$\mathbb{E}[X] = \int_{\Omega} xp(x)dx, \quad (\text{ただし } \Omega \text{ は見本空間全体を表す})$$

で，**分散**は平均を用いて

$$V(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

で定義される．

実際のデータを扱う場合には分布が与えられる訳ではないので，データを使って平均や分散の近似値を求める必要がある．

まず，n 個の 1 次元データ $x_1, x_2, \ldots, x_n$ が与えられた場合を考えよう．**標本平均**は

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

標本分散は

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

で定義される．なお標本平均からの差の平方和を S_x と書くことがある．

標本分散の期待値は分散と一致しないので，これを修正した**不偏分散**

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

が用いられることも多い．

標準偏差 (standard deviation) は

$$s_x = \sqrt{V_x} \qquad (\text{多くの場合は不偏分散を使う})$$

で定義される．

データを処理する前に**標準化**，あるいは**正規化**と呼ばれる操作が行われることがある．これはデータを平均 0，分散 1 に変数変換 (1 次式) することで，

$$y_i = \frac{x_i - \bar{x}}{s_x}$$

で与えられる．

次に，n 個の 2 次元データ $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ が与えられた場合を考える．

標本平均は成分毎に 1 次元と同様に計算すればよく，**標本共分散**は

$$C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}S_{xy},$$

不偏共分散は

$$C_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}S_{xy}$$

で与えられる．

また，2 つの変数間の関係を示す量として**相関係数**

$$r_{xy} = \frac{C_{xy}}{\sqrt{V_x V_y}} = \frac{S_{xy}}{\sqrt{S_x S_y}}$$

が用いられることもある．これは標準化したデータの共分散であり，その大きさは

$$-1 \leq r_{xy} \leq 1$$

を満たす (Schwarz の不等式から容易に証明できる)．相関係数の正負に関しては

- 正になるとき: x が増えると y も増える傾向がある
- 負になるとき: x が増えると y は減る傾向がある

ことが言える．

1.4.6 ベイズの定理

よく用いられる**ベイズの公式**は

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}$$

であり，条件付き確率分布 $P(X|Y)$ から逆向きの条件付き確率分布 $P(Y|X)$ を求める公式であるが，これ以外にもさまざまな表現がある．

例 1.1. A 先生は大の野球ファンで，球団 H の勝敗で翌日の機嫌が左右されるとしよう．よくよく調べた結果

- 球団 H が勝つと 90%の確率で機嫌が良い
- 球団 H が負けると 70%の確率で機嫌が悪い

が成り立っているとする．一方，球団 H の勝率は現在のところ

- 球団 H は 60%の確率で勝つ
- 球団 H は 40%の確率で負ける

となっているとする．
このとき以下の確率を求めなさい．

- A 先生が機嫌が良いときに球団 H が勝った確率は？
- A 先生が機嫌が悪いときに球団 H が負けた確率は？

1.5 ベクトルと行列による微分

1.5.1 ベクトルによる微分

d 次元ベクトル $\boldsymbol{a} = (a_1, a_2, \ldots, a_d)^\mathsf{T}$ と $\boldsymbol{b} = (b_1, b_2, \ldots, b_d)^\mathsf{T}$ を考える．
ベクトル $\boldsymbol{a}$ の関数 $f(\boldsymbol{a})$ の微分を

$$\frac{\partial f}{\partial \boldsymbol{a}} = \begin{pmatrix} \frac{\partial f}{\partial a_1} \\ \vdots \\ \frac{\partial f}{\partial a_d} \end{pmatrix} \begin{matrix} \leftarrow 1 \text{ 次元} \\ \\ \leftarrow d \text{ 次元} \end{matrix}$$

と書くことにする．

例 1.2. $f(\boldsymbol{a}) = \boldsymbol{b}^\mathsf{T}\boldsymbol{a} = \boldsymbol{a}^\mathsf{T}\boldsymbol{b}$ のとき，$\frac{\partial f}{\partial \boldsymbol{a}}$ を求める．まず成分で考える．

$$\frac{\partial f}{\partial a_i} = \frac{\partial}{\partial a_i}(a_1 b_1 + \cdots + a_i b_i + \cdots + a_d b_d) = b_i.$$

したがって

$$\frac{\partial f}{\partial \boldsymbol{a}} = \begin{pmatrix} b_1 \\ \vdots \\ b_i \\ \vdots \\ b_d \end{pmatrix} = \boldsymbol{b}$$

となり，

$$\left.\begin{aligned} \frac{\partial}{\partial \boldsymbol{a}}(\boldsymbol{a}^\mathsf{T}\boldsymbol{b}) &= \boldsymbol{b} \\ \frac{\partial}{\partial \boldsymbol{a}}(\boldsymbol{b}^\mathsf{T}\boldsymbol{a}) &= (\boldsymbol{b}^\mathsf{T})^\mathsf{T} = \boldsymbol{b} \end{aligned}\right\}$$

というルールがあることがわかる．

例 1.3. $d \times d$ 行列 A を用いて定義される関数 $f(\boldsymbol{a}) = \boldsymbol{a}^\mathsf{T} A \boldsymbol{a}$ の微分 $\frac{\partial f}{\partial \boldsymbol{a}}$ を求めよ．

1.5.2 行列による微分

$d \times d$ 行列 A を

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1d} \\ a_{21} & a_{22} & \dots & a_{2d} \\ \dots & \dots & \dots & \dots \\ a_{d1} & a_{d2} & \dots & a_{dd} \end{pmatrix}$$

とし，A の関数 $f(A)$ の微分を

$$\frac{\partial f}{\partial A} = \begin{pmatrix} \frac{\partial f}{\partial a_{11}} & \frac{\partial f}{\partial a_{12}} & \dots & \frac{\partial f}{\partial a_{1d}} \\ \frac{\partial f}{\partial a_{21}} & \frac{\partial f}{\partial a_{22}} & \dots & \frac{\partial f}{\partial a_{2d}} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f}{\partial a_{d1}} & \frac{\partial f}{\partial a_{d2}} & \dots & \frac{\partial f}{\partial a_{dd}} \end{pmatrix}$$

(d 次元 × d 次元)

と書くことにする．

例 1.4. d 次元ベクトル $\boldsymbol{b}$ を用いて定義される関数

$$f(A) = \boldsymbol{b}^{\mathsf{T}} A \boldsymbol{b} = \sum_{i,j=1}^{d} b_i a_{ij} b_j$$

の微分を考える．成分で考えると

$$\frac{\partial f}{\partial a_{ij}} = \frac{\partial}{\partial a_{ij}} \sum_{i,j=1}^{d} b_i a_{ij} b_j = b_i b_j$$

となるので,

$$\frac{\partial}{\partial A} \boldsymbol{b}^{\mathsf{T}} A \boldsymbol{b} = \begin{pmatrix} b_1 b_1 & b_1 b_2 & \dots & b_1 b_d \\ b_2 b_1 & b_2 b_2 & \dots & b_2 b_d \\ \dots & \dots & \dots & \dots \\ b_d b_1 & b_d b_2 & \dots & b_d b_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \dots & b_d \end{pmatrix} = \boldsymbol{b}\boldsymbol{b}^{\mathsf{T}}.$$

例 1.5. $d \times d$ 行列 B を用いて定義される関数

$$f(A) = \operatorname{tr} AB = \sum_{i,j=1}^{d} a_{ij} b_{ji}$$

の微分を考える．成分では

$$\frac{\partial f}{\partial a_{ij}} = b_{ji}$$

となるので,

$$\frac{\partial}{\partial A} \operatorname{tr} AB = \begin{pmatrix} b_{11} & b_{21} & \dots & b_{d1} \\ b_{12} & b_{22} & \dots & b_{d2} \\ \dots & \dots & \dots & \dots \\ b_{1d} & b_{2d} & \dots & b_{dd} \end{pmatrix} = B^{\mathsf{T}}.$$

例 1.6. 行列のトレースに関して

$$\begin{aligned}&\operatorname{tr} AB = \operatorname{tr}(AB)^{\mathsf{T}} = \operatorname{tr} B^{\mathsf{T}}A^{\mathsf{T}},\\ &\operatorname{tr} AB = \operatorname{tr} BA,\\ &\operatorname{tr} A^{\mathsf{T}}B^{\mathsf{T}} = \operatorname{tr} B^{\mathsf{T}}A^{\mathsf{T}},\end{aligned}$$

が成り立つことを確かめよ.

上の行列のトレースの性質から

$$\begin{aligned}\frac{\partial}{\partial A}\operatorname{tr} AB &= B^{\mathsf{T}}\\ \frac{\partial}{\partial A}\operatorname{tr} BA &= B^{\mathsf{T}}\\ \frac{\partial}{\partial A}\operatorname{tr} A^{\mathsf{T}}B^{\mathsf{T}} &= B^{\mathsf{T}}\\ \frac{\partial}{\partial A}\operatorname{tr} B^{\mathsf{T}}A^{\mathsf{T}} &= B^{\mathsf{T}}\end{aligned}$$

ことが容易に確かめられる.
また

$$\boldsymbol{b}^{\mathsf{T}}A\boldsymbol{b} = \operatorname{tr}\boldsymbol{b}^{\mathsf{T}}A\boldsymbol{b} = \operatorname{tr} A\boldsymbol{b}\boldsymbol{b}^{\mathsf{T}}$$

となることから

$$\frac{\partial}{\partial A}\boldsymbol{b}^{\mathsf{T}}A\boldsymbol{b} = \frac{\partial}{\partial A}\operatorname{tr} A\boldsymbol{b}\boldsymbol{b}^{\mathsf{T}} = \left(\boldsymbol{b}\boldsymbol{b}^{\mathsf{T}}\right)^{\mathsf{T}} = \boldsymbol{b}\boldsymbol{b}^{\mathsf{T}}$$

となる，例の計算結果が矛盾しないことが確かめられる.

例 1.7. 行列 A の行列式を $|A|$ と書くとき

$$\frac{\partial |A|}{\partial A}$$

を求めよ. (ヒント: 余因子行列を上手く使うと簡単に計算できる.)

2 回帰分析

2.1 目的と考え方

2.1.1 目的

回帰分析 (regression analysis) とは, **説明変数** (explanatory variable) によって**目的変数** (response variable) を予測するための関係式

$$(\text{目的変数}) = f(\text{説明変数})$$

を構成し，変数間の関係を明らかにするための分析法である．説明変数，目的変数は，それぞれ**独立変数** (independent variable), **従属変数** (dependent variable) と呼ばれることもある．

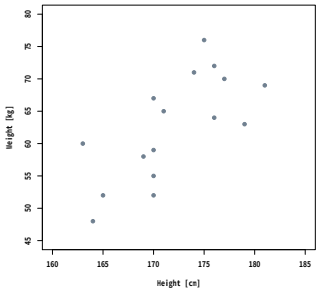
例 2.1. 身長から体重を推測するためには，データに基づいて

$$(\text{体重}) = f(\text{身長})$$

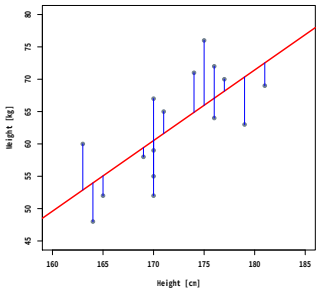
という関係式を構成する必要がある．一般の関数を考えることもできるが，1 次式 (説明変数の線形結合) を用いることが多い．

身長と体重のデータ

	身長 [cm]	体重 [kg]
1	164	48
2	170	55
3	171	65
4	174	71
5	176	64
6	181	69
7	170	52
8	176	72
9	170	59
10	165	52
11	169	58
12	170	67
13	179	63
14	163	60
15	177	70
16	175	76



身長と体重の散布図



回帰直線 (赤) と残差 (青)

2.1.2 観測データ

回帰分析では，説明変数と目的変数の組を 1 つの観測値 (データ点，標本; datum, sample) と考え，n 個の観測値からなるデータ集合 (標本集合; data set, sample set) を対象とした分析を行う．このとき，i 番目のデータを $(\boldsymbol{x}_i, y_i)$, $i = 1, 2, \ldots, n$ で表すことにする．

説明変数 一般に幾つあってもよいが，1 つかそれ以外で区別することがある．

1 次元の場合 単回帰 (simple regression)

多次元の場合 重回帰 (multiple regression)

目的変数 幾つあってもよいが，各変数毎に分析すれば良いので，多くの場合 1 次元を考えれば十分である．

以下では一般の場合を扱うために説明変数は p 次元とし，ベクトルとスカラを区別するためにベクトルは

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathsf{T}}, \qquad (\mathsf{T} \text{ は転置を表す})$$

のように太字 (bold face) で表す．なお，以下では特に断わらない限りベクトルは列 (縦) ベクトルとして扱う．

2.1.3 確率モデル

説明変数から目的変数がどのように生成されたかを表す生成モデルとして，回帰式と加法的な誤差を考慮した以下の**線形回帰** (linear regression) モデルを考える．

$$\underbrace{y_i}_{\text{目的変数}} = \underbrace{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}_{\text{回帰式}} + \underbrace{\epsilon_i}_{\text{誤差}}, \quad i = 1, \ldots, n$$

回帰式 (regression function) ここでは説明変数の線形結合を考える．

誤差 (error) 回帰式では説明しきれないデータの不確定性を表す．

線形でない場合を一般に非線形回帰と呼ぶ．多数のパラメタを持つ柔軟な非線形モデルによる非線形回帰を用いれば，説明変数と目的変数の間の複雑な関係を表すことができるが，一方でパラメタを決定するための計算が繁雑になる，あるいは精度の良いパラメタ推定を行うためには大量のデータが必要となるなど新たな問題点が生じることにも注意が必要である．なお，例えば x_{ik}^2 や $x_{ik} \times x_{ih}$ といった 2 次式や，$\log x_{ik}$ などの非線形関数で変換した説明変数を新たな説明変数としてモデルに加えることにすれば，線形回帰モデルでも説明変数と目的変数の間のある程度の非線形関係を表すことができる．

2.2 計算法

2.2.1 最適性の指針

目的変数の実測値と回帰式による予測値を比較して回帰式を決定するためには，以下の 2 つの方策が考えられる:

最小二乗法 (least squares) 残差の平方和を最小化する回帰式を選択する，

最尤法 (maximum likelihood) 残差の確率分布を仮定して尤度が最大となる回帰式を選択する．

誤差の分布が，説明変数によらず同一の正規分布に従うとすれば，この二つの基準は等価となることが以下のように示される．

まず，回帰式のパラメタ (母数) を纏めて $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathsf{T}}$ とベクトルで書くことにする．観測値と回帰式の予測値 (あてはめ値) の残差は

$$e_i(\boldsymbol{\beta}) = y_i - \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \right), \quad i = 1, \ldots, n$$

であり，パラメタ $\boldsymbol{\beta}$ の関数となる．

さて，誤差が説明変数には依存せず，平均 0，分散 σ^2 の正規分布に従うと仮定する．その確率密度は

$$p(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\epsilon_i^2/2\sigma^2}$$

で表され，誤差 ϵ が微小な区間 $[z, z+\delta]$ の間に発生する確率は

$$\Pr(z < \epsilon < z+\delta) = \int_z^{z+\delta} p(\epsilon) d\epsilon \simeq p(z)\delta$$

となる．n 個の観測データが独立に得られているとき，誤差 $\epsilon_i,\ i = 1, \ldots, n$ がそれぞれ区間 $[z_i, z_i+\delta_i]$ に含まれる同時確率は

$$\Pr(z_1 < \epsilon_1 < z_1+\delta_1, \ldots, z_n < \epsilon_n < z_n+\delta_n) = \prod_{i=1}^{n} p(z_i)\delta_i$$

である．したがって観測データの誤差が $\epsilon_i,\ i = 1, \ldots, n$ で表されるとき，観測データの生成される確率は密度関数の積 $\prod_{i=1}^{n} p(\epsilon_i)$ に比例すると考えてよい．

ところでパラメタ $\boldsymbol{\beta}$ が真の値となるとき，残差 $e_i(\boldsymbol{\beta})$ と誤差 ϵ_i は一致するので，残差の起こりやすさを上記の誤差の確率密度で評価した

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(e_i(\boldsymbol{\beta}))$$

は $\boldsymbol{\beta}$ が真の値に近いほど大きくなることが期待される．関数 $l(\boldsymbol{\beta})$ を，パラメタ $\boldsymbol{\beta}$ の**尤度関数** (likelihood function) と呼ぶ．尤度関数は，パラメタ $\boldsymbol{\beta}$ において観測値 $(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, n$ が出現する確率と考えることができ，その値を最大とするパラメタを持つモデルは観測データを生成する尤もらしい (最も妥当な) モデルであると考えられる．なお数値計算上，多数の関数の積を扱うことは精度の問題などでも好ましくないため，単調関数である対数関数で尤度関数を変換したもの

$$L(\boldsymbol{\beta}) = \log l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log p(e_i(\boldsymbol{\beta}))$$

を考えるのが一般的である．これを**対数尤度関数** (log likelihood function) と呼ぶ．

さて，ここで**残差平方和** (sum of squared errors, sum of squared residuals) を

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i(\boldsymbol{\beta})^2 = \sum_{i=1}^{n} \Big\{ y_i - (\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}) \Big\}^2$$

と書くことにする．これを用いると対数尤度関数は

$$\begin{aligned} L(\boldsymbol{\beta}) = \log l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log p(e(\boldsymbol{\beta})_i) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} e_i(\boldsymbol{\beta})^2 + (\boldsymbol{\beta} \text{ と無関係な項}) \\ &= -\frac{1}{2\sigma^2} S(\boldsymbol{\beta}) + (\boldsymbol{\beta} \text{ と無関係な項}) \end{aligned}$$

と纏めることができる．この式から，対数尤度関数 $L(\boldsymbol{\beta})$ の最大化 (最尤法) と残差平方和 $S(\boldsymbol{\beta})$ の最小化 (最小二乗法) は等価となり，最適なパラメタ $\hat{\boldsymbol{\beta}}$ は

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$$

で与えられる．逆に誤差に関する仮定が成立しなならば，2 つの方策は一般に異なる解を与える．

2.2.2 最小二乗推定量

残差平方和 S の最小値を与えるパラメタ $\boldsymbol{\beta}$ の条件 (必要条件) は

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = \left(\frac{\partial S}{\partial \beta_0}, \frac{\partial S}{\partial \beta_1}, \ldots, \frac{\partial S}{\partial \beta_p} \right)^{\mathsf{T}} = \mathbf{0}$$

で与えられる．ただし $\mathbf{0}$ は $p+1$ 次元の零ベクトルである．求める回帰式の最適な係数 $\boldsymbol{\beta}$ は上記の条件を満たすのもの中にあることがわかる (実は 1 つに決まる)．

データをまとめて並べた行列を定義し，残差平方和をデータの行列とパラメタのベクトルを用いて表すことを考える．

$$X = \overset{p+1 \text{ 次元}}{\begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}} \; n \text{ 次元} = \begin{pmatrix} 1 & \boldsymbol{x}_1^{\mathsf{T}} \\ 1 & \boldsymbol{x}_2^{\mathsf{T}} \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n^{\mathsf{T}} \end{pmatrix}$$

$$Y = \overset{1 \text{ 次元}}{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}} \; n \text{ 次元} = (y_1, y_2, \ldots, y_n)^{\mathsf{T}}$$

$$\boldsymbol{\beta} = \overset{1 \text{ 次元}}{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}} \; p+1 \text{ 次元} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathsf{T}}$$

行列 X は**計画行列** (design matrix) と呼ばれる．このとき残差は

$$\boldsymbol{e}(\boldsymbol{\beta}) = \begin{pmatrix} e_1(\boldsymbol{\beta}) \\ e_2(\boldsymbol{\beta}) \\ \vdots \\ e_n(\boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \beta_0 + \sum_{j=1}^{p} \beta_j x_{1j} \\ \beta_0 + \sum_{j=1}^{p} \beta_j x_{2j} \\ \vdots \\ \beta_0 + \sum_{j=1}^{p} \beta_j x_{nj} \end{pmatrix} = Y - X\boldsymbol{\beta}$$

となることから，残差平方和は

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} e_i(\boldsymbol{\beta})^2 = \boldsymbol{e}(\boldsymbol{\beta})^{\mathsf{T}} \boldsymbol{e}(\boldsymbol{\beta}) \\ &= (Y - X\boldsymbol{\beta})^{\mathsf{T}} (Y - X\boldsymbol{\beta}) \end{aligned}$$

で表される．

次に偏微分係数を纏めてベクトル表示するために，ベクトル $\boldsymbol{\beta}$ の関数 $f(\boldsymbol{\beta})$ いくつかの特別な場合について，その微分を求めておく．以下では

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = \left(\frac{\partial f}{\partial \beta_0}, \frac{\partial f}{\partial \beta_1}, \ldots, \frac{\partial f}{\partial \beta_p} \right)^{\mathsf{T}}$$

と書くことにする．さて $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)^{\mathsf{T}}$ に対して $f(\boldsymbol{\beta}) = \boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\alpha}$ のとき

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = \boldsymbol{\alpha}$$

となる．また，$(p+1) \times (p+1)$ 型行列 A に対して $f(\boldsymbol{\beta}) = \boldsymbol{\beta}^{\mathsf{T}} A \boldsymbol{\beta}$ のとき

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = A\boldsymbol{\beta} + (\boldsymbol{\beta}^{\mathsf{T}} A)^{\mathsf{T}} = (A + A^{\mathsf{T}})\boldsymbol{\beta}$$

となる．ただし，$A + A^{\mathsf{T}}$ は行列 A の対称な成分を取り出していることに注意する．

これらを用いると，残差平方和が最小となるパラメタ $\boldsymbol{\beta}$ の条件を簡単に表現することができる．残差平方和を展開すると

$$\begin{aligned} S(\boldsymbol{\beta}) &= (Y - X\boldsymbol{\beta})^{\mathsf{T}} (Y - X\boldsymbol{\beta}) \\ &= Y^{\mathsf{T}}Y - \boldsymbol{\beta}^{\mathsf{T}}X^{\mathsf{T}}Y - Y^{\mathsf{T}}X\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathsf{T}}X^{\mathsf{T}}X\boldsymbol{\beta} \end{aligned}$$

となるので，これを $\boldsymbol{\beta}$ で微分して

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2X^{\mathsf{T}}X\boldsymbol{\beta} - 2X^{\mathsf{T}}Y = 0$$

すなわち

$$X^{\mathsf{T}}X\boldsymbol{\beta} = X^{\mathsf{T}}Y$$

を満たす $\boldsymbol{\beta}$ を求めれば良い．この式を**正規方程式** (normal equation) という．ここで $(X^{\mathsf{T}}X)^{\mathsf{T}} = X^{\mathsf{T}}X$ より $X^{\mathsf{T}}X$ は対称行列であり，計画行列 X の Gram 行列 (Gram matrix; Gramian) という．正

規方程式は Gram 行列 $X^\mathsf{T}X$ が正則 (逆行列を持つ) であれば簡単に解けて最適な係数 $\hat{\boldsymbol{\beta}}$ は

$$\hat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$$

で与えられる．これを**最小二乗推定量** (least square estimator) という．

目的変数の予測値 (あてはめ値) を $\hat{y}_i$ と書くことにすれば，パラメタ $\boldsymbol{\beta}$ における予測値は

$$\hat{Y} = \overset{1\text{ 次元}}{\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}}_{n\text{ 次元}} = X\boldsymbol{\beta}$$

となる．特に最小二乗推定量を用いた予測値は

$$\hat{Y} = X\hat{\boldsymbol{\beta}} = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$$

であるので，予測値 $\hat{Y}$ は観測データの目的変数 Y の線形和で記述されることがわかる．

2.2.3 中心化を用いた表現

前節で求めた最適な線形回帰式

$$\begin{aligned} y &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \\ &= (1, \boldsymbol{x}^\mathsf{T})\hat{\boldsymbol{\beta}} \end{aligned}$$

は説明変数と目的変数の標本平均を通ることが以下のようにして確かめられる．まず計画行列 X の性質として

$$X^\mathsf{T}X(X^\mathsf{T}X)^{-1}X^\mathsf{T} = X^\mathsf{T}$$

となることから，左辺の一番左の行列 X^T と右辺の行列の第 1 行を比較して

$$\mathbb{1}^\mathsf{T}X(X^\mathsf{T}X)^{-1}X^\mathsf{T} = \mathbb{1}^\mathsf{T}$$

となることがわかる．ただし $\mathbb{1}$ は全成分が 1 である n 次元列ベクトルである．次に説明変数と目的変数の標本平均を以下で定義する．

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \qquad (\boldsymbol{x}\text{ の標本平均，}p\text{ 次元ベクトル})$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (y\text{ の標本平均，スカラ})$$

このとき

$$\begin{aligned} \mathbb{1}^\mathsf{T}X &= n(1, \bar{\boldsymbol{x}}^\mathsf{T}) \\ \mathbb{1}^\mathsf{T}Y &= n\bar{y} \end{aligned}$$

であることに注意すると

$$\begin{aligned}(1, \bar{\boldsymbol{x}}^\mathsf{T})\hat{\boldsymbol{\beta}} &= \frac{1}{n}\mathbb{1}^\mathsf{T} X\hat{\boldsymbol{\beta}} \\ &= \frac{1}{n}\mathbb{1}^\mathsf{T} X(X^\mathsf{T}X)^{-1}X^\mathsf{T}Y \\ &= \frac{1}{n}\mathbb{1}^\mathsf{T} Y = \bar{y}\end{aligned}$$

となる．したがって回帰式は標本平均 $(\bar{\boldsymbol{x}}, \bar{y})$ を通ることがわかる．このことより，回帰式を構成する際には，標本平均が中心となるようにデータ全体を平行移動した上で傾きだけ求めれば良いことが示唆される．このときモデルの本質的な自由度は p 次元となる．

上記の議論から線形回帰式は説明変数 $\boldsymbol{x}$ と目的変数 y の標本平均を通ることがわかったので，以降の計算の意味を理解するための別の表現を用意しておく．そのために説明変数および目的変数からそれぞれの標本平均を引いた以下の量で行列 X, Y を再定義しておく:

$$X = \begin{pmatrix} \boldsymbol{x}_1^\mathsf{T} - \bar{\boldsymbol{x}}^\mathsf{T} \\ \boldsymbol{x}_2^\mathsf{T} - \bar{\boldsymbol{x}}^\mathsf{T} \\ \vdots \\ \boldsymbol{x}_n^\mathsf{T} - \bar{\boldsymbol{x}}^\mathsf{T} \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ n \text{ 次元}, \end{matrix}$$

(p 次元)

$$Y = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ n \text{ 次元} \end{matrix} = (y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})^\mathsf{T}.$$

(1 次元)

この操作を**中心化** (centering) という．また $\boldsymbol{\beta}$ を定数項を除いた形

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \begin{matrix} \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ p \text{ 次元} \end{matrix} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$$

(1 次元)

で定義しなおしておく．さらに中心化した予測値 $\hat{y}_i - \bar{y}$ をまとめて

$$\hat{Y} = \begin{pmatrix} \hat{y}_1 - \bar{y} \\ \hat{y}_2 - \bar{y} \\ \vdots \\ \hat{y}_n - \bar{y} \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ n \text{ 次元} \end{matrix} = (\hat{y}_1 - \bar{y}, \hat{y}_2 - \bar{y}, \ldots, \hat{y}_n - \bar{y})^\mathsf{T}$$

(1 次元)

と書く．パラメタ $\boldsymbol{\beta}$ における中心化した予測値と説明変数の関係は

$$\hat{Y} = X\boldsymbol{\beta}$$

である．

ここで，X,Y は中心化されているので

$$\begin{aligned} X^{\mathsf{T}}X &= S_x && (\boldsymbol{x} \text{ の標本平均からの偏差の平方和}) \\ &= nV_x && (\boldsymbol{x} \text{ の標本分散共分散行列} \times n) \\ X^{\mathsf{T}}Y &= S_{xy} && (\boldsymbol{x} \text{ と } y \text{ の標本平均からの偏差の平方和}) \\ &= nV_{xy} && (\boldsymbol{x} \text{ と } y \text{ の標本共分散行列} \times n) \\ Y^{\mathsf{T}}Y &= S_y && (y \text{ の標本平均からの偏差の平方和}) \\ &= nV_y && (y \text{ の標本共分散行列} \times n) \end{aligned}$$

となり，分散共分散がそのまま計算できる表現になっていることに注意する．

練習問題 (2)

これらを用いると実測値と予測値の残差は

$$\boldsymbol{e}(\boldsymbol{\beta}) = \begin{pmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{pmatrix} = Y - \hat{Y} = Y - X\boldsymbol{\beta}$$

と書け，また残差平方和は

$$S(\boldsymbol{\beta}) = \boldsymbol{e}(\boldsymbol{\beta})^{\mathsf{T}}\boldsymbol{e}(\boldsymbol{\beta}) = (Y - X\boldsymbol{\beta})^{\mathsf{T}}(Y - X\boldsymbol{\beta})$$

と表されるので，偏微分を用いたこれまでの議論と同様に

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2X^{\mathsf{T}}X\boldsymbol{\beta} - 2X^{\mathsf{T}}Y = 0$$

すなわち

$$X^{\mathsf{T}}X\boldsymbol{\beta} = X^{\mathsf{T}}Y$$

が最適なパラメタの満たす正規方程式となる．これより，Gram 行列 $X^{\mathsf{T}}X$ が正則であれば最適パラメタは

$$\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y = V_x^{-1}V_{xy}$$

と書けるので，係数は説明変数の標本分散共分散行列 $V_{\boldsymbol{x}}$ と，説明変数と目的変数の標本共分散行列 $V_{\boldsymbol{x}y}$ のみで表されることがわかる．

2.3 分析の評価

2.3.1 残差の分解

目的変数が回帰式によってどのくらい説明されるようになったのかをみるために，目的変数の分散をいくつかの量に分解する．目的変数の平均まわりの平方和，すなわち偏差の平方和 (分散 $\times n$) は

$$\begin{aligned} S_y &= Y^{\mathsf{T}}Y \\ &= (Y - \hat{Y} + \hat{Y})^{\mathsf{T}}(Y - \hat{Y} + \hat{Y}) \\ &= (Y - \hat{Y})^{\mathsf{T}}(Y - \hat{Y}) + \hat{Y}^{\mathsf{T}}(Y - \hat{Y}) + (Y - \hat{Y})^{\mathsf{T}}\hat{Y} + \hat{Y}^{\mathsf{T}}\hat{Y} \end{aligned}$$

と分解されるが，予測値が

$$\begin{aligned}\hat{Y} &= X\hat{\boldsymbol{\beta}} \\ &= X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y\end{aligned} \tag{2.1}$$

と書けることから, 予測値の平均からの偏差の平方和は

$$\begin{aligned}\hat{Y}^{\mathsf{T}}\hat{Y} &= \hat{\boldsymbol{\beta}}^{\mathsf{T}}X^{\mathsf{T}}X\hat{\boldsymbol{\beta}} \\ &= Y^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1}(X^{\mathsf{T}}X)(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y \\ &= Y^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y \\ &= \hat{Y}^{\mathsf{T}}Y = Y^{\mathsf{T}}\hat{Y}\end{aligned}$$

となることに注意すると

$$\begin{aligned}&(\text{第 1 項}) = (Y - X\hat{\boldsymbol{\beta}})^{\mathsf{T}}(Y - X\hat{\boldsymbol{\beta}}) = S(\hat{\boldsymbol{\beta}}) \\ &(\text{第 2 項}) = \hat{Y}^{\mathsf{T}}Y - \hat{Y}^{\mathsf{T}}\hat{Y} = 0 \\ &(\text{第 3 項}) = 0 \qquad\qquad (\text{第 2 項と同様}) \\ &(\text{第 4 項}) = \hat{\boldsymbol{\beta}}^{\mathsf{T}}X^{\mathsf{T}}X\hat{\boldsymbol{\beta}} = S_r(\hat{\boldsymbol{\beta}})\end{aligned}$$

となる．結局，目的変数の偏差の平方和は

$$\begin{aligned}S_y &= S(\hat{\boldsymbol{\beta}}) + S_r(\hat{\boldsymbol{\beta}}) \\ &= (\text{回帰で説明できない残差}) + (\text{回帰で説明できる残差})\end{aligned}$$

と分解できることがわかる．分析の評価においては，この関係が利用される．なお，右辺を $\hat{\boldsymbol{\beta}}$ で評価することが明らかな場合は省略して S, S_r と書くこともある．

2.3.2 寄与率

残差平方和の分解より回帰式による説明力の 1 つの基準として

$$R^2 = \frac{S_r}{S_y} = \frac{Y^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y}{Y^{\mathsf{T}}Y}$$

$$= \frac{S_y - S}{S_y} \qquad \left(\frac{(\text{回帰で説明できる残差平方和})}{(\text{目的変数が本質的に持つ残差平方和})}\right)$$

$$= 1 - \frac{S}{S_y} = 1 - \frac{S/n}{S_y/n} \qquad \left(1 - \frac{(\text{残差の標本分散})}{(\text{目的変数の標本分散})}\right)$$

という量を考えることができる．これを**寄与率** (proportion of the variance) あるいは**決定係数** (coefficient of determination, R^2, R squared) という．

特に単回帰の場合は

$$R^2 = \frac{S_{xy}^2}{S_x S_y} = \frac{V_{xy}^2}{V_x V_y} = r_{xy}^2$$

となり，**相関係数** (correlation coefficient) r_{xy} の 2 乗となる．

このままの値を評価に使う場合もあるが，本来は

$$R^2 = 1 - \frac{(\text{残差の分散})}{(\text{目的変数の分散})}$$

を計算したいので，それぞれの分散の推定値を不偏分散で置き換えて

$$\bar{R}^2 = 1 - \frac{S/(n-p-1)}{S_y/(n-1)} \qquad \left(1 - \frac{(\text{残差の不偏分散})}{(\text{目的変数の不偏分散})}\right)$$

と定義することがある．これを**自由度調整済寄与率**あるいは**自由度調整済決定係数** (adjusted R^2, adjusted R squared, R bar squared) という．ただし，$n-p-1$ は残差の分散の計算をするときに必要な自由度である．これらの自由度は以下のようにして求めることができる．まず，目的変数の不偏分散は目的変数の偏差の平方和を $n-1$ (データ数 -1) で割ったもの ($n-1$ となっているのは平均の計算に 1 自由度使っているからと考えればよい) である．回帰の分散の自由度 (回帰式は標本平均を通ることがわかっているので，本質的には傾きだけ決めればよい) は回帰式の次数 p であるので，目的変数の分散の自由度は回帰の分散の自由度 (回帰の次数) と残差の分散の自由度 ϕ に分解され

$$n-1 = p + \phi$$

が成り立つ．したがって

$$\phi = n-p-1$$

となる．

単回帰のときには $p = 1$ なので，$\phi = n-2$ となり，n がある程度大きければ自由度の調整はあまり影響はないが，p の大きな重回帰のときには自由度の調整が重要となることがある．一般に次数の大きな回帰式ほど残差平方和は小さくなるため，次数の異なる回帰式を比較検討する場合には調整した寄与率で比較する必要がある．

2.3.3 ハット行列

式 (2.1) で見たように，予測値は観測値およびその標本平均の線形和で表される．ここでは観測値と予測値の関係を表す別の表現を求めておく．

まず，n 個の観測値，予測値および標本平均を並べたベクトルを以下で定義しておく．

$$\begin{aligned}
\boldsymbol{y} &= (y_1, y_2, \ldots, y_n)^\mathsf{T} \\
\hat{\boldsymbol{y}} &= (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^\mathsf{T} \\
\bar{\boldsymbol{y}} &= (\bar{y}, \bar{y}, \ldots, \bar{y})^\mathsf{T}
\end{aligned}$$

さて，中心化した行列に対して一般に $\mathbb{1}^\mathsf{T} X = 0$ または $X^\mathsf{T}\mathbb{1} = 0$ となることに注意する．また，行列 M を

$$M = \frac{1}{n}\mathbb{1}\mathbb{1}^\mathsf{T}$$

と定義すると，標本平均ベクトルは

$$\bar{\boldsymbol{y}} = \mathbb{1}\bar{y} = \frac{1}{n}\mathbb{1}\mathbb{1}^{\mathsf{T}}\boldsymbol{y} = M\boldsymbol{y}$$

と書くことができることに注意する．このとき，予測値ベクトルは

$$\begin{aligned}\hat{\boldsymbol{y}} &= \hat{\boldsymbol{y}} - \bar{\boldsymbol{y}} + \bar{\boldsymbol{y}} \\ &= X\hat{\boldsymbol{\beta}} + \bar{\boldsymbol{y}} = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}(\boldsymbol{y} - \bar{\boldsymbol{y}}) + \bar{\boldsymbol{y}} \\ &= X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\left\{\boldsymbol{y} - \frac{1}{n}\mathbb{1}\mathbb{1}^{\mathsf{T}}\boldsymbol{y}\right\} + \frac{1}{n}\mathbb{1}\mathbb{1}^{\mathsf{T}}\boldsymbol{y} \\ &= X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\boldsymbol{y} + M\boldsymbol{y}\end{aligned}$$

と書くことができる．ここで行列 H を

$$H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}} + M$$

で定義すると，観測値と予測値の関係は

$$\hat{\boldsymbol{y}} = H\boldsymbol{y}$$

と書くことができる．つまり目的変数の予測値 (推定値) を説明変数のみからなる行列 H と目的変数の実測値の積として表すことができる．この行列 H を**ハット行列** (hat matrix) という．行列 H にはいくつかの特徴的な性質がある．まず H は対称で

$$HX = X, \quad H\mathbb{1} = \mathbb{1}$$

が成り立つ．また，H は冪等 (idempotent)

練習問題 (4)

$$H^2 = H, \quad (I - H)^2 = I - H$$

となるので射影行列であることも確認できる．

練習問題 (5)

なお，ハット行列は中心化していない計画行列

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{x}_1^{\mathsf{T}} \\ 1 & \boldsymbol{x}_2^{\mathsf{T}} \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n^{\mathsf{T}} \end{pmatrix}$$

を用いて表すこともできる．この場合，標本平均を計算するためのベクトル $\mathbb{1}$ が計画行列の中に含まれるため，ハット行列は計画行列のみで

$$H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$$

と表わされる．

練習問題 (6)

2.3.4 残差と標準化誤差

残差 (residual) とは，観測値と回帰式から計算される予測値 (あてはめ値) の差

$$e_i(\boldsymbol{\beta}) = y_i - \left(\beta_0 + \sum_{k=1}^{p}\beta_k x_{ik}\right), \quad i = 1, \dots, n$$

であった．推定値 $\hat{\boldsymbol{\beta}}$ における残差

$$\hat{\epsilon}_i = e_i(\hat{\boldsymbol{\beta}}) = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

の和に関しては

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0$$

が成り立つ．これは節 2.3.3 の結果を用いると

$$\hat{\boldsymbol{\epsilon}} = Y - \hat{Y} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (I - H)\boldsymbol{y}$$

となることから，

$$\begin{aligned} \mathbb{1}^{\mathsf{T}}\hat{\boldsymbol{\epsilon}} &= \mathbb{1}^{\mathsf{T}}(I - H)\boldsymbol{y} \\ &= (\mathbb{1}^{\mathsf{T}} - \mathbb{1}^{\mathsf{T}})\boldsymbol{y} \end{aligned}$$

となり容易に確かめることができる．

さて，モデルが正しく

$$y_i = \bar{y} + (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n$$

が成り立っている，すなわち

$$\boldsymbol{y} = \bar{\boldsymbol{y}} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

であるとする．このとき，

$$\begin{aligned} \hat{\boldsymbol{\epsilon}} &= (I - H)\boldsymbol{y} \\ &= (I - H)\bar{\boldsymbol{y}} + (I - H)X\boldsymbol{\beta} + (I - H)\boldsymbol{\epsilon} \\ &= (I - H)\mathbb{1}\bar{y} + (X - X)\boldsymbol{\beta} + (I - H)\boldsymbol{\epsilon} \\ &= (I - H)\boldsymbol{\epsilon} \end{aligned}$$

となるので，残差は誤差の線形和で表されることがわかる．

さて，残差と誤差のこの関係を用いて，残差の統計的な性質を考えてみることにしよう．まず，容易に確認できるように残差の平均は

$$\mathbb{E}[\hat{\boldsymbol{\epsilon}}] = (I - H)\mathbb{E}[\boldsymbol{\epsilon}] = 0$$

である．残差の分散共分散は誤差 $\boldsymbol{\epsilon}$ が独立で平均 0，分散 σ^2 の正規分布に従うことから

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\epsilon}}) &= \mathbb{E}[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^{\mathsf{T}}] \\ &= (I - H)\ \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}]\ (I - H) \\ &= (I - H)\sigma^2 I(I - H) \\ &= \sigma^2(I - H) \end{aligned}$$

となる．各データごとに見ると，データ i の分散はハット行列の (i, i) 成分を h_{ii} と書くことにすれば

$$\mathrm{Var}(\hat{\epsilon}_i) = \mathbb{E}[\hat{\epsilon}^2] = (1 - h_{ii})\sigma^2$$

となる．一般に $h_{ii} > 0$ であるから，上記の計算より各データごとの残差は誤差の真の分散より平均的には小さく，過小評価されることがわかる．また，その評価の度合は説明変数の配置に依存して決まることがわかる．

これらの依存性を取り除くために，以下のように正規化した残差

$$e_i^s = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

を定義することができる．これを**標準化残差** (standard residual) という．ただし $\hat{\sigma}^2$ は σ^2 の推定値で，通常は不偏分散

$$\hat{\sigma}^2 = \frac{1}{n-p-1} S(\hat{\boldsymbol{\beta}}) = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

を用いて推定する．データがモデルに従っているのであれば，標準化残差は自由度 $n-p-1$ の t-分布に従うことが期待されるので，これを用いて残差の異常に大きい外れ値を検出することができる．

練習問題 (7)

一方，上記の定義による標準化では $\hat{\sigma}$ に $\hat{\epsilon}_i$ が含まれる．このとき e_i^s と $\hat{\sigma}$ に依存関係が残るため，これを取り除いた分散の推定

$$\hat{\sigma}_{(-i)}^2 = \frac{1}{n-p-2} \sum_{k=1,k\neq i}^{n} \hat{\epsilon}_k^2$$

を考え，

$$e_i^t = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}$$

を定義することがある．これを**スチューデント化残差** (studentized residual) という．なお，

$$(n-p-2)\hat{\sigma}_{(-i)}^2 + \hat{\epsilon}_i^2 = (n-p-1)\hat{\sigma}^2$$

という関係があるので，スチューデント化の計算は容易に実行できる．

練習問題 (8)

2.3.5 最小二乗推定量の性質

真の回帰係数を $\boldsymbol{\beta}$ とし，観測データには独立な誤差が重畳している

$$Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathsf{T}},\ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

とする．このとき推定量 $\hat{\boldsymbol{\beta}}$ の平均と分散はそれぞれ

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y] \\
&= \mathbb{E}[(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}(X\boldsymbol{\beta}+\boldsymbol{\epsilon})] \\
&= \mathbb{E}[\boldsymbol{\beta} + (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\boldsymbol{\epsilon}] \\
&= \boldsymbol{\beta} \\
\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathsf{T}}] \\
&= \mathbb{E}\left[\left((X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\boldsymbol{\epsilon}\right)\left((X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\boldsymbol{\epsilon}\right)^{\mathsf{T}}\right] \\
&= (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}]X(X^{\mathsf{T}}X)^{-1} \\
&= \sigma^2(X^{\mathsf{T}}X)^{-1}
\end{aligned}$$

と計算される．誤差 ϵ_i が独立な正規分布に従うことから，その和として表される推定量 $\hat{\boldsymbol{\beta}}$ も正規分布に従う:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\big(\boldsymbol{\beta}, \sigma^2 (X^\mathsf{T} X)^{-1}\big).$$

また，行列 $(X^\mathsf{T}X)^{-1}$ の対角成分を $\xi_i, i = 1, \ldots, p$ とすると，各成分の分布は平均 β_i，分散 $\sigma^2\xi_i$(標準偏差 $\sigma\sqrt{\xi_i}$) の正規分布に従う．この分布の標準偏差を用いると推定量の各成分の精度の指標を与えることができるが，誤差の分散 σ^2 は通常未知なので，別途推定する必要がある．一般的な方法としては不偏分散

$$\hat{\sigma}^2 = \frac{1}{n-p-1} S(\hat{\boldsymbol{\beta}}) \qquad \text{(残差平方和/自由度)}$$

用いることになる．不偏標準偏差 $\hat{\sigma}$ と行列 $(X^\mathsf{T}X)^{-1}$ の対角成分 ξ_i を用いて定義される

$$\hat{\sigma}\sqrt{\xi_i} = \sqrt{\frac{S(\hat{\boldsymbol{\beta}})\xi_i}{n-p-1}}$$

を $\hat{\beta}_i$ の**標準誤差** (standard error) と呼ぶ．

なお，ハット行列 H の定義より

$$\begin{aligned} \operatorname{tr} H &= \operatorname{tr} X(X^\mathsf{T}X)^{-1}X^\mathsf{T} + \operatorname{tr} M \\ &= \operatorname{tr}(X^\mathsf{T}X)^{-1}X^\mathsf{T}X + \operatorname{tr} M = p + 1 \end{aligned}$$

が成り立つので，不偏分散の平均は

$$\mathbb{E}[\hat{\sigma}^2] = \frac{1}{n-p-1} \operatorname{tr} \mathbb{E}[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^\mathsf{T}] = \frac{1}{n-p-1} \operatorname{tr}(I-H)\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\mathsf{T}] = \sigma^2$$

となり，たしかに不偏であることが確認できる．また，分散で正規化した残差平方和 $S(\hat{\boldsymbol{\beta}})$ は自由度 $n-p-1$ の χ^2-分布に従うこと

$$\frac{S(\hat{\boldsymbol{\beta}})}{\sigma^2} = (n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$$

が知られているので，不偏分散のより詳しい性質を調べるにはこれを利用すればよい．

2.3.6 テコ比

節 2.3.3 で求めたハット行列 H を用いた実測値と予測値の関係から，行列 H の ij 成分 h_{ij} を用いて成分毎に書くと

$$\begin{aligned} \hat{y}_i &= \sum_{j=1}^{n} h_{ij} y_j \\ &= h_{ii} y_i + (\text{その他の } y_j \text{ に関する項}) \end{aligned}$$

であり，予測値 $\hat{y}_i$ は y_i による項とそれ以外の y_j, $j \neq i$ による項に分離することができる．このとき，予測値 $\hat{y}_i$ に含まれる y_i の係数

$$h_{ii} = \frac{1}{n} + (X(X^\mathsf{T}X)^{-1}X^\mathsf{T})_{ii}$$

を**テコ比** (leverage) と言う．これは y_i を推定する際の y_i に依存する度合を表す量であり，

テコ比が大きい y_i を使って y_i を予測している
すなわち y_i の予測には回帰式があまり役に立たず，他のデータからは予測し難いデータである

テコ比が小さい y_i を使わずに y_i を予測している
すなわち回帰式がうまく機能し，y_i がなくても他のデータから予測しやすデータである

ことを示している．なお，テコ比は $\boldsymbol{x}_i$ にのみに依存しているので，テコ比で評価している目的変数 y_i の予測の難しさは，説明変数 $\boldsymbol{x}_i$ の配置のみに基づくものであることがわかる．

特に単回帰 (説明変数が 1 次元) の場合にはテコ比は

$$h_{ii} = \frac{(x_i - \bar{x})^2}{S_x} + \frac{1}{n}$$

と書けることから，各データのテコ比の間には

$$\sum_{i=1}^{n} h_{ii} = 2, \quad \frac{1}{n} \leq h_{ii} \leq 1$$

という関係が成り立つことが直接確かめられる．また，重回帰においては行列 H の定義より

練習問題 (3)

$$\begin{aligned}\sum_{i=1}^{n} h_{ii} &= \operatorname{tr} H = \operatorname{tr} X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}} + \operatorname{tr} M \\ &= \operatorname{tr}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}X + \operatorname{tr} M \\ &= p + n \times \frac{1}{n} = p + 1\end{aligned}$$

が成り立つことがわかる．

2.3.7 Cook の距離

Cook の距離 (Cook's distance) は各データについて定義される量で，データ $(\boldsymbol{x}_i, y_i)$ が回帰式にどのくらい影響力をもっているか，あるいはデータ全体の中でどのくらい特異であるのかを示す指標となる．

$\hat{y}_j$: データを全て使った y_j の予測値
$\hat{y}_{j(-i)}$: データ $(\boldsymbol{x}_i, y_i)$ 以外を全て使った y_j の予測値

とすると，データ $(\boldsymbol{x}_i, y_i)$ に関する Cook の距離 D_i は

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(-i)})^2}{\underbrace{p}_{\text{回帰式の自由度}}} \times \underbrace{S/\phi}_{\text{残差の 2 乗平均}}$$

で定義される (自由度として $p+1$ を用いる流儀もある)．直感的にはデータ $(\boldsymbol{x}_i, y_i)$ の存在が他のデータの予測に及ぼす影響の度合を表しており，これを予測の誤差の 2 乗平均で正規化して評価していると考えることができる．

一方

$\hat{\boldsymbol{\beta}}$: データを全て使ったパラメタの推定値

$\hat{\boldsymbol{\beta}}_{(-i)}$: データ $(\boldsymbol{x}_i, y_i)$ 以外を全て使ったパラメタの推定値

として

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}}\right)^{\mathsf{T}} \mathrm{Cov}(\hat{\boldsymbol{\beta}})^{-1} \left(\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}}\right)}{p}$$

で与える別の定義もある．この定義は，節 2.3.5 で議論した推定量 $\hat{\boldsymbol{\beta}}$ の性質から

$$\begin{aligned} D_i &= \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}\right)^{\mathsf{T}} \mathrm{Cov}(\hat{\boldsymbol{\beta}})^{-1} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}\right)}{p} \\ &= \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}\right)^{\mathsf{T}} \sigma^{-2} (X^{\mathsf{T}} X) \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}\right)}{p} \\ &= \frac{\left(X\hat{\boldsymbol{\beta}} - X\hat{\boldsymbol{\beta}}_{(-i)}\right)^{\mathsf{T}} \left(X\hat{\boldsymbol{\beta}} - X\hat{\boldsymbol{\beta}}_{(-i)}\right)}{\sigma^2 p} \\ &= \frac{\left(\hat{Y} - \hat{Y}_{(-i)}\right)^{\mathsf{T}} \left(\hat{Y} - \hat{Y}_{(-i)}\right)}{\sigma^2 p} \end{aligned}$$

と書き直すことができる．ただし $\hat{Y}, \hat{Y}_{(-i)}$ はそれぞれ $\hat{y}_j, \hat{y}_{j(-i)}$ をならべてできる行列とする．このとき誤差の分散 σ^2 の値が必要となるが，これをその推定量である残差の不偏分散で置き換えると予測値を用いた定義と等しくなる．

2.3.8 多重共線性

説明変数 (前述したように独立変数と呼ぶこともある) が複数あるとき，それらは独立であることが望ましい．目的変数を効率良く説明するためには，説明変数が出来るだけ異なる情報を持っていた方が良いことは直感的にも明らかであろう．ところが実際のデータでは説明変数の独立性が保証されることは少なく，変数のいくつかは似たような挙動をすることになる．このような説明変数間の従属性，特に線形従属性は実データの解析においては重大な問題を引き起こすことがある．

例えば，目的変数 y が 3 つの説明変数 x_1, x_2, x_3 によって

$$y = x_1 + x_2 + x_3 + \epsilon$$

と表されるとする．ここで x_1 と x_2 は線形従属な関係にあり，極端な例であるが，

$$x_1 = x_2$$

が成り立つとしよう．このとき上の目的変数と説明変数の関係は

$$\begin{aligned} y &= 1.5 \times x_1 + 0.5 \times x_2 + x_3 + \epsilon \\ y &= 2.0 \times x_1 \qquad\qquad + x_3 + \epsilon \\ y &= \qquad\quad + 2.0 \times x_2 + x_3 + \epsilon \\ y &= 3.0 \times x_1 - 1.0 \times x_2 + x_3 + \epsilon \end{aligned}$$

など様々な同値表現を持ち，解の一意性が失われることになる．実際には完全な線形従属性が成り立つ訳ではなく，雑音成分 ξ を含んだ形で

$$x_1 = x_2 + \xi$$

のように表される近似的な従属関係となることが多いため，どのような解が得られるかはデータに依存する．雑音成分に依存して，例えば上の 4 番目の例のように係数の正負が逆転 (x_2 が y に対して負の影響を与える) した不適切な解が得られてしまうこともある．このような独立変数間の線形従属関係は**多重共線性** (colinearity) と呼ばれる．

さて，節 2.3.5 で示したように，回帰係数の推定精度 (推定量の分散共分散行列) は

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^\mathsf{T} X)^{-1}$$

で表される．ここで回帰係数の第 1 成分 β_1 の分散は，上記の行列の第 $(1,1)$ 成分になることに注意しよう．行列 X を説明変数の第 1 成分とそれ以外にブロック化して

$$X = \begin{pmatrix} Y_1 & X_1 \end{pmatrix}, \quad Y_1 = \begin{pmatrix} x_{11} - \bar{x}_1 \\ x_{21} - \bar{x}_1 \\ \vdots \\ x_{n1} - \bar{x}_1 \end{pmatrix}, \quad X_1 = \begin{pmatrix} x_{12} - \bar{x}_2 & \dots & x_{1p} - \bar{x}_p \\ x_{22} - \bar{x}_2 & \dots & x_{2p} - \bar{x}_p \\ \cdots & \cdots & \cdots \\ x_{n2} - \bar{x}_2 & \dots & x_{np} - \bar{x}_p \end{pmatrix}$$

と書くことにする．このとき

$$\begin{aligned} (X^\mathsf{T} X)^{-1} &= \left(\begin{pmatrix} Y_1^\mathsf{T} \\ X_1^\mathsf{T} \end{pmatrix} \begin{pmatrix} Y_1 & X_1 \end{pmatrix} \right)^{-1} \\ &= \begin{pmatrix} Y_1^\mathsf{T} Y_1 & Y_1^\mathsf{T} X_1 \\ X_1^\mathsf{T} Y_1 & X_1^\mathsf{T} X_1 \end{pmatrix} \end{aligned}$$

となるので，ブロック行列の逆行列表現から回帰係数の第 1 成分 β_1 の分散は

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \left(Y_1^\mathsf{T} Y_1 - Y_1^\mathsf{T} X_1 (X_1^\mathsf{T} X_1)^{-1} X_1^\mathsf{T} Y_1 \right)^{-1}$$

となることがわかる．一方，x_1 を目的変数，それ以外を説明変数としたときの決定係数 R_1^2 は $X \to X_1, Y \to Y_1$ と対応づければ良いので

練習問題 (4)

$$R_1^2 = \frac{Y_1^\mathsf{T} X_1 (X_1^\mathsf{T} X_1)^{-1} X_1^\mathsf{T} Y_1}{Y_1^\mathsf{T} Y_1}$$

と表され，

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{Y_1^\mathsf{T} Y_1} \times \frac{1}{1 - R_1^2}$$

と書くことができる．説明変数の第 1 成分が他の成分とは独立な場合，他の成分では第 1 成分を表すことはできないので決定係数は $R_1^2 = 0$ となり，上記の分散はこのとき最小となる．

さて，以上の議論は第 1 成分に限らず成り立つ．すなわち，回帰係数の第 k 成分の最小分散を $\mathrm{Var}(\hat{\beta}_k)_{\min}$，決定係数を含む係数項を

$$\mathrm{VIF}_k = \frac{1}{1 - R_k^2}$$

と書くことにすれば

$$\mathrm{Var}(\hat{\beta}_k) = \mathrm{Var}(\hat{\beta}_k)_{\min} \times \mathrm{VIF}_k$$

と表すことができる．係数項 VIF_K は，多重共線性が推定量をどれだけ不安定にするかを表す指標となり，これを variance inflation factor (VIF) と呼ぶ．

実用上は，VIF 値が 10 を越えた場合は深刻な多重共線性が起きていると考えられる．また，4 を越えた場合には注意が必要で，説明変数間の関係を検証すべきであると言われている．

2.3.9 回帰係数の t 統計量

誤差が正規分布に従う場合，t-分布を用いて推定された係数の有用性について検定を行うことができる．

節 2.3.5 で見たように，係数の推定量 $\hat{\boldsymbol{\beta}}$ は正規分布に，誤差の不偏分散 $\hat{\sigma}^2$(の定数倍) は自由度 $n-p-1$ の χ^2-分布に従う．

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (X^\mathsf{T} X)^{-1})$$

$$(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$$

ここで係数の分散共分散行列の不偏推定量の (i,i) 成分を

$$(\hat{\sigma}^2 (X^\mathsf{T} X)^{-1})_{ii} = \hat{\sigma}^2 \xi_i^2$$

と書くことにすれば，以下で定義される統計量 T は自由度 $n-p-1$ の t-分布に従うことが示される．

$$T = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\xi_i} \sim \mathcal{T}(n-p-1)$$

もちろん β_i は未知であるので統計量 T は一般には計算することができないが，仮説「$\beta_i = 0$ である」のもとでは計算することができる．すなわち，

帰無仮説 $\beta_i = 0$ である (β_i は不要である)

対立仮説 $\beta_i \neq 0$ である

とし，データから推定された $\hat{\beta}_i$ と $\hat{\sigma}\xi_i$ を用いて

$$t = \frac{\hat{\beta}_i}{\hat{\sigma}\xi_i}$$

を求め，p-値 (両側検定)

$$\begin{aligned} p &= \Pr(|T| > |t| \mid T \sim \mathcal{T}(n-p-1)) \\ &= 2\int_{|t|}^{\infty} p(x)dx \end{aligned}$$

を計算し，説明変数の第 i 成分の要不要について検定を行うことができる．ただし，$p(x)$ は自由度 $n-p-1$ の t-分布の確率密度関数である．

2.3.10 回帰モデルの F 統計量

同様に誤差が正規分布に従う場合，残差の分解から F-分布を用いて回帰モデルそのものの有用性について検定を行うことができる．

節 2.3.1 での議論から

$$\frac{S_y}{\sigma^2} = \frac{S(\hat{\boldsymbol{\beta}})}{\sigma^2} + \frac{S_r(\hat{\boldsymbol{\beta}})}{\sigma^2}$$

と分解できる．ここで右辺第 1 項は節 2.3.5 で見たように，

$$\frac{S(\hat{\boldsymbol{\beta}})}{\sigma^2} \sim \chi^2(n-p-1)$$

である．また第 2 項は

$$S_r(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}^{\mathsf{T}}(X^{\mathsf{T}}X)\hat{\boldsymbol{\beta}}$$

であるが，$\hat{\boldsymbol{\beta}}$ が上記の p 次元正規分布に従うことから，$\boldsymbol{\beta} = 0$ の場合に限り

$$\frac{S_r(\hat{\boldsymbol{\beta}})}{\sigma^2} \sim \chi^2(p)$$

となることがわかる．したがって「$\boldsymbol{\beta} = 0$ である (回帰式は不要である)」という帰無仮説のもとで以下の統計量 F は自由度 $p, n-p-1$ の F-分布に従うことが示される．

練習問題 (10)

$$F = \frac{S_r/p}{S/(n-p-1)} \sim \mathcal{F}(p, n-p-1)$$

係数の検定と同様にデータから推定された $\hat{\boldsymbol{\beta}}$ を用いて計算される S_r と S を用いて

$$f = \frac{S_r(\hat{\boldsymbol{\beta}})/p}{S(\hat{\boldsymbol{\beta}})/(n-p-1)}$$

を求め，p-値 (片側検定)

$$\begin{aligned} p &= \Pr(F > f \mid F \sim \mathcal{F}(p, n-p-1)) \\ &= \int_f^\infty p(x)dx \end{aligned}$$

計算することによって，回帰式そのものの要不要について検定を行うことができる．ただし，$p(x)$ は自由度 $p, n-p-1$ の F-分布の確率密度関数である．

2.3.11 信頼区間と予測区間

節 2.3.7 で述べたように，誤差が正規分布に従うという仮定のもとでは回帰係数の精度は

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta} \\ \mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \sigma^2(X^{\mathsf{T}}X)^{-1} \end{aligned}$$

となる．

ここで推定された回帰係数 $\hat{\boldsymbol{\beta}}$ を用いて，新しいデータ $(\boldsymbol{x}, y)$ の予測を行うことを考える．ただし，ここで扱う X, Y を用いた回帰式では説明変数，目的変数ともに標本平均からの差 $(\boldsymbol{x} - \bar{\boldsymbol{x}}, y - \bar{y})$ を考える必要があるが，以下では簡単のためこれを $(\boldsymbol{x}, y)$ で表すものとする．

$$
\begin{aligned}
y &= \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta} + \epsilon && \text{(真の値)} \\
\hat{y} &= \boldsymbol{x}^{\mathsf{T}}\hat{\boldsymbol{\beta}} && \text{(推定された回帰係数による予測値)} \\
\tilde{y} &= \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta} && \text{(最適な予測値)}
\end{aligned}
$$

とすると，残差の平均と分散は

$$
\begin{aligned}
\mathbb{E}[y - \hat{y}] &= \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta} - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta} = 0 \\
\mathrm{Var}(y - \hat{y}) &= \mathrm{Var}(\boldsymbol{x}^{\mathsf{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \epsilon) \\
&= \mathrm{Var}(\boldsymbol{x}^{\mathsf{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})) + \mathrm{Var}(\epsilon) \\
&= \sigma^2 \boldsymbol{x}^{\mathsf{T}} \mathrm{Cov}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\boldsymbol{x} + \mathrm{Var}(\epsilon) \\
&= \underbrace{\sigma^2 \boldsymbol{x}^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\boldsymbol{x}}_{\hat{\boldsymbol{\beta}}\text{の精度による誤差}} + \underbrace{\sigma^2}_{\text{もともとある誤差}}
\end{aligned}
$$

と計算できる．このことから推定量 $\hat{y}$ には 2 種類の誤差が含まれていることがわかる．

ひとつは推定値の精度による誤差で，データから推定された回帰係数と真の回帰係数の差に依存するものである．この分散は $\boldsymbol{x}$ に依存して

$$
\mathbb{E}[(\tilde{y} - \hat{y})^2] = \sigma^2 \boldsymbol{x}^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\boldsymbol{x} = \sigma_c(\boldsymbol{x})^2
$$

と計算され，誤差 ϵ の正規性から

$$
\tilde{y} - \hat{y} \sim \mathcal{N}(0, \sigma_c(\boldsymbol{x})^2)
$$

となる．このため，与えられた確率 α に対し

$$
\Pr(-C_\alpha \leq Z \leq C_\alpha \mid Z \sim \mathcal{N}(0,1)) = \alpha \tag{2.2}
$$

となる定数 C_α を用いると，

$$
\Pr(-C_\alpha \leq (\tilde{y} - \hat{y})/\sigma_c(\boldsymbol{x}) \leq C_\alpha) = \alpha
$$

とすることができる．すなわち，確率 α で区間

$$
\Big[\hat{y} - C_\alpha \sigma_c(\boldsymbol{x}),\ \hat{y} + C_\alpha \sigma_c(\boldsymbol{x})\Big]
$$

の中に最適な予測値 $\tilde{y}$ が存在する (最適な予測値が入る区間をこのように構成したとき，この構成方法は確率 α で正しい) ことになる．これを**信頼区間** (confidence interval) という．

一方，データから推定された回帰式による予測値と真の値との差の分散は

$$
\mathbb{E}[(y - \hat{y})^2] = \sigma^2 \boldsymbol{x}^{\mathsf{T}}(X^{\mathsf{T}}X)^{\mathsf{T}}\boldsymbol{x} + \sigma^2 = \sigma_p(\boldsymbol{x})^2
$$

と計算され，誤差の正規性から

$$y - \hat{y} \sim \mathrm{N}(0, \sigma_p(\boldsymbol{x})^2)$$

となる．同様に確率 α で区間

$$\Big[\hat{y} - C_\alpha \sigma_p(\boldsymbol{x}),\ \hat{y} + C_\alpha \sigma_p(\boldsymbol{x})\Big]$$

の中に真の値 y が存在する (真の値が入る区間をこのように構成したとき，この構成方法は確率 α で正しい) ことになる．これを**予測区間** (prediction interval) という．

ただし，以上の議論は σ^2 の値がわかっていることを仮定しているので，正規分布を用いて区間幅 C_α が決められている．実際の問題では σ^2 の推定量として残差の不偏分散

$$\hat{\sigma}^2 = \frac{1}{n-p-1} S(\hat{\boldsymbol{\beta}})$$

を用い，対応する予測値の分散はそれぞれ

$$\begin{aligned}
\hat{\sigma}_c(\boldsymbol{x})^2 &= \hat{\sigma}^2 \boldsymbol{x}^\mathsf{T} (X^\mathsf{T} X)^{-1} \boldsymbol{x} \\
\hat{\sigma}_p(\boldsymbol{x})^2 &= \hat{\sigma}^2 \boldsymbol{x}^\mathsf{T} (X^\mathsf{T} X)^{-1} \boldsymbol{x} + \hat{\sigma}^2
\end{aligned}$$

を用いることになる．このため統計量は分散の推定誤差も考慮した裾の重い分布に従う．具体的には，$(\tilde{y} - \hat{y})/\hat{\sigma}_c$ および $(y - \hat{y})/\hat{\sigma}_p$ の分布は正規分布ではなく，いずれも自由度 $n-p-1$ の t-分布に従う:

$$\frac{\tilde{y} - \hat{y}}{\hat{\sigma}_c(\boldsymbol{x})} \sim \mathrm{T}(n-p-1), \qquad \frac{y - \hat{y}}{\hat{\sigma}_p(\boldsymbol{x})} \sim \mathrm{T}(n-p-1).$$

区間幅を決める C_α は式 (2.2) の正規分布を自由度 $n-p-1$ の t-分布の密度関数で置き換えて計算することになる．

$$\Pr(-C_\alpha \le Z \le C_\alpha \mid Z \sim \mathrm{T}(n-p-1)) = \alpha$$

これを用いてそれぞれの区間は

$$\begin{aligned}
&\Big[\hat{y} - C_\alpha \hat{\sigma}_c(\boldsymbol{x}),\ \hat{y} + C_\alpha \hat{\sigma}_c(\boldsymbol{x})\Big] && \text{(信頼区間)} \\
&\Big[\hat{y} - C_\alpha \hat{\sigma}_p(\boldsymbol{x}),\ \hat{y} + C_\alpha \hat{\sigma}_p(\boldsymbol{x})\Big] && \text{(予測区間)}
\end{aligned}$$

で与えられる

2.4 解析の事例

2.5 補遺

回帰分析を理解する上で重要ながら，本稿ではまだ書き切れていない項目には以下ようなものがある．

- t-検定による係数や切片の信頼性 (significance) の考え方
- 分散分析による説明変数・モデルの選択

- AIC 情報量規準による説明変数・モデルの選択
- PLS (partial least squares), PCR (principal component regression), RMA (reduced major axis) などのその他の回帰手法
- GAM (lgeneralized additive model), NN (neural network) などの非線形回帰手法

2.5.1 標本平均を通ることの別証

2.2.2 節の計算をブロック毎にもう少し細かく見てみることにする．

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \qquad (\boldsymbol{x} \text{ の標本平均，} p \text{ 次元ベクトル})$$

$$\overline{\boldsymbol{x}^2} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}} \qquad (\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \text{ の標本平均，} p \times p \text{ 行列})$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (y \text{ の標本平均，スカラ})$$

$$\overline{\boldsymbol{x}y} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i \qquad (\boldsymbol{x}y \text{ の標本平均，} p \text{ 次元ベクトル})$$

と定義しておくと

$$X^{\mathsf{T}}X = n \overset{p+1 \text{ 次元}}{\begin{pmatrix} 1 & \bar{\boldsymbol{x}}^{\mathsf{T}} \\ \bar{\boldsymbol{x}} & \overline{\boldsymbol{x}^2} \end{pmatrix}} \; p+1 \text{ 次元}$$

$$X^{\mathsf{T}}Y = n \overset{1 \text{ 次元}}{\begin{pmatrix} \bar{y} \\ \overline{\boldsymbol{x}y} \end{pmatrix}} \; p+1 \text{ 次元}$$

と書くことができる．また，$\boldsymbol{x}$ の標本分散共分散行列 V_x や $\boldsymbol{x}$ と y の標本共分散行列 V_{xy} はそれぞれ

$$V_x = \overline{\boldsymbol{x}^2} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathsf{T}} \qquad (p \times p \text{ 行列})$$

$$V_{xy} = \overline{\boldsymbol{x}y} - \bar{\boldsymbol{x}}\bar{y} \qquad (p \times 1 \text{ 行列})$$

となることに注意すると $X^{\mathsf{T}}X/n$ の逆行列は

$$\begin{pmatrix} 1 & \bar{\boldsymbol{x}}^{\mathsf{T}} \\ \bar{\boldsymbol{x}} & \overline{\boldsymbol{x}^2} \end{pmatrix}^{-1} = \overset{p+1 \text{ 次元}}{\begin{pmatrix} 1 + \bar{\boldsymbol{x}}^{\mathsf{T}} V_x^{-1} \bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}^{\mathsf{T}} V_x^{-1} \\ -V_x^{-1}\bar{\boldsymbol{x}} & V_x^{-1} \end{pmatrix}} \; p+1 \text{ 次元}$$

練習問題 (1)

で与えられる．これより回帰式の最適な係数は

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \frac{1}{n} \begin{pmatrix} 1 + \bar{\boldsymbol{x}}^{\mathsf{T}} V_x^{-1} \bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}^{\mathsf{T}} V_x^{-1} \\ -V_x^{-1}\bar{\boldsymbol{x}} & V_x^{-1} \end{pmatrix} n \begin{pmatrix} \bar{y} \\ \overline{\boldsymbol{x}y} \end{pmatrix} \\ &= \begin{pmatrix} \bar{y} + \bar{\boldsymbol{x}}^{\mathsf{T}} V_x^{-1} \bar{\boldsymbol{x}}\bar{y} - \bar{\boldsymbol{x}}^{\mathsf{T}} V_x^{-1} \overline{\boldsymbol{x}y} \\ -V_x^{-1}\bar{\boldsymbol{x}}\bar{y} + V_x^{-1}\overline{\boldsymbol{x}y} \end{pmatrix} \\ &= \begin{pmatrix} \bar{y} - \bar{\boldsymbol{x}}^{\mathsf{T}} V_x^{-1} V_{xy} \\ V_x^{-1} V_{xy} \end{pmatrix} \end{aligned}$$

で表わされる．このとき，求めた線形回帰式による目的変数の予測値を $\hat{y}$ とすれば

$$\begin{aligned}\hat{y} &= \left(1, \boldsymbol{x}^{\mathsf{T}}\right)\hat{\boldsymbol{\beta}} \\ &= \bar{y} - \bar{\boldsymbol{x}}^{\mathsf{T}} V_x^{-1} V_{xy} + \boldsymbol{x}^{\mathsf{T}} V_x^{-1} V_{xy} \\ &= \bar{y} + (\boldsymbol{x} - \bar{\boldsymbol{x}})^{\mathsf{T}} V_x^{-1} V_{xy}\end{aligned}$$

すなわち

$$\hat{y} - \bar{y} = (\boldsymbol{x} - \bar{\boldsymbol{x}})^{\mathsf{T}} V_x^{-1} V_{xy}$$

となるので，最適な線形回帰式は $\boldsymbol{x}$ と y の標本平均を通ることがわかる．

練習問題

(1) $X^{\mathsf{T}}X/n$ の逆行列をブロックに分解して，a をスカラ，$\boldsymbol{b}$ を p 次元ベクトル，C を $p \times p$ 行列として $\left(\begin{smallmatrix} a & b^{\mathsf{T}} \\ b & C \end{smallmatrix}\right)$ で表すことにすれば，

$$\begin{pmatrix} 1 & \bar{\boldsymbol{x}}^{\mathsf{T}} \\ \bar{\boldsymbol{x}} & \overline{\boldsymbol{x}^2} \end{pmatrix} \begin{pmatrix} a & \boldsymbol{b}^{\mathsf{T}} \\ \boldsymbol{b} & C \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I_p \end{pmatrix}$$

を満たさなければいけない．ただし，I_p は $p \times p$ の単位行列とする．ブロック毎に整理すると

$$\begin{aligned} a + \bar{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{b} &= 1 \\ \boldsymbol{b}^{\mathsf{T}} + \bar{\boldsymbol{x}}^{\mathsf{T}} C &= 0 \\ \bar{\boldsymbol{x}} a + \overline{\boldsymbol{x}^2}\boldsymbol{b} &= 0 \\ \bar{\boldsymbol{x}}\boldsymbol{b}^{\mathsf{T}} + \overline{\boldsymbol{x}^2} C &= I_p \end{aligned}$$

となるので，これを解いて $X^{\mathsf{T}}X/n$ の逆行列を求めなさい (C から求めるとよい)．

(2) 説明変数 $\boldsymbol{x}$ と目的変数 y に対して，n 個のデータ $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$ が与えられているとき，その標本分散・共分散が，標本平均を用いて定義した行列 X と Y を用いてそれぞれ $X^{\mathsf{T}}X/n, Y^{\mathsf{T}}Y/n, X^{\mathsf{T}}Y/n$ で表されることを示しなさい．

(3) 単回帰におけるテコ比の間に

$$\sum_{i=1}^{n} h_{ii} = 2, \quad \frac{1}{n} \leq h_{ii} \leq 1$$

という関係が成り立つことを示しなさい．

(4) 行列 H において

$$HX = X, \quad H\mathbb{1} = \mathbb{1}$$

が成り立つことを示しなさい．

(5) 行列 H において

$$H^2 = H, \quad (I-H)^2 = I-H$$

が成り立つことを示しなさい.

(6) 中心化していない計画行列と中心化した計画行列を用いたハット行列の定義が一致することを示しなさい.

(7) 不偏分散

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2$$

が不偏となることを

$$\sum_{i=1}^{n} \mathbb{E}[e_i^2] = \mathrm{tr}(I-H)$$

であることを用いて示しなさい.

(8) スチューデント化誤差における関係式

$$(n-p-2)\hat{\sigma}^2_{(-i)} = (n-p-1)\hat{\sigma}^2 - e_i^2$$

を示しなさい.

(9) 以下のブロック行列において D の逆行列が存在するとき，次の式が成り立つことを示しなさい.

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} (A-BD^{-1}C)^{-1} & -(A-BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A-BD^{-1}C)^{-1} & D^{-1}+D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}$$

(10) 係数の推定量 $\hat{\boldsymbol{\beta}}$ が以下の正規分布に従うとき

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(X^\mathsf{T}X)^{-1})$$

$$Z = \frac{S_r(\hat{\boldsymbol{\beta}})}{\sigma^2} = \frac{\hat{\boldsymbol{\beta}}^\mathsf{T}(X^\mathsf{T}X)\hat{\boldsymbol{\beta}}}{\sigma^2}$$

が自由度 p の χ^2-分布に従うことを確かめよ.

3 主成分分析

3.1 目的と考え方

3.1.1 目的

主成分分析 (principal component analysis; PCA) とは，線形変換を用いて多数の変量を縮約して表現する低次元の量 (指標または特徴量) を見付けることによって，変数間の関係を明らかにするための分析法である．陽には見えない隠れている特徴 (これを主成分と呼ぶ) を，観測された変量の線形結合で表す方法とも言える．

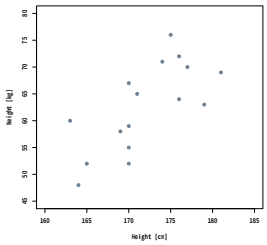
例 3.1. 身長と体重の散布図が拡がっている方向は，直感的には体の大きさ (体格) を表していると考えることができる．身長と体重を統合して，体格を表す新しい量

$$(\text{体格}) = f(\text{身長}, \text{体重})$$

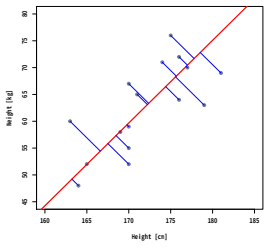
を合成することによって，変量間の関係を数量化して考えることができる．通常は変量の線形結合を考える．

身長と体重のデータ

	身長 [cm]	体重 [kg]
1	164	48
2	170	55
3	171	65
4	174	71
5	176	64
6	181	69
7	170	52
8	176	72
9	170	59
10	165	52
11	169	58
12	170	67
13	179	63
14	163	60
15	177	70
16	175	76



身長と体重の散布図

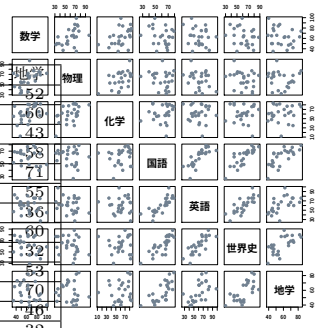


主成分方向 (赤) と残差 (青)

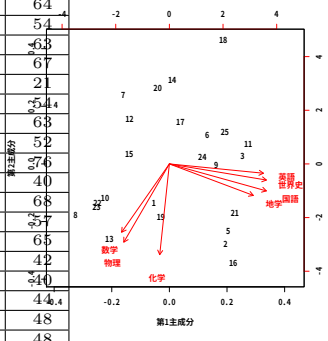
例 3.2. 国語，英語，数学，物理，化学，世界史，地学の 7 つの試験を 100 名の学生に対して行ったとしよう．学生の成績がどのように分布しているかを知るには，7 次元空間に各学生の成績をプロットして図示すれば良いが，この空間を我々が実感することは難しい．しかしながら 7 次元空間の中に巧い 2 次元空間を探し出してその上に全データを投影 (射影) することができれば，学生の成績を 2 次元上の分布として近似的にではあるが把握することができる．

成績データ

番号	国語	英語	数学	物理	化学	世界史	地学
1	46	53	40	45	57	43	52
2	55	55	19	30	61	44	60
3	50	52	27	31	43	61	43
4	58	60	33	44	50	54	58
5	75	91	60	98	53	73	71
6	51	60	29	44	1	55	55
7	38	51	35	68	60	48	36
8	49	43	37	29	48	58	60
9	25	31	35	26	41	35	32
10	61	74	48	35	53	64	53
11	65	65	55	83	43	72	70
12	43	51	66	53	39	43	46
13	34	49	51	49	51	21	32
14	48	55	66	73	71	68	74
15	34	42	43	36	54	38	39
16	64	67	55	58	61	63	62
17	50	67	65	65	69	59	64
18	56	50	54	48	63	53	54
19	71	52	47	57	69	56	63
20	45	52	55	61	58	41	67
21	50	49	68	58	83	47	21
22	40	35	65	60	56	46	54
23	65	49	63	50	50	64	63
24	37	40	53	38	25	42	52
25	69	78	60	66	67	90	76
26	41	56	83	66	87	46	40
27	69	55	39	70	39	58	68
28	44	39	63	47	35	50	57
29	78	79	26	37	33	65	65
30	11	47	65	74	76	37	42
31	31	45	57	47	38	39	40
32	49	67	44	49	56	62	44
33	50	49	30	36	49	50	48
34	62	70	44	57	79	57	48
35	54	48	46	45	54	40	36



データの散布図



`biplot` によるデータの表示

主成分分析は，高次元に分布する観測データの性質を理解するために

- 本来の特徴をできる限り保持する低次元空間を見付ける (**次元縮約**)
- データに内在する重要な指標を抜き出し，余計な雑音を抑える (**特徴抽出，雑音除去**)

という 2 つの側面を持つ．主に前者の目的が強調されることが多いが，後者は例えば主成分分析を行ってから回帰分析を行う**主成分回帰** (principal component regression; PCR) といった考え方において利用されている．以下ではこの目的に適う巧い射影方向の求め方を，射影されたデータのばらつき，あるいは残差のばらつきという観点から捉え，主成分分析の考え方を概説する．

3.1.2 観測データ

多次元データの取り扱いは回帰分析とは異なり，変量に説明変数と目的変数の区別はなく，全てが説明変数に相当する．以下では 1 つの観測値 (データ点) は p 次元とし，n 個の観測データが得られたとする．このとき i 番目のデータを $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathsf{T}}$, $i = 1, 2, \ldots, n$ で表すことにする．

3.1.3 確率モデル

主成分分析では通常明示的な生成モデルを考えないが，以下のような仮想的な確率モデル (確率的主成分分析モデル，あるいは因子分析モデルの一種) を想定することがある．

各観測データ $\boldsymbol{x}_i$ は，背後に隠れていて見えない d 次元 $(d \leq p)$ の特徴量 $\boldsymbol{s}_i = (s_{i1}, s_{i2}, \ldots, s_{id})^{\mathsf{T}}$ によって，以下の式で生成されていると考える．

$$\boldsymbol{x}_i = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 s_{i1} + \boldsymbol{\alpha}_2 s_{i2} + \cdots + \boldsymbol{\alpha}_d s_{id} + \boldsymbol{\epsilon}_i,\ i = 1, 2, \ldots, n$$

ただし $\boldsymbol{\alpha}_k = (\alpha_{k1}, \alpha_{k2}, \ldots, \alpha_{kp})^{\mathsf{T}},\ k = 1, 2, \ldots, d$ の次元はデータと同じく p 次元で，$\boldsymbol{\alpha}_0$ はデータの平均，$\boldsymbol{\alpha}_k,\ k = 1, \ldots, d$ は互いに直交する d 個のベクトルである．また $\boldsymbol{\epsilon}_i$ は誤差を表す．すなわち，観測データは p 次元であるが，本質的な情報は d 次元部分空間にのみ拡がっており，これに雑音が重畳して観測されていると考えることになる．言い方を変えると，特徴量の方向 $\boldsymbol{\alpha}_k, k = 1, 2, \ldots, d$ は互いに直交して拡がっているという非現実的なモデルを考えていると思ってもよい．

なお，$\boldsymbol{\epsilon}$ の分布については，通常は p 次元の等方的な誤差 (分散が単位行列の定数倍) を考える．また，誤差 $\boldsymbol{\epsilon}$ の分散は特徴量 $\boldsymbol{s}$ の分散に比べて十分に小さい状況を想定して近似的に解くことになる．この仮定のため上記で述べた生成モデルは厳密性を欠くことになる．

3.1.4 特徴量を再構成するための指針

以下の解析では，主に $d = 1$ の場合を考え，その後一般化を行う．まず $d = 1$ の場合には，

$$\boldsymbol{x}_i = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 s_{i1} + \boldsymbol{\epsilon}_i,\ i = 1, 2, \ldots, n$$

であるので，$\boldsymbol{\alpha}_1$ を方向ベクトルとする直線上に観測データを直交射影して，1 次元の特徴量 s を再現するための新しい量を作ることを考えればよい．特徴量 s と雑音 $\boldsymbol{\epsilon}$ の分散の大小関係の仮定から，どの方向に射影するのが良いかを決めるための指針としては，以下の 2 通りの考え方が重要である．

1. 射影により作られた新しい量ができるだけ広く分布して，各観測データ間の違いを十分に表すことができる．
   $\Rightarrow$ 射影された量のばらつき (平方和または分散で測ればよい) を最大にする．

2. 射影するときに捨て去られる情報をできるだけ少なくする．
   $\Rightarrow$ 残差のばらつき (平方和または分散) を最小にする．

一見異なる基準のように思われるが，実はこの二つは等価である．これは後に詳しく計算するように正射影の性質から

(元のデータのばらつき) = (射影された値のばらつき)+(残差のばらつき)

が成り立つからである．

3.2 計算法

3.2.1 準備

分散は平行移動に関して不変であるので，射影方向の良否を評価する基準として分散 (あるいは残差平方和) を用いる場合には，観

測データを平行移動して標本平均を中心としたもの，すなわち平均が 0 となるように平行移動したものを考えるのが計算上都合が良い．

n 個の観測データ $\boldsymbol{x}_i, i = 1, 2, \ldots, n$ による標本平均を

$$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

で表す．標本平均で平行移動したデータをまとめて行列で表すこととし，これを

$$X = \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & \cdots & x_{2p} - \bar{x}_p \\ \cdots & \cdots & \cdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{np} - \bar{x}_p \end{pmatrix} \begin{matrix} \\ \\ \\ \end{matrix} \begin{matrix} p \text{ 次元} \\ \\ \\ n \text{ 次元} \end{matrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathsf{T}} - \bar{\boldsymbol{x}}^{\mathsf{T}} \\ \boldsymbol{x}_2^{\mathsf{T}} - \bar{\boldsymbol{x}}^{\mathsf{T}} \\ \cdots \\ \boldsymbol{x}_n^{\mathsf{T}} - \bar{\boldsymbol{x}}^{\mathsf{T}} \end{pmatrix}$$

と書くことにする．この操作を**中心化** (centering) という．このとき，全ての要素が 1 である p 次元ベクトル $\mathbb{1} = (1, 1, \ldots, 1)^{\mathsf{T}}$ を用いると $\mathbb{1}^{\mathsf{T}} X = 0$ または $X^{\mathsf{T}} \mathbb{1} = 0$ となることに注意する．これは中心化されたデータの平均が 0 であることを意味する．以下ではデータは中心化されていることを仮定し，単に

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{matrix} \\ \\ \\ \end{matrix} \begin{matrix} p \text{ 次元} \\ \\ \\ n \text{ 次元} \end{matrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathsf{T}} \\ \boldsymbol{x}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathsf{T}} \end{pmatrix}$$

と書くこととする．

射影方向を表すために p 次元の単位ベクトル

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix}, \ \|\boldsymbol{\alpha}\| = 1 \quad (\text{または} \boldsymbol{\alpha}^{\mathsf{T}} \boldsymbol{\alpha} = 1 \text{と書くこともできる})$$

を考える．

このときデータ点 $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^{\mathsf{T}}$ を，原点を通り $\boldsymbol{\alpha}$ を方向ベクトルとする直線上へ直交射影した点は

$$\boldsymbol{z} = (\boldsymbol{x} \cdot \boldsymbol{\alpha}) \boldsymbol{\alpha} = (\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\alpha}) \boldsymbol{\alpha}$$

と書くことができるので，n 個の観測データを射影した点はまとめて

$$Z = \begin{pmatrix} \boldsymbol{z}_1^{\mathsf{T}} \\ \boldsymbol{z}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{z}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathsf{T}} \boldsymbol{\alpha} \boldsymbol{\alpha}^{\mathsf{T}} \\ \boldsymbol{x}_2^{\mathsf{T}} \boldsymbol{\alpha} \boldsymbol{\alpha}^{\mathsf{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathsf{T}} \boldsymbol{\alpha} \boldsymbol{\alpha}^{\mathsf{T}} \end{pmatrix} = X \boldsymbol{\alpha} \boldsymbol{\alpha}^{\mathsf{T}}$$

と表すことができる．

一方，観測データと射影の間の残差は

$$\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{z}$$

であるから，n 個の観測データの残差はまとめて

$$R = X - Z = X(I - \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathsf{T}})$$

と表すことができる．このとき $\boldsymbol{x}$ から $\boldsymbol{z}$ への射影行列 $I - \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathsf{T}}$ は冪等 (idempotent)

$$(I - \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathsf{T}})(I - \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathsf{T}}) = I - \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathsf{T}}$$

となっていることに注意する．これは射影したものをもう一度射影しても変わらないという性質を表している．

3.2.2 射影のばらつきの最大化

射影されたデータのばらつきは，射影された直線上でのデータ点の分散で評価することができる．今，原点を通る直線上への射影を考えているので，射影されたデータ点 $\boldsymbol{z}_i^{\mathsf{T}} = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathsf{T}}$ は $\|\boldsymbol{\alpha}\| = 1$ であることに注意すれば $z_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\alpha}$ を座標値とする数直線上にあると考えることができる．この座標の標本平均は

$$\begin{aligned}\bar{z} &= \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\alpha} \\ &= \frac{1}{n}\mathbb{1}^{\mathsf{T}} X\boldsymbol{\alpha} = 0\end{aligned}$$

であるから，座標値の標本分散は

$$\begin{aligned}\overline{V}_z &= \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z}) = \frac{1}{n}\sum_{i=1}^{n} z_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\alpha}\right)^2 = \frac{1}{n}\boldsymbol{\alpha}^{\mathsf{T}} X^{\mathsf{T}} X\boldsymbol{\alpha}\end{aligned}$$

となる．したがって射影のばらつきの最大化を考えるには

$$S(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^{\mathsf{T}} X^{\mathsf{T}} X\boldsymbol{\alpha}$$

の最大化を考えれば良い．

3.2.3 残差のばらつきの最小化

観測データと射影されたデータの関係から，各データ点における残差はまとめて

$$R = X - Z = X(I - \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathsf{T}})$$

と書くことができる．残差は $p-1$ 次元の部分空間に拡がっているが，その平均は 0 となるので，ばらつきの評価としては残差べ

練習問題 (1)

クトルの長さの分散を考えることにする．残差ベクトル $\boldsymbol{r}$ の長さを $r = \|\boldsymbol{r}\| = \sqrt{\boldsymbol{r}^\mathsf{T}\boldsymbol{r}}$ で表すことにすれば

$$\begin{aligned}
\overline{V}_r &= \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{r}_i^\mathsf{T}\boldsymbol{r}_i \\
&= \frac{1}{n}\operatorname{tr} RR^\mathsf{T} = \frac{1}{n}\operatorname{tr} R^\mathsf{T}R \\
&= \frac{1}{n}\operatorname{tr}(I - \boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T})X^\mathsf{T}X(I - \boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T}) \\
&= \frac{1}{n}\left(\operatorname{tr} X^\mathsf{T}X - 2\operatorname{tr}\boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T}X^\mathsf{T}X + \operatorname{tr}\boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T}X^\mathsf{T}X\boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T}\right) \\
&= \frac{1}{n}\left(\operatorname{tr} X^\mathsf{T}X - \boldsymbol{\alpha}^\mathsf{T}X^\mathsf{T}X\boldsymbol{\alpha}\right)
\end{aligned}$$

となる．なお計算の途中で $\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{\alpha} = \|\boldsymbol{\alpha}\|^2 = 1$，および $\operatorname{tr} AB = \operatorname{tr} BA$ を用いた．

したがって残差のばらつきの最小化を考えるには

$$\operatorname{tr} X^\mathsf{T}X - \boldsymbol{\alpha}^\mathsf{T}X^\mathsf{T}X\boldsymbol{\alpha} = \operatorname{tr} X^\mathsf{T}X - S(\boldsymbol{\alpha})$$

を $\boldsymbol{\alpha}$ に関して最小化，すなわち

$$S(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\mathsf{T}X^\mathsf{T}X\boldsymbol{\alpha}$$

の最大化を考えれば良い．したがって射影のばらつきの最大化と等価になる．

3.2.4 主成分分析における固有値問題

以上から，主成分分析の問題は $\boldsymbol{\alpha}$ を単位ベクトル ($\|\boldsymbol{\alpha}\| = 1$ あるいは $\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{\alpha} = 1$) として $S(\boldsymbol{\alpha})$ を最大化する問題に帰着できることがわかった．これを解くにはラグランジュ関数を

$$L(\boldsymbol{\alpha}, \lambda) = \boldsymbol{\alpha}^\mathsf{T}X^\mathsf{T}X\boldsymbol{\alpha} + \lambda(1 - \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{\alpha})$$

と定義して鞍点条件を求めればよい．$\boldsymbol{\alpha}$ に関する偏微分は

$$\frac{\partial L}{\partial \boldsymbol{\alpha}} = 2X^\mathsf{T}X\boldsymbol{\alpha} - 2\lambda\boldsymbol{\alpha} = 0$$

となるので，結局 $\boldsymbol{\alpha}$ を求めるには

$$X^\mathsf{T}X\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$$

という固有値問題を解けば良い．このとき

$$S(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\mathsf{T}X^\mathsf{T}X\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{\alpha} = \lambda$$

であり，$S(\boldsymbol{\alpha})$ を最大化するためには λ をできるだけ大きくすればよいので，最大固有値を取ればよい．したがって最大固有値 (第 1 固有値) に対応する固有ベクトルが解となることがわかる．

以上の議論は単一の射影方向を考えたもので，得られた $\boldsymbol{\alpha}$ は第 1 主成分方向 (主成分軸) に対応する．この方向に射影した値 $z = \boldsymbol{x}^\mathsf{T}\boldsymbol{\alpha}$ (原点からの符号付き長さ) はそのデータ点の**第 *1* 主成分**あるいは**主成分得点** (principal component score) と呼ばれる．

以上の議論は以下のようにして第 2 主成分以降にも拡張される．第 1 主成分方向 $\boldsymbol{\alpha}_1$ が求められたら，元の観測値からその主成分方向の成分を取り除いた残差

$$\boldsymbol{r}_i = \boldsymbol{x}_i - (\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\alpha}_1)\boldsymbol{\alpha}_1$$

を考えることができるので，これに対して主成分分析を行えばよい．第 1 主成分を除いたという意味で残差をまとめて

$$X_{(-1)} = X - X\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^{\mathsf{T}} = X(I - \boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^{\mathsf{T}})$$

と書くことにする．これを新たなデータと見做して主成分方向を求めるためには固有値問題

$$X_{(-1)}^{\mathsf{T}}X_{(-1)}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$$

を考えればよい．以下では $X^{\mathsf{T}}X$ の k 番目に大きな固有値 (第 k 固有値) を λ_k，それに対応する長さ 1 の固有ベクトルを $\boldsymbol{\alpha}_k$ と書くことにする．このとき，異なる固有ベクトルは直交する

$$\boldsymbol{\alpha}_k^{\mathsf{T}}\boldsymbol{\alpha}_j = 0, \quad k \neq j$$

ことと

$$X_{(-1)}^{\mathsf{T}}X_{(-1)} = (I - \boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^{\mathsf{T}})X^{\mathsf{T}}X(I - \boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^{\mathsf{T}})$$

であることに注意すれば，

$$\begin{aligned} X_{(-1)}^{\mathsf{T}}X_{(-1)}\boldsymbol{\alpha}_1 &= 0 \\ X_{(-1)}^{\mathsf{T}}X_{(-1)}\boldsymbol{\alpha}_k &= \lambda_k\boldsymbol{\alpha}_k \quad (k \neq 1) \end{aligned}$$

となることが容易に確かめられる．これから $X_{(-1)}^{\mathsf{T}}X_{(-1)}$ の最大固有値は λ_2 となるので，残差 $X_{(-1)}$ に対する主成分方向は，$X^{\mathsf{T}}X$ の第 2 固有ベクトルを選べばよい．よって X の第 2 主成分は第 2 固有ベクトル $\boldsymbol{\alpha}_2$ であることがわかる．

練習問題 (2)

この操作を 3 番目の固有ベクトル，4 番目の固有ベクトルと順次繰り返していけば，結局 $X^{\mathsf{T}}X$ の第 k 固有ベクトルが第 k 主成分となることがわかる．

練習問題 (3)

また，各固有ベクトル $\boldsymbol{\alpha}_k$ は互いに直交するので，主成分得点 $z_i^k = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\alpha}_k$ は k と j が違うとき

$$\frac{1}{n}\sum_{i=1}^{\mathsf{T}} z_i^k z_i^j = \frac{1}{n}(X\boldsymbol{\alpha}_k)^{\mathsf{T}}(X\boldsymbol{\alpha}_j) = \boldsymbol{\alpha}_k^{\mathsf{T}}X^{\mathsf{T}}X\boldsymbol{\alpha}_j = \lambda_j\boldsymbol{\alpha}_k^{\mathsf{T}}\boldsymbol{\alpha}_j = 0$$

となり，z^k と z^j は無相関になっていることが判る．

3.3 分析の評価

3.3.1 寄与率

回帰分析のところで考察した**寄与率** (proportion of the variance) を，より一般的な形で述べると

$$(\text{寄与率}) = \frac{(\text{その方法で説明できるばらつき})}{(\text{データ全体のばらつき})}$$

となる．これを主成分分析の場合に当て嵌めると

$$(\text{寄与率}) = \frac{(\text{主成分が説明しているばらつき})}{(\text{全体のばらつき})}$$

として定義すれば良いことがわかる．

残差のばらつきと同様に，データ全体のばらつきとしては中心化した各データベクトルの長さの2乗和を考えればよい．すなわち

$$\begin{aligned}(\text{全体のばらつき}) &= \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_i \\ &= \operatorname{tr} X X^{\mathsf{T}} \qquad (n \times n \text{ 行列}) \\ &= \operatorname{tr} X^{\mathsf{T}} X \qquad (p \times p \text{ 行列})\end{aligned}$$

となる．ここで行列 $X^{\mathsf{T}}X$ のスペクトル分解 (固有値と固有ベクトルによる行列の分解表現)

$$X^{\mathsf{T}}X = \lambda_1 \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1^{\mathsf{T}} + \lambda_2 \boldsymbol{\alpha}_2 \boldsymbol{\alpha}_2^{\mathsf{T}} + \cdots + \lambda_p \boldsymbol{\alpha}_p \boldsymbol{\alpha}_p^{\mathsf{T}} = \sum_{k=1}^{p} \lambda_k \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^{\mathsf{T}}$$

を使うと

$$\begin{aligned}\operatorname{tr} X^{\mathsf{T}}X &= \operatorname{tr}\left(\sum_{k=1}^{p} \lambda_k \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^{\mathsf{T}}\right) = \sum_{k=1}^{p} \lambda_k \operatorname{tr} \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^{\mathsf{T}} = \sum_{k=1}^{p} \lambda_k \operatorname{tr} \boldsymbol{\alpha}_k^{\mathsf{T}} \boldsymbol{\alpha}_k \\ &= \sum_{k=1}^{p} \lambda_k = \lambda_1 + \lambda_2 + \cdots + \lambda_p\end{aligned}$$

となる．一方，第 k 主成分の方向ベクトルを $\boldsymbol{\alpha}_k$ とすると，この方向に射影したデータのばらつきは

$$\begin{aligned}(\text{第 } k \text{ 主成分のばらつき}) &= \sum_{i=1}^{n} (\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\alpha}_k)^2 = \boldsymbol{\alpha}_k^{\mathsf{T}} X^{\mathsf{T}} X \boldsymbol{\alpha}_k \\ &= \lambda_k\end{aligned}$$

となる．

以上より，第 k 主成分の**寄与率**は

$$(\text{第 } k \text{ 主成分の寄与率}) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}$$

となり，また第 k 成分まで使った主成分分析の**累積寄与率** (cumulative proportion of the variance) は

$$(\text{第 } k \text{ 主成分までの累積寄与率}) = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

で与えられる．

この値が各主成分の説明力の評価と，いくつの主成分を用いるべきかの基準として用いられる．一般に，累積寄与率が 80%程度の主成分を使って分析を行うことが多い．

3.3.2 主成分負荷量

主成分方向の係数の大きさを見ることによって，元の特徴量が得られた成分に対してどの程度貢献あるいは影響しているかを知ることができる．しかしながら，特徴量のスケールによって係数の大きさは変化するため，例え特徴量が正規化 (平均 0，分散 1) にされていたとしても，そのまま大きさを比較することは適切でない場合もある．特徴量のスケールによらず，影響の度合を調べる方法として相関係数を用いるものがある．これを**主成分負荷量** (principal component loading) と呼ぶ．

第 k 成分が 1 で，残りが全て 0 であるベクトルを $\boldsymbol{e}_k$ と書くことにする．n 個のデータの第 k 主成分の得点は $X\boldsymbol{v}_k$，第 j 特徴量は $X\boldsymbol{e}_j$ とベクトル標記することができるので，その相関係数は

$$\begin{aligned}\mathrm{Cor}(X\boldsymbol{v}_k, X\boldsymbol{e}_j) &= \frac{(X\boldsymbol{v}_k)^{\mathsf{T}} X\boldsymbol{e}_j}{\sqrt{(X\boldsymbol{v}_k)^{\mathsf{T}} X\boldsymbol{v}_k}\sqrt{(X\boldsymbol{e}_j)^{\mathsf{T}} X\boldsymbol{e}_j}} \\ &= \frac{\boldsymbol{v}_k^{\mathsf{T}} X^{\mathsf{T}} X\boldsymbol{e}_j}{\sqrt{\boldsymbol{v}_k^{\mathsf{T}} X^{\mathsf{T}} X\boldsymbol{v}_k}\sqrt{\boldsymbol{e}_j^{\mathsf{T}} X^{\mathsf{T}} X\boldsymbol{e}_j}} \\ &= \frac{\lambda_k \boldsymbol{v}_k^{\mathsf{T}} \boldsymbol{e}_j}{\sqrt{\lambda_k}\sqrt{(X^{\mathsf{T}}X)_{jj}}}\end{aligned}$$

となる．特に X を正規化した場合には $X^{\mathsf{T}}X$ の対角成分は全て 1，すなわち $(X^{\mathsf{T}}X)_{jj} = 1$ となるので，第 k 主成分に対する第 j 特徴量の負荷量は

$$(\boldsymbol{l}_k)_j = \sqrt{\lambda_k}(\boldsymbol{v}_k)_j$$

となり，第 k 主成分に対する負荷量をベクトル表示すると

$$\boldsymbol{l}_k = \sqrt{\lambda_k}\boldsymbol{v}_k = \sigma_k \boldsymbol{v}_k \quad \text{(特異値による表現)}$$

となる．したがって，同じ主成分に対する各特徴量の影響の度合は固有ベクトルの成分比を見ればよいが，同じ特徴量の各主成分への影響の度合は固有値の平方根で重み付けする必要があることがわかる．

3.3.3 biplot

階数 r の $n \times p$ 型行列 X は

$$X = U\Sigma V^{\mathsf{T}} \tag{3.1}$$

の形に分解される．ここで U は $n \times n$ 型直交行列，V は $p \times p$ 型直交行列であり，$O_{s,t}$ を $s \times t$ 型零行列，D を $r \times r$ 型体格行列として Σ は

$$D = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{pmatrix}, \quad \sigma_1 \geq \sigma_2 \geq \sigma_r > 0$$

$$\Sigma = \begin{pmatrix} D & O_{r,p-r} \\ O_{n-r,r} & O_{n-r,m-r} \end{pmatrix} \begin{matrix} \\ n \text{ 次元} \end{matrix}$$

$$\begin{matrix} p \text{ 次元} \\ \end{matrix}$$

のような $n \times p$ 型の行列である．行列 D の対角成分 $\sigma_k, k = 1, \ldots, r$ を行列 X の**特異値** (singular value)，式 (3.1) を行列 X の**特異値分解** (singular value decomposition)，と呼ぶ．このとき

$$\begin{aligned} X^{\mathsf{T}}X &= (U\Sigma V^{\mathsf{T}})^{\mathsf{T}}(U\Sigma V^{\mathsf{T}}) \\ &= V\Sigma^{\mathsf{T}}U^{\mathsf{T}}U\Sigma V^{\mathsf{T}} = V\Sigma^{\mathsf{T}}\Sigma V^{\mathsf{T}} \end{aligned}$$

であり，

$$\Sigma^{\mathsf{T}}\Sigma = \begin{pmatrix} \sigma_1^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_r^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

となるので，行列 V の第 k 列ベクトルを $\boldsymbol{v}_k$，

$$\lambda_k = \begin{cases} \sigma_k^2, & k \leq r \\ 0, & k > r \end{cases}$$

とすれば，

$$\begin{aligned} X^{\mathsf{T}}X\boldsymbol{v}_k &= V\Sigma^{\mathsf{T}}\Sigma V^{\mathsf{T}}\boldsymbol{v}_k \\ &= \lambda_k \boldsymbol{v}_k \end{aligned}$$

であり，$X^{\mathsf{T}}X$ の固有値は行列 X の特異値の平方，固有ベクトルは行列 V の列ベクトル，すなわち $\boldsymbol{\alpha}_k = \boldsymbol{v}_k$ となることがわかる．

練習問題 (4)

行列 U の第 k 列ベクトルを $\boldsymbol{u}_k$ とすれば，

$$X = \sum_k \boldsymbol{u}_k \sigma_k \boldsymbol{v}_k^{\mathsf{T}}$$

と表すことができる．この関係を用いると，第 k 主成分と第 j 主成分を用いた行列 X の近似 X' は

$$X \simeq X' = \boldsymbol{u}_k \sigma_k \boldsymbol{v}_k^{\mathsf{T}} + \boldsymbol{u}_j \sigma_j \boldsymbol{v}_j^{\mathsf{T}}$$

で表されることがわかる．主成分分析の結果を視覚化するために用いられる `biplot` 表示は，この分解を利用したデータと主成分方向の散布図である．X のばらつきをできるだけ保持するという意味で最適な近似は $k = 1, j = 2$ のときに与えられることは示したが，寄与率に応じて適当な次数までの主成分を用いることもある．さて $0 \leq s \leq 1$ として上記の近似式は

$$X' = GH^{\mathsf{T}}, \quad G = \underbrace{\begin{pmatrix} \sigma_k^{1-s}\boldsymbol{u}_k & \sigma_j^{1-s}\boldsymbol{u}_j \end{pmatrix}}_{\text{2 次元}} \text{ } n \text{ 次元}, \quad H = \underbrace{\begin{pmatrix} \sigma_k^{s}\boldsymbol{v}_k & \sigma_j^{s}\boldsymbol{v}_j \end{pmatrix}}_{\text{2 次元}} p \text{ 次元}$$

練習問題 (5)

と書くことができる．行列 G の各行は各データに 2 次元の座標を与え，行列 H の各行は各変量に 2 次元の座標を与えることに注意しよう．これらの 2 次元を用いて散布図を作成したものが `biplot` となる．散布図のスケールを制御するパラメタ s としては 0,1 または 1/2 が用いられることが多い．

3.4 解析の事例

3.5 補遺

本稿で取り上げられなかった項目は以下の通りである.

- 確率的主成分分析 (probabilistic PCA)
- カーネル主成分分析 (kernel PCA)
- 正準相関分析 (canonical correlation analysis; CCA)
- 因子分析 (factor analysis; FA) との関係
- 独立成分分析 (independent component analysis; ICA) との関係

練習問題

(1) 残差 R の平均が 0 となることを示しなさい.

(2) 行列 $X_{(-1)}$ に対して

$$X_{(-1)}^{\mathsf{T}} X_{(-1)} \boldsymbol{\alpha}_1 = 0$$
$$X_{(-1)}^{\mathsf{T}} X_{(-1)} \boldsymbol{\alpha}_k = \lambda_k \boldsymbol{\alpha}_k \quad (k \neq 1)$$

が成り立つことを示しなさい.

(3) X の第 k 主成分が, $X^{\mathsf{T}}X$ の第 k 固有ベクトル $\boldsymbol{\alpha}_k$ となることを確かめなさい.

(4) $X^{\mathsf{T}} X \boldsymbol{v}_k = \lambda_k \boldsymbol{v}_k$ となることを確かめなさい.

(5) $X' = GH^{\mathsf{T}}$ が X のばらつきを保持する最適な近似であることを説明しなさい.

4 判別分析

4.1 目的と考え方

4.1.1 目的

判別分析 (discriminant analysis) は，多次元の特徴量に与えられた**クラスラベル** (または**カテゴリ**; class label, category) を予測するための関係式を構成する方法である．以下では特徴量を表す確率変数を X，クラスを表す確率変数を C とする．特徴量 X としては p 次元の実数値を考える．2 値判別の場合はクラス C として 2 つのラベル (1/0，$+1/-1$，a/b，$\circ/\times$ など) を考える．一方，多値判別の場合は多数のラベルを対象とした分類を考える．具体的には，特徴量を入力，クラスを出力とする判別関数を構成し，特徴量の値から新しいデータがどのクラスに属するかを推測することになる．判別関数を構成する考え方としては，確率分布に基づいて構成する方法と，確率分布を陽に考えず誤判別率などの評価関数を決めて最適化問題として解く方法があるが，本稿では前者を扱うことにする．

例 4.1. 年齢ごと (15,20,25,30 歳) の身長・体重から，50 歳のときに心臓病の有無 ($\circ$=“心臓病がない”，$\times$=“心臓病がある”) を調べたとする．既知のデータを使って判別関数

$$(\text{心臓病の有無}) = f(\text{年齢ごとの身長・体重})$$

を作ることによって，新規のデータに対して心臓病の有無を予測することを考える．

	15 歳		20 歳		25 歳		30 歳		50 歳
	身長 [cm]	体重 [kg]	身長 [cm]	体重 [kg]	身長 [cm]	体重 [kg]	身長 [cm]	体重 [kg]	心臓病
A	150	41	170	65	170	67	170	72	$\circ$
B	165	50	166	55	166	55	166	55	$\times$
C	$\cdots$							$\cdots$	$\times$
$\vdots$	$\vdots$	$\vdots$							$\vdots$
Z	160	50	163	60	170	60	170	64	?

4.1.2 観測データ

取り扱うデータは特徴量とクラスラベルの n 組で，i 番目のデータを $(\boldsymbol{x}_i, c_i)$ で表すことにする．

特徴量 一般に幾つあってもよい．以下では p 次元とし，$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathsf{T}}$ で表す．

クラスラベル データが属するクラスを表す．最も基本的 (単純) な場合は 2 つのクラスを考える (2 値判別問題)．2 値から多値への拡張の方法はいろいろあるが，講義では 2 値の場合のみを扱う．

4.1.3 確率モデル

クラスによらず特徴量の分布が同じ場合には，特徴量からクラスを見分けることはできないので，ここでは各クラス毎に，特徴量が異なる分布に従っていることを仮定する．分布としては様々なものが考えられるが，後に述べる線形判別関数と 2 次判別関数の構成を考える際には特徴量の分布として多次元正規分布を仮定した計算を行う．

判別の手続きの中では以下の 2 つの条件付確率を考える必要がある．

- データから各クラスに属する特徴量の確率モデルを推定する．

  $$(\text{クラスで条件付けた特徴量の分布}) = P(X = \boldsymbol{x}|C = c)$$

- 特徴量の値からどのクラスである確率が大きいかを考える．

  $$(\text{特徴量で条件付けたクラスの分布}) = P(C = c|X = \boldsymbol{x})$$

特徴量の条件付確率 $P(X = \boldsymbol{x}|C = c)$ からクラスの条件付確率 $P(C = c|X = \boldsymbol{x})$ を求めるためには，次節に詳しく述べるように Bayes の定理 (公式) を用いる．

4.1.4 事後確率

特徴量が与えられたときに計算されるクラスの条件付確率を**事後確率** (posterior probability) と呼ぶ．

例として以下の単純な場合を考えてみよう．

例 4.2. まず，心臓病になる人の比率を

50 歳のときに心臓が良い人　50%
悪い人　50%

と仮定する．これは確率分布として

$$P(C = \circ) = \frac{1}{2}, \qquad P(C = \times) = \frac{1}{2}$$

と表される．

それぞれのクラスに属するデータから特徴量の条件付確率 $P(X|C)$ を推定する．

心臓が良い人のデータ: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$　$\Rightarrow$　$P(X|C = \circ)$
心臓が悪い人のデータ: $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n$　$\Rightarrow$　$P(X|C = \times)$

判別分析においては，クラスの周辺確率 (特徴量とクラスの同時確率から特徴量を積分消去して周辺化した確率) を**事前確率** (prior probability) と呼ぶ．これは特徴量が与えられる前に予測できるクラスの確率である．クラスで条件付けた特徴量の条件付確率 $P(X|C)$ から事後確率 $P(C|X)$ を求めるには以下のよう *Bayes* **の定理** (Bayes' theorem) を用いればよい．

特徴量 X とクラス C の同時分布は，クラスの周辺分布 (事前確率) とクラスで条件付けた特徴量の分布の積で

$$P(X, C) = P(C)P(X|C)$$

と表される．これを用いると，特徴量 X が与えられたときのクラス C の確率である事後確率 $P(C|X)$ は Bayes の定理により

$$P(C|X) = \frac{P(X,C)}{P(X)} = \frac{P(X,C)}{\sum_C P(X,C)} = \frac{P(C)P(X|C)}{\sum_C P(C)P(X|C)}$$

で計算される．判別は 2 つの事後確率の大きさを比べて行えばよい．

例 4.3. 2 値判別の問題では 2 つの事後確率

$$P(C = \circ|X = \boldsymbol{x})$$
$$P(C = \times|X = \boldsymbol{x})$$

をそれぞれ直接計算する必要はなく，その大小関係を知るには比のみを計算すればよい．例えば上記の例では事前確率が各クラス 1/2 であることを用いれば

$$\frac{P(C = \circ|X = \boldsymbol{x})}{P(C = \times|X = \boldsymbol{x})} = \frac{P(C = \circ)P(X = \boldsymbol{x}|C = \circ)}{P(C = \times)P(X = \boldsymbol{x}|C = \times)}$$
$$= \frac{P(X = \boldsymbol{x}|C = \circ)}{P(X = \boldsymbol{x}|C = \times)} \gtrless 1, \qquad \begin{cases} > 1, & C = \circ \\ < 1, & C = \times \end{cases}$$

のように，条件付確率の比と 1 の大小関係で判別すれば良い．

より一般には $P(C = \circ) \neq P(C = \times)$ $(\neq \frac{1}{2})$ ($\circ$ と $\times$ の出現確率が 50%-50%でない) となるが，この場合には右辺の基準 1 を事前確率の比 (分子分母が逆となるので注意) で置き換えれば良い．すなわち

$$\frac{P(C = \circ)P(X = \boldsymbol{x}|C = \circ)}{P(C = \times)P(X = \boldsymbol{x}|C = \times)} \gtrless 1$$

あるいは事前分布を移項して

$$\frac{P(X = \boldsymbol{x}|C = \circ)}{P(X = \boldsymbol{x}|C = \times)} \gtrless \frac{P(C = \times)}{P(C = \circ)}, \qquad \begin{cases} >, & C = \circ \\ <, & C = \times \end{cases}$$

とすればよい．例えば $P(C = \times) = \frac{2}{3}$ の場合右辺の比の値は 2 となる．このルールで $\circ$ と判定されるのは，$\circ$ で条件付けた確率が $\times$ より 2 倍以上大きい場合だけなので，$\circ$ と判定され難くなることがわかる．

4.2 計算法

以下では，線形判別および 2 次判別の考え方を扱う．まずは特徴量 X が 1 次元で，各クラスで条件付けた分布が正規分布となる場合を想定し，具体的な判別ルールを構成する．その後，多次元の場合を考えることにする．

4.2.1 等しい分散を持つ 1 次元正規分布の場合

特徴量 X が 1 次元正規分布に従っているとする．1 次元の特徴量の例としては，例えば前節の例のデータのうち 30 歳のときの身長と体重に着目して BMI (Body Mass Index) を計算し，これを特

徴量 X として心臓病の有無の判別するといった場合を考えることができる．さらにこの 1 次元の特徴量が正規分布に従っていると仮定できる場合には，判別ルールは以下のように具体的に書き下すことができる．

2 つのクラス $(0/1, +1/-1, c_1/c_2$ 等何でもよいが，ここでは視覚的に区別しやすい $\circ/\times$ で表す) で条件付けたそれぞれの特徴量が，分散は同じであるが平均が異なる正規分布 (密度関数)

$$P(X = x|C = \circ) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)$$
$$P(X = x|C = \times) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right)$$

に従っているとする．事前確率が $P(C = \circ) = P(C = \times)$ となる場合には，特徴量 $X = x$ に対する事後確率の比は特徴量の条件付確率の比 $R(x)$

$$\begin{aligned} R(x) &= \frac{P(X = x|C = \circ)}{P(X = x|C = \times)} \\ &= \exp\left(-\frac{1}{2\sigma^2}\left\{(x-\mu_1)^2 - (x-\mu_2)^2\right\}\right) \\ &= \exp\left(\frac{1}{2\sigma^2}\left\{2x(\mu_1-\mu_2) - (\mu_1^2-\mu_2^2)\right\}\right) \end{aligned}$$

となるので，$R(x)$ が 1 より大きいか小さいか調べることによって判別を行うことができる．ここでさらに $F(x) = \log R(x)$ とすると

$$R(x)\begin{cases} > 1, & C = \circ \\ < 1, & C = \times \end{cases} \quad \Leftrightarrow \quad F(x)\begin{cases} > 0, & C = \circ \\ < 0, & C = \times \end{cases}$$

となるので，$F(x)$ の正負で判別を行うことができる．このとき

$$\begin{aligned} F(x) &= \frac{1}{2\sigma^2}\left\{2x(\mu_1-\mu_2) - (\mu_1^2-\mu_2^2)\right\} \\ &= \frac{\mu_1-\mu_2}{\sigma^2}\left(x - \frac{\mu_1+\mu_2}{2}\right) \end{aligned}$$

であるから，$\mu_1 - \mu_2 > 0$ のときには

$$\tilde{F}(x) = x - \frac{\mu_1+\mu_2}{2} \gtrless 0, \qquad \begin{cases} > 0, & C = \circ \\ < 0, & C = \times \end{cases}$$

$\mu_1 - \mu_2 < 0$ のときには

$$\tilde{F}(x) = x - \frac{\mu_1+\mu_2}{2} \lesseqgtr 0, \qquad \begin{cases} < 0, & C = \circ \\ > 0, & C = \times \end{cases}$$

によって判定すればよいことがわかる．

この判別ルールは x に関する 1 次 (線形) 式で表されるので**線形判別分析** (linear discriminant analysis) と呼ばれ，関数 $F(x)$ あるいは $\tilde{F}(x)$ を**線形判別関数** (linear discriminant function) という．

一方 $P(C=\circ) \neq P(C=\times)$ の場合には

$$\frac{P(C=\circ)}{P(C=\times)}R(x) \gtrless 1 \quad \Leftrightarrow \quad F(x) \gtrless \log\frac{P(C=\times)}{P(C=\circ)} = -\log\frac{P(C=\circ)}{P(C=\times)}$$

となるので,

$$\log\frac{P(C=\circ)}{P(C=\times)} = \gamma$$

とおいて

$$F(x) = \frac{\mu_1-\mu_2}{\sigma^2}\left(x - \frac{\mu_1+\mu_2}{2}\right) + \gamma$$

あるいは

$$\tilde{F}(x) = x - \frac{\mu_1+\mu_2}{2} + \frac{\sigma^2}{\mu_1-\mu_2}\gamma$$

を判別関数として判定すればよい．右辺の γ を含んだ項は，$\circ$ と $\times$ の出現頻度の偏りに応じて $\circ$ が多ければ $\circ$ と判定しやすく，$\times$ が多ければ $\times$ と判定しやすくなるようなバイアスとして働く．

4.2.2 異なる分散を持つ 1 次元正規分布の場合

2 つのクラスが平均・分散ともに異なる正規分布

$$P(X=x|C=\circ) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)$$
$$P(X=x|C=\times) = \frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)$$

に従う場合にも同様に計算できる．$P(C=\circ)=P(C=\times)$ のとき条件付確率の比 $R(x)$ は

$$\begin{aligned} R(x) &= \frac{\sigma_2}{\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}\right) \\ &= \frac{\sigma_2}{\sigma_1}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}+\frac{x-\mu_2}{\sigma_2}\right)\left(\frac{x-\mu_1}{\sigma_1}-\frac{x-\mu_2}{\sigma_2}\right)\right) \end{aligned}$$

となる．判別関数はこれの対数を取ることにより得られ

$$\begin{aligned} F(x) &= \log R(x) \\ &= \log\frac{\sigma_2}{\sigma_1} - \frac{1}{2}\left\{\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{(x-\mu_2)^2}{\sigma_2^2}\right\} \end{aligned}$$

となるので, x に関する 2 次式となる. これを *2* **次判別関数** (quadratic discriminant function) と呼ぶ．なお事前分布に偏りがある場合は線形判別と同様にバイアス γ を加えればよい．

4.2.3 等しい分散を持つ多次元正規分布の場合

X が p 次元の変量の場合には多次元の分布を考える必要がある．多次元正規分布の密度関数は

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

と書くことができる．ただし Σ は X の分散共分散行列で $p \times p$ の対称行列である．

ここで，クラス $C = \circ$ および $C = \times$ で条件付けた特徴量 X の分布が，ともに分散 Σ で，平均がそれぞれ $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ の多次元正規分布に従っているとする．計算はやや繁雑になるが，前節と同様にして判別関数は

$$F(\boldsymbol{x}) = \boldsymbol{a}^{\mathsf{T}}\boldsymbol{x} + b$$

という形になることがわかる．ここで $\boldsymbol{a}$ は共通の分散共分散行列 Σ と各クラスの平均ベクトル $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ で決まるベクトルであり，b は分散行列，平均値，および事前確率で決まるスカラーである．

練習問題 (1)

4.2.4 異なる分散を持つ多次元正規分布の場合

クラス $C = \circ$ および $C = \times$ で条件付けた特徴量 X の分布がそれぞれ平均 $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$，分散 Σ_1, Σ_2 の多次元正規分布に従っているとする．判別関数は

$$F(\boldsymbol{x}) = (\boldsymbol{x} \text{ の } 2 \text{ 次形式}) + (\text{定数})$$

という形になる．

練習問題 (2)

4.3 分析の評価

4.3.1 誤り率

判別の良さを測る最も単純な基準は**誤り率** (error rate) で，

$$(\text{誤り率}) = \frac{(\text{誤って判別されたデータ数})}{(\text{全データ数})}$$

で定義される．

2 値判別問題の場合，誤り率はその誤り方によって更に詳細に分けて考えることができる．判別したいラベル (例えば「病気であること」) を陽性 (positive) とし，以下の 4 つを考える．

- 正しく陽性と判定: **真陽性** (true positive; TP)
- 誤って陽性と判定: **偽陽性** (false positive; FP) (第 I 種過誤)
- 誤って陰性と判定: **偽陰性** (false negative; FN) (第 II 種過誤)
- 正しく陰性と判定: **真陰性** (true negative; TN)

偽陽性および偽陰性は検定で広く用いられる第 I 種過誤と第 II 種過誤に相当する．また，判別分析の結果を評価するに，上記の分類でデータ数を以下のように並べた正誤表

		真値	
		陽性	陰性
予測値	陽性	真陽性 (true positive)	偽陽性 (false positive)
	陰性	偽陰性 (false negative)	真陰性 (true negative)

が用いられる．パターン認識や機械学習の分野では，これの転置にあたる表

		予測値	
		陽性	陰性
真値	陽性	真陽性 (true positive)	偽陰性 (false negative)
	陰性	偽陽性 (false positive)	真陰性 (true negative)

が用いられることが多く，これを**混同行列** (confusion matrix) または**誤差行列** (error matrix) という．

これらの数を元に判別の評価基準はいろいろ定義されているが，基本的な量としては以下のものが良く用いられる．

$$(\text{真陽性率}) = \frac{TP}{TP+FN} \qquad \text{(true positive rate)}$$

$$(\text{真陰性率}) = \frac{TN}{FP+TN} \qquad \text{(true negative rate)}$$

$$(\text{適合率}) = \frac{TP}{TP+FP} \qquad \text{(precision)}$$

$$(\text{正答率}) = \frac{TP+TN}{TP+FP+TN+FN} \qquad \text{(accuracy)}$$

なお真陽性率は**感度** (sensitivity) あるいは**再現率** (recall)，真陰性率は**特異度** (specificity)，正答率は**精度**とも呼ばれ，分野によって名称が異なる場合があるので注意すること．

また最近では再現率 (真陽性率) と適合率の調和平均を用いた *F*-**値** (F-measure, F-score)

$$F_1 = \frac{2}{1/(\text{再現率}) + 1/(\text{適合率})} = \frac{2TP}{2TP+FP+FN} \quad \text{(調和平均)}$$

$$F_\beta = \frac{\beta^2+1}{\beta^2/(\text{真陽性率}) + 1/(\text{適合率})} \qquad \text{(重み付き調和平均)}$$

や，*Matthews correlation coefficient* (MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

も用いられる．

練習問題 (3)

4.3.2 訓練誤差と予測誤差

判別関数は通常有限個の既知のデータを用いて構成するが，この既知のデータに対する誤り率を**訓練誤差** (training error)，未知のデータを用いて求めた誤り率を**予測誤差**または**汎化誤差** (predictive error, generalization error) と呼んで区別することがある．

判別関数を構成するために用いた有限個のデータによって計算された訓練誤差は真の誤り率より良くなることが多い．これは，特定の既知データの判別に特化している可能性があるためであり，これを**過適応** (over-fitting) あるいは**過学習** (over-training) と言う．実際のデータ解析においては未知データに対する判別性能が重要であるから，予測誤差を用いて評価すべきである．

一般に予測誤差を評価するためには，判別関数を構成するための**訓練データ** (training data) とは別に，予測精度を評価するための**試験データ** (test data) が必要となる．このため，収集したデータを判別関数の推定と精度評価のための訓練データと試験データに分割して用いることがある．しかしながら，取得できるデータ数に限りがある場合は，データの分割の仕方に依存して予測誤差の評価が大きく変わってしまうことがある．こうしたデータ分割の偏りによる精度評価の偏りを避ける方法として**交叉検証法** (交差と書くこともある; cross-validation; CV) がある．交叉検証法ではまず n 個のデータを k ブロックにランダムに分割する．第 i ブロックを除いた $k-1$ ブロックのデータで判別関数を推定し，除いておいた第 i ブロックのデータで予測誤差を評価する．これを $i=1,\ldots,k$ で繰り返し，k 個の予測誤差を得た上で，その平均や分散などの集約統計量を用いて予測誤差を評価する．推定からは除き，予測誤差の評価に用いるデータを**検証データ** (validation data) と呼び，k 分割するこの方法を k-**重交叉検証法** (k-fold cross-validation; k-vold CV) と言う．この特殊な形として $k=n$ としたものがあり，これを *leave-one-out* **法** (leave-one-out cross-validation; LOO-CV) と言う．

4.3.3 ROC 曲線

判別問題では事前確率に大きな偏りを持つものを扱う場合が多く，この場合単純な誤り率だけでは評価が不十分なことがある．例えばある心臓病の発生確率が 0.1%だとしよう．どんなデータに対しても「心臓病でない」と判定すればこの判別の誤り率は 0.001(0.1%) ということになる．さらにこうした問題では，心臓病である人を心臓病でないと誤ることによる損失は，心臓病でない人を心臓病であると誤ることによる損失に比べて著しく大きく，両者を同列で比較することは好ましくない．事前確率の偏りは，理論的にはその比率によって補正されているが，これは誤りの損失までも考慮したものになっていない．

こうした問題における判別ルールの良さ，あるいは判別の難しさを測る基準の一つとして**受信者動作特性曲線** (receiver operating characteristic curve; ROC curve) がある．線形判別，2 次判別をはじめとして，2 値判別ルールは

$$F(x)-c\begin{cases}>0, & C=\circ\\ <0, & C=\times\end{cases}$$

と書けるものが多い．一般に c は特徴量の分布の母数や事前分布などから適切なものが与えられるが，c を自由に動かした上で次の二つの量を考える．

$$\mathrm{TPR}(c) = (C = \circ \text{を正しく判別した比率})$$
$$\mathrm{FPR}(c) = (C = \times \text{を誤って判別した比率})$$

ここで TPR は**真陽性率** (true positive rate)，FPR は**偽陽性率** (false positive rate) である．また，定義から陽性率も偽陽性率もクラス事前分布によらないことに注意する．可能な範囲で c を動かし，$\mathrm{FPR}(c)$ を x 座標，$\mathrm{TPR}(c)$ を y 座標として描かれる曲線が ROC 曲線となる．

簡単な場合として同じ分散を持つ 2 つのクラス条件付正規分布を考える．2 つの分布が同じ平均を持ち，全く重なっている場合には c によらず $\mathrm{TPR}(c) = \mathrm{FPR}(c)$ となるので，ROC 曲線は $(0,0)$ と $(1,1)$ を結ぶ直線となる．一方 2 つの平均が大きく隔っている場合には，$\mathrm{TPR}(c) > 0$ となる c ではほぼ $\mathrm{FPR}(c) = 0$ が成り立ち，逆に $\mathrm{FPR}(c) > 0$ となる c では $\mathrm{TPR}(c) = 1$ が成り立つので，ROC 曲線は $(0,0)$, $(0,1)$, $(1,1)$ を順に結んだ折れ線となる．一般に ROC 曲線は $(0,0)$ と $(1,1)$ を結ぶ右肩上りの曲線となるが，この曲線と x 軸で囲まれた面積が広いほど良い判別を行うことができると考えられる．この面積のことを AUC (area under the ROC curve) という．

4.4 解析の事例

4.5 補遺

本稿では判別分析の中で基本的な線形判別と 2 次判別を取り上げたが，これ以外にまだ書き切れていない項目には以下ようなものがある．

- Fisher の判別分析の基本的な考え方
- ロジスティック回帰による判別
- プロビット回帰による判別
- Support Vector Machine
- Jacknife, Bootstrap, Cross-Validation などのリサンプリングによる判別性能の評価

練習問題

(1) クラスで条件付けた特徴量の分布が，等しい分散を持つ多次元正規分布の場合の判別関数の具体的な形を求めよ．

(2) クラスで条件付けた特徴量の分布が，異なる分散を持つ多次元正規分布の場合の判別関数の具体的な形を求めよ．

(3) F-値や MCC が評価しようとしている判別方法の特性を，具体的な例 (どのような場合にこれらの値は高くなり，どのような場合に低くなるのか？) を考えながら説明せよ．

クラスタ分析 5

5.1 目的と考え方

5.1.1 目的

クラスタ分析 (cluster analysis) は，データ間に定義された距離に基づいて近いデータ同士を同じグループ (クラスタ) に属するものとして分類していく方法である．

グループの構成の仕方は大きく二種類に分けることができる．

- **階層的方法** — データ点およびグループの間に距離を定義し，近いものから順にクラスタを形成しながら，あるいは近いもの同士がクラスタ内に残るように分割しながら，グループ化していく方法．
  代表的なものとして，デンドログラム (樹形図) を作る凝集的クラスタリングがある．
- **非階層的方法** — グループ内に含まれるデータから決められるグループの領域と，実際にそれに含まれるデータに矛盾が生じないようにグループ化していく方法．
  代表的なものとして k-**平均法** (k-means) がある．

本講義では階層的クラスタリングの中でも凝集的方法を扱う．

例 5.1. 価格帯や顧客の評価に応じて同様な特徴を持つレストランのグループ分けを行うことを考える．

店名	価格	サービス	いごこち	味	···
A	3	4	4	3	···
B	5	4	4	5	···
C	3	5	5	4	···
⋮					

例えば高田馬場のラーメン店のグループ化を考えてみよう．

5.1.2 考え方

凝集的階層クラスタリングを行う基本的な操作は以下のようになる．

1. データ・クラスタ間の距離を求める．
   - (クラスタに属さない) データ点とデータ点の距離
   - (クラスタに属さない) データ点とクラスタの距離
   - クラスタとクラスタの距離

   の測り方を決め，それぞれを求める必要がある．なお，通常1つのデータ点もクラスタと考え，データ点とクラスタの距離にはクラスタとクラスタの距離の測り方を適用する．

2. 最も近い 2 つ (データ点とデータ点，データ点とクラスタ，クラスタとクラスタのいずれの場合もあり得る) を統合し，新たなクラスタとする．

3. クラスタ数が目的の数になるまでに (1)，(2) の手続きを繰り返してデータを順次クラスタに統合していく．

5.2 データ間の距離

以下で扱うデータはベクトルとして表現されているとし，2 つのベクトル $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^p$ の距離を $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ で表す．また各ベクトルの成分は $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jp})^T$ で表すものとする．

代表的なデータ間の距離としては以下のようなものがある．

5.2.1 ユークリッド距離 (Euclidean distance)

一般的な平方距離．各成分の差の 2 乗和の平方根 (2 ノルム)．

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

5.2.2 最大距離 (maximum distance)

各成分の差の中の最大値．

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \max\bigl\{|x_{i1} - x_{j1}|, \ldots, |x_{ip} - x_{jp}|\bigr\}.$$

5.2.3 マンハッタン距離 (Manhattan distance)

格子状に引かれた路に沿って移動するときの距離．各成分の差の絶対値の和 (1 ノルム)．

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = |x_{i1} - x_{j1}| + \cdots + |x_{ip} - x_{jp}|.$$

5.2.4 キャンベラ距離 (Canberra distance)

原点付近での違いを強調するように各座標軸における位置を考慮してマンハッタン距離を修正した距離．各成分の絶対値の和に対する成分の差の絶対値の和．

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{|x_{i1} - x_{j1}|}{|x_{i1}| + |x_{j1}|} + \cdots + \frac{|x_{ip} - x_{jp}|}{|x_{ip}| + |x_{jp}|}.$$

ただし分母が 0 になる項は通常取り除く (0 として計算する)．

5.2.5 ミンコフスキー距離 (Minkowski distance)

ユークリッド距離を q 乗に一般化した距離．各成分の差の q 乗和の q 乗根 (q ノルム)．

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \bigl\{|x_{i1} - x_{j1}|^q + \cdots + |x_{ip} - x_{jp}|^q\bigr\}^{1/q}.$$

5.2.6 バイナリー距離 (binary distance)

質的変数を対象とした距離．各成分を 0 と 1 の 2 値とし，少なくともどちらか一方が 1 である成分数に対する一方のみが 1 である成分数の比率．

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \frac{\min(x_{i1}, x_{j1}) + \cdots + \min(x_{ip}, x_{jp})}{\max(x_{i1}, x_{j1}) + \cdots + \max(x_{ip}, x_{jp})}.$$

これらの距離は，データの性質に応じて適宜使い分ける必要がある．また，対象の各成分をそのまま用いてその物理的な意味合いを積極的に利用する場合もあるが，逆に大きさの違い，例えば長さの測定で単位の取り方を mm (ミリメートル) にするか cm (センチメートル) にするかなどによって分析結果が異なってしまうことを避けたい場合には，全ての成分が平均 0，分散 1 となるように正規化するなど，データを前処理する必要がある．

5.3 クラスタ間の距離

ここでは階層的クラスタリングにおいて用いられる代表的なクラスタ間の距離の定義の仕方を説明する．一般にクラスタ間の距離は，各クラスタに含まれるデータ点同士の距離をどのように使うかによって決められる．このとき，データ点の距離から陽に定義する方法と，クラスタを統合したときに成り立つクラスタ間の距離の関係を用いて再帰的に定義する方法がある．

以下ではいくつかのデータ点からなる 2 つのクラスタを

$$C_a = \{\boldsymbol{x}_i | i \in \Lambda_a\}, \quad C_b = \{\boldsymbol{x}_j | j \in \Lambda_b\}$$

としたとき，この 2 つのクラスタ間の距離を $D(C_a, C_b)$ で表すものとする．

まず各クラスタに含まれるデータ点同士の距離を用いてクラスタ間の距離を書き下すとともに，2 つのクラスタを統合した際に統合とは無関係な他のクラスタとの関係がどのようになるかを考える．後者のクラスタ間の関係は，クラスタ C_a, C_b，およびこれらを統合した $C_a \cup C_b$ と，他のクラスタ C_c との距離で記すことができるため，距離の直感的な意味を捉え易く，実際の計算には都合のよい場合が多い．

5.3.1 最短距離法 (単連結法, single linkage method)

2 つのクラスタに属する対象のうち，最も近い対象間の距離をクラスタ間の距離とする方法である．

$$D(C_a, C_b) = \min_{\boldsymbol{x}_i \in C_a,\, \boldsymbol{x}_j \in C_b} d(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

統合前後のクラスタ間の関係は

$$D(C_a \cup C_b, C_c) = \min\bigl\{D(C_a, C_c), D(C_b, C_c)\bigr\}$$

となる．

5.3.2 最長距離法 (完全連結法, complete linkage method)

2 つのクラスタに属する対象のうち，最も遠い対象間の距離をクラスタ間の距離とする方法である．

$$D(C_a, C_b) = \max_{\boldsymbol{x}_i \in C_a,\, \boldsymbol{x}_j \in C_b} d(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

統合前後のクラスタ間の関係は

$$D(C_a \cup C_b, C_c) = \max\bigl\{D(C_a, C_c), D(C_b, C_c)\bigr\}$$

となる．

5.3.3 群平均法 (average linkage method)

2 つのクラスタに属する対象間のすべての組み合わせの距離を求め，その平均値をクラスタ間の距離とする方法である．

$$D(C_a, C_b) = \frac{1}{|C_a||C_b|} \sum_{\boldsymbol{x}_i \in C_a,\, \boldsymbol{x}_j \in C_b} d(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

ただし，$|C_a|, |C_b|$ でそれぞれのクラスタ内の要素の数を表すこととする．

統合前後のクラスタ間の関係は

$$D(C_a \cup C_b, C_c) = \frac{|C_a| D(C_a, C_c) + |C_b| D(C_b, C_c)}{|C_a| + |C_b|}.$$

となる．

5.3.4 McQuitty 法 (McQuitty's method)

2 つのクラスタを統合した場合に，他のクラスタからの距離を統合前のクラスタ間の距離の単純平均で算出する方法．群平均法では統合するクラスタの要素数に応じた重みで新しい距離が計算されるが，この重みを要素数によらないものとしたと考えることができる．データ間の距離を用いて書き下すことは難しいため，統合後のクラスタ間の距離は以下の式で再帰的に計算することによって求める．

$$D(C_a \cup C_b, C_c) = \frac{D(C_a, C_c) + D(C_b, C_c)}{2}.$$

5.3.5 重心法 (centroid method)

各クラスタに含まれるデータ点の重心 (平均値) を用いてクラスタ間の距離を定義する方法．クラスタ C_a の重心を $\bar{\boldsymbol{x}}_{\boldsymbol{a}}$, C_b の重心を $\bar{\boldsymbol{x}}_{\boldsymbol{b}}$ とすると

$$D(C_a, C_b) = d(\bar{\boldsymbol{x}}_a, \bar{\boldsymbol{x}}_b)$$

であるので，この距離が最も小さい 2 つのクラスタを順次統合していくことになる．

重心 (平均値) そのものはユークリッド距離のもとで意味を持つ概念であり，データ間の距離をユークリッド距離で定義する場合には統合後の距離について

$$D(C_a \cup C_b, C_c)^2 = \frac{|C_a| D(C_a, C_c)^2 + |C_b| D(C_b, C_c)^2}{|C_a| + |C_b|} - \frac{|C_a||C_b| D(C_a, C_b)^2}{(|C_a| + |C_b|)^2}$$

という関係が成り立つ．このため統合後のクラスタ間の距離は，クラスタの重心を求めずに計算することができる．なお，データ間の距離さえ与えられていれば，この式を再帰的に用いればクラスタ間の距離を計算することができるため，データ間の距離がユークリッド距離でない一般の距離の場合にも重心法を適用することができる．

5.3.6 メディアン法 (median method)

重心法と同様にクラスタの代表点同士の距離を用いて，クラスタ間の距離を定義するが，統合されたクラスタの代表点をもとのクラスタの代表点の中点とする方法．中点も一般にはユークリッド距離のもとで意味を持つ概念であり，データ間の距離をユークリッド距離で定義する場合には統合後の距離について

$$D(C_a \cup C_b, C_c)^2 = \frac{D(C_a, C_c)^2 + D(C_b, C_c)^2}{2} - \frac{D(C_a, C_b)^2}{4}$$

という関係が成り立ち，これにもとづいて再帰的にクラスタ間の距離を計算する．これは重心法の式で $|C_a| = |C_b|$ とした場合と考えることもでき，重心法と同様にデータ間の距離が一般の距離の場合でもメディアン法を用いることはできる．

5.3.7 ウォード法 (Ward's method)

新しい対象がクラスタに加わる時，最も広がりが小さく抑えられるクラスタを最も近いクラスタとする方法．クラスタの広がりを評価するにはクラスタの重心から各データ点までのユークリッド距離の平方和を考える．このときクラスタ C_a と C_b を統合した際の重心を

$$\bar{\boldsymbol{x}}_{a \cup b} = \frac{|C_a| \bar{\boldsymbol{x}}_a + |C_b| \bar{\boldsymbol{x}}_b}{|C_a| + |C_b|}$$

と書き，ユークリッド距離を d で表すと，クラスタ間の距離は

$$\begin{aligned} D(C_a, C_b) &= \sum_{\boldsymbol{x}_k \in C_a \cup C_b} d(\boldsymbol{x}_k, \bar{\boldsymbol{x}}_{a \cup b})^2 - \sum_{\boldsymbol{i}_k \in C_a} d(\boldsymbol{x}_i, \bar{\boldsymbol{x}}_a)^2 - \sum_{\boldsymbol{j}_k \in C_b} d(\boldsymbol{x}_j, \bar{\boldsymbol{x}}_b)^2 \\ &= \frac{d(\bar{\boldsymbol{x}}_a, \bar{\boldsymbol{x}}_b)^2}{\frac{1}{|C_a|} + \frac{1}{|C_b|}} \end{aligned}$$

によって定義されるため，この距離が最も小さい 2 つのクラスタを統合する．

統合前後のクラスタ間の関係は

$$D(C_a \cup C_b, C_c) = \frac{|C_a| + |C_c|}{|C_a| + |C_b| + |C_c|} D(C_a, C_c) + \frac{|C_b| + |C_c|}{|C_a| + |C_b| + |C_c|} D(C_b, C_c) - \frac{|C_c|}{|C_a| + |C_b| + |C_c|} D(C_a, C_b)$$

となり，重心の計算を行わなくてもクラスタ間の距離は再帰的に計算することができる．このため，データ間の距離が一般の距離の場合にも，上の式を用いてウォード法を適用することができる．

5.4 分析の評価

獲得されたクラスタ構造を評価する指標はいくつか提案されている．

5.4.1 凝集係数

凝集的クラスタリングの場合に用いられる凝集係数 (agglomerative coefficient) は以下のように定義される．まず，データ $\boldsymbol{x}_i$ が最初に統合されたクラスタを C とするとき，データ $\boldsymbol{x}_i$ とクラスタ C の距離を d_i とする．

$$d_i = D(\boldsymbol{x}_i, C)$$

また，アルゴリズムの最後に統合された 2 つのクラスタ C', C'' の距離を D とする．

$$D = D(C', C'')$$

凝集係数 AC は $l_i = 1 - d_i/D$ の平均値として

$$AC = \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{d_i}{D}\right)$$

と定義される．定義より $0 \leq AC \leq 1$ となり，1 に近いほどクラスタ構造が明瞭であることが示唆される．上記の l_i をデータ毎に並べた棒グラフを banner plot と呼ぶが，凝集係数は banner plot の面積比と考えることもできる．

5.4.2 シルエット係数

推定したクラスタ構造を評価する指標としてはシルエット係数 (silhouette coefficient) がある．まず，データ $\boldsymbol{x}_i$ が含まれているクラスタに対して，$\boldsymbol{x}_i$ を除いたクラスタ C^1 とデータ $\boldsymbol{x}_i$ の距離を d_i^1 とする．

$$d_i^1 = D(\boldsymbol{x}_i, C^1 \setminus \boldsymbol{x}_i)$$

また，データ $\boldsymbol{x}_i$ が含まれているクラスタ以外で，$\boldsymbol{x}_i$ に一番近いクラスタ C^2 との距離を d_i^2 とする．

$$d_i^2 = D(\boldsymbol{x}_i, C^2)$$

シルエット係数 S_i は

$$S_i = \frac{d_i^2 - d_i^1}{\max(d_i^1, d_i^2)}$$

で定義される．定義より $-1 \leq S_i \leq 1$ となり，1 に近いほどそのデータは適切にクラスタリングされていることが示唆される．クラスタリング全体の良さを評価するには S_i の平均を用いればよい．

5.4.3 鎖効果

最短距離法や最長距離法は最初に一度だけ全ての 2 点間の距離を計算すれば良いので計算時間が短縮できるが，はずれ値などの影響を受けやすく，対象が順番に 1 つの大きなクラスタに統合されていく鎖効果と呼ばれる悪い判別が起こりやすい．一方，群平均法やウォード法は鎖効果が起こりにくいと言われるが，クラスタを作る度に距離の計算をし直す必要があるため，大規模なデータに対しては計算時間が増えることになる．

計算量はデータ数に依存するが，一般には群平均法やウォード法が良く用いられる．

5.5 解析の事例

6 多次元尺度構成法

6.1 多次元尺度構成法の目的と考え方

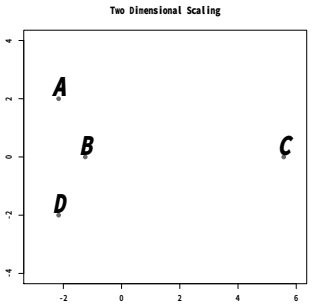
6.1.1 目的

多次元尺度構成法 (multi dimentional scaling; MDS) は類似度 (あるいは非類似度) が定義された対象を低次元の (わかりやすい) 空間に埋め込む方法で，データの可視化の一つの方法である．

例 6.1. データ間の距離が以下のように与えられていたとき，これらの距離関係の辻褄ができるだけ合うように各データを 2 次元空間に配置することを考える．

対象間の距離のデータ

	A	B	C	D
A	0			
B	3	0		
C	8	7	0	
D	4	3	8	0



2 次元空間における埋め込み

データ点を埋め込んだ空間での距離 (または類似性) が，元々与えられている距離 (類似性) の関係とできるだけ矛盾しないように各対象に適切な座標を割り当てることになる．ただし，回転や反転の不定性があることに注意する．

6.1.2 距離の定義

データ i と j に類似度 (similarity) s^*_{ij} あるいは非類似度 (dissimilarity) d^*_{ij} が与えられているとする．例えば，類似度の場合

$$s^*_{ij} > s^*_{ik} \text{ なら，} j \text{ の方が } k \text{ より } i \text{ に似ている}$$

ことを意味する．

非類似度の場合はそのまま距離として用いればよいが，類似度が与えられている場合は上記の関係から基準となる距離 d^*_{ij} を作る．基本的なルールとしては

$$s^*_{ij} = s^*_{kl} \Rightarrow d^*_{ij} = d^*_{kl}$$
$$s^*_{ij} < s^*_{kl} \Rightarrow d^*_{ij} \geq d^*_{kl}$$

を考える．つまり

- 類似度が同じなら，距離も同じ．
- 類似度が高い方が，距離が近いか同じ (遠くなることはない)．

を満たすように d^* を設定する．

6.1.3 計量 MDS (Torgerson の方法)

もともとデータはある次元のベクトルとして表され，データ間の距離が与えられている場合を想定する．

ここでは対象間の距離としてユークリッド距離を考える．対象は p 次元空間にあるとして 2 点を $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$，$\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jp})^T$ とすると，$\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ の距離は

$$d_{ij} = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$$

で表される．

ここで 2 点 $\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ の内積

$$z_{ij} = \sum_{k=1}^{p} x_{ik} x_{jk}$$

を考える．データの 1 つ (簡単のためデータの番号を 0 とする) を原点 (座標を $\boldsymbol{x}_0 = 0$ に設定) とすると，内積は以下のように距離で表される．

$$z_{ij} = \frac{1}{2} \left(d_{i0}^2 + d_{j0}^2 - d_{ij}^2 \right)$$

したがって，適当なデータを原点とすれば，与えられた距離 d^* からデータ間の内積を求めることができる．なお原点としては，他の点までの平均距離が最小となる中心点 (medoid)

$$\boldsymbol{x}_{\mathrm{med}} = \arg\min_i \sum_{j \neq i} d^*(i, j)$$

を用いることが多い．

データ数が n のとき，原点とするデータ 0 以外の点の座標を並べた行列

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_{n-1}^T \end{pmatrix}$$

を考えると，点 $\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ の内積 z_{ij} を ij 成分とする対称行列は

$$Z = (z_{ij}) = (\boldsymbol{x}_i^T \boldsymbol{x}_j) = XX^T$$

となるので, X を求めるには Z の分解を考えればよい. Z の Eckart-Young 分解 (固有値と固有ベクトルによる表現)

$$Z = S\Lambda S^T \quad S \text{ は直交，} \Lambda \text{ は対角}$$

を用いると X の 1 つの表現は

$$X = S\Lambda^{\frac{1}{2}}$$

となる．ただし，任意の直交行列を R とすると，

$$X' = S\Lambda^{\frac{1}{2}} R$$

としても $X'X'^T = XX^T = Z$ となるので，X には直交行列の不定性があり一意には決まらないことに注意する．

なお，データ数を n としたとき，この X は最大でも $n-1$ 次元の座標しか得られない．次元縮約を行うためには Λ の大きいものから必要な数だけ用いる．例えば 2 次元に縮約して表現する場合には，大きな固有値 2 つとそれに対応する固有ベクトルのみを用いて X を構成すればよい．

例 6.2. 都市の間の距離情報から座標を推定する．(解析の事例で詳しく述べる)

6.1.4 非計量 MDS (Kruskal の方法)

データ i に座標 $\boldsymbol{x}_i$ を与え，i と j の距離をミンコフスキー距離

$$d_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q \right)^{\frac{1}{q}}$$

で測ることを想定し，これが与えられた距離 d^* と出来るだけ矛盾しないように座標 $\boldsymbol{x}_i$ を求めることを考える．基準となる距離 d^* と割り当てた座標による距離 d との違いを評価するために以下で定義されるストレスを考える．

$$S = \sqrt{\frac{\sum_{i,j=1}^{n} \left(d_{ij} - d_{ij}^* \right)^2}{\sum_{i,j=1}^{n} d_{ij}^2}}$$

この評価量の分子および分母はそれぞれ以下の意味を持つ

- 分子は割り当てられた座標から計算される距離と基準の距離との差が小さいほど良い (類似度の関係をできるだけ再現するため).
- 分母は大きいほど良い (距離が 0 に縮退しないようにするため).

あとは S を最小とする $\{\boldsymbol{x}_i, i = 1, 2, \ldots, n\}$ を求める最適化の問題となる．

求められた座標が良いか悪いかの評価は達成されたストレスの値に基づいて行えばよいが，以下の Kruskal の判定基準が一般には用いられる．

S	評価
0.2	良くない
0.1	悪くはない
0.05	良い
0.025	非常に良い
0	完全

6.2 解析の事例

7 時系列解析

7.1 時系列のモデル

ある時間区間に渡り記録された観測値の系列を時系列と呼ぶ．単なる観測値の集合ではなく，時間軸に沿って並べられた観測データの関係性が重要であり，それをどのように記述するかが時系列解析の目的となる．

統計における時系列の分析では，何らかの確率過程 (時間を添字として持つ確率変数列; $X(t),\ t = \ldots, -2, -1, 0, 1, 2, \ldots$) を用いてデータをモデル化し，解析を行うことになる．代表的なモデルとしては，自己回帰過程 (AR)，移動平均過程 (MA), 自己回帰移動平均過程 (ARMA) などがある．以下にいくつかの簡単な確率過程をまとめる (定常過程・非定常過程の定義については次節を参照のこと).

7.1.1 白色雑音

最も単純な定常過程.

$$X(t) = \varepsilon(t)$$

ただし $\varepsilon(t)$ は以下の性質を満す確率変数である.

$$E(\varepsilon(t)) = 0$$
$$V(\varepsilon(t), \varepsilon(s)) = \sigma^2 \delta(t - s)$$

特に $\varepsilon(t)$ が正規分布に従うとき，白色ガウス雑音という．なお，以下の説明で現れる $\varepsilon(t)$ は白色ガウス雑音とする.

7.1.2 トレンドのある確率過程

平均値が時間とともに変動する典型的な非定常過程.

$$X(t) = \underbrace{\mu + \alpha \times t}_{\text{トレンド}} + \varepsilon(t)$$

ここでは 1 次式のトレンドを考えているが，高次の多項式や非線形関数 (指数関数，対数関数など) を考えることもある.

7.1.3 ランダムウォーク

分散が時間とともに変動する非定常過程の一つ.

$$X(t) = X(t-1) + \varepsilon(t)$$

ランダムウォークは最も単純な構造の非定常過程であるが，その階差 $Y(t) = X(t) - X(t-1)$ は定常過程になるので，時系列のモデルとしては基本的で重要なモデルである.

7.1.4 次数 1 の自己回帰過程 ($AR(1)$)

$|a| < 1$ なら定常過程，それ以外なら非定常過程.

$$X(t) = aX(t-1) + \varepsilon(t)$$

ランダムウォークを一般化したものとして捉えることができる.

7.1.5 自己回帰移動平均過程 ($ARMA(p,q)$)

係数に依存して定常・非定常の性質は変化する.

$$\begin{aligned} X(t) = a(0) + a(1)X(t-1) + \cdots + a(p)X(t-p) \\ + b(1)\varepsilon(t-1) + \cdots + b(q)\varepsilon(t-q) + \varepsilon(t) \end{aligned}$$

時系列の時間構造を考える上で基本となるモデルである．次数 p の自己回帰過程 (AR 過程) は $AR(p) = ARMA(p,0)$，次数 q の移動平均過程 (MA 過程) は $MA(q) = ARMA(0,q)$ である.

7.1.6 一般化自己回帰条件付分散変動過程 ($GARCH(p,q)$)

係数に依存して定常・非定常の性質は変化する.

$$\begin{aligned} X(t) &= \sigma(t)\varepsilon(t) \\ \sigma(t)^2 &= \alpha(0) + \alpha(1)X(t-1)^2 + \cdots + \alpha(q)X(t-q)^2 \\ &\quad + \beta(1)\sigma(t-1)^2 + \cdots + \beta(p)\sigma(t-p)^2 \end{aligned}$$

時系列の分散が，自己回帰モデルに従って変動するモデルで，金融時系列などのモデルに使われることが多い．なお GARCH は generalized autoregressive conditional heteroskedasticity の略語である．次数 q の自己回帰条件付分散変動過程 (ARCH 過程) は $ARCH(q) = GARCH(0,q)$ である.

例 7.1. 以下のようにしていくつかの確率過程を作ってみることができる (図 7.1 参照).

```
### いくつかの確率過程を擬似乱数で生成する

##
T <- 1000 # 時系列の長さ
## 白色ガウス雑音 (平均 0, 分散 1)
x.wn <- ts(rnorm(T,mean=0,sd=1))

## トレンドのある確率過程 X(t)=1+0.01*t+wn
x.tr <- ts(1+0.01*(1:T)+rnorm(T))

## ランダムウォーク
x.rw <- ts(cumsum(rnorm(T,mean=0,sd=1)))
## 漸化式に従って作る場合には以下のようにする
## x.rw <- double(T) # 長さ T の配列を確保
## x.rw[1] <- rnorm(1,mean=0,sd=1) # 初期値をランダムに生成
## for(t in 2:T) { # 漸化式に従って時系列を生成
##   x.rw[t] <- x.rw[t-1] + rnorm(1,mean=0,sd=1)
```

```
18  ## }
19  ## x.rw <- ts(x.rw) # ts class に変換
20  ##
21
22  ## AR(2),MA(2),ARMA(2,2)
23  a0 <- 0
24  a <- c(0.8,-0.5) # AR の係数
25  b <- c(0.6,0.3) # MA の係数
26  e <- rnorm(T)  # 雑音系列
27  x.arma <- double(T) # 長さ T の配列を確保
28  x.arma[1:2] <- rnorm(2,mean=0,sd=1) # 初期値をランダムに
  生成
29  for(t in 3:T) {
30    x.arma[t] <-
31        a0 + a %*% x.arma[t-(1:2)] + b %*% e[t-
  (1:2)] + e[t]
32  }
33  x.arma <- ts(x.arma)
34  ##
35  x.ar <- double(T)
36  x.ar[1:2] <- rnorm(2,mean=0,sd=1)
37  for(t in 3:T) {
38    x.ar[t] <- a0 + a %*% x.arma[t-(1:2)] + e[t]
39  }
40  x.ar <- ts(x.ar)
41  ##
42  x.ma <- double(T)
43  x.ma[1:2] <- rnorm(2,mean=0,sd=1)
44  for(t in 3:T) {
45    x.ma[t] <- b %*% e[t-(1:2)] + e[t]
46  }
47  x.ma <- ts(x.ma)
48
49  ## ARCH(2)
50  a0 <- 0.1
51  a <- c(0.5,0.2) # ARCH の係数
52  e <- rnorm(T)  # 雑音系列
53  x.arch <- double(T) # 長さ T の配列を確保
54  x.arch[1:2] <- rnorm(2,mean=0,sd=1) # 初期値をランダムに
  生成
55  for(t in 3:T) {
56    x.arch[t] <- e[t]*sqrt(a0+a %*% x.arch[t-(1:2)]^2)
57  }
58  x.arch <- ts(x.arch)
59
60  ## いくつか表示してみる
61  ts.plot(x.wn,x.tr,x.rw,x.arma,x.arch,col=2:6,
62          main="various random processes")
63  legend("topleft",inset=.05,
64         legend=c("white noise","trend","random walk",
65             "ARMA(2,2)","ARCH(2)"), col=2:6,lty=1,lwd=2)
66  ## 同じ雑音系列による AR,MA,ARMA
67  ts.plot(x.arma,x.ar,x.ma,col=2:4,
68          main="random processes with identical noise")
69  legend("topleft",inset=.05,
70         legend=c("ARMA(2,2)","AR(2)","MA(2)"),
71         col=2:4,lty=1,lwd=2)
72  ## 自己相関関数を比較
73  opar <- par(mfrow=c(3,1))
74  acf(x.arma)
75  acf(x.ar)
76  acf(x.ma)
```

```
77  par(opar)
78  ## lag plot を比較
79  lag.plot(x.arma,lags=4)
80  lag.plot(x.ar,lags=4)
81  lag.plot(x.ma,lags=4)
```

(tsa-ex1.r)

7.2 定常性

7.2.1 弱定常性と強定常性

時系列の性質を考える上で重要な概念の 1 つに定常性がある.

定義 7.2. 確率過程 $X(t)$ の分布が時刻 t によらないとき，強定常であるという.

強定常性の概念は自然ではあるが，相関などの 2 次の統計量に基づく実際の分析においては強すぎる条件であるので，通常は以下の弱い意味での定常性を用いる.

定義 7.3. 確率過程 $X(t)$ が

$$
\begin{aligned}
&E(X(t)) = \mu, && \text{(平均が時間によらない)}\\
&V(X(t)) = \sigma^2, && \text{(分散が時間によらない)}\\
&V(X(t), V(s)) = \gamma(t-s), && \text{(共分散は時差のみによる)}
\end{aligned}
$$

という性質を持つとき弱定常であるという.

非定常な系列は時間とともに確率過程を記述するパラメタが変化して扱いが難しくなる. 例えば，先に述べたようにランダムウォークは単純な構造の非定常過程であるが，以下のような実験で非定常性が弊害を持つことが確かめられる.

例 7.4. 2 つのランダムウォークは独立であるにも関わらず，多くの場合見せ掛けの強い相関を持つことが知られている.

```
1   ### 非定常系列における擬相関
2
3   ## 独立な 2 つのランダムウォークを作る
4   T <- 1000 # 時系列の長さ
5   rw1 <- ts(cumsum(rnorm(T)))
6   rw2 <- ts(cumsum(rnorm(T)))
7   ## 表示
8   ts.plot(rw1,rw2,col=c("red","blue"),
9           main="two independent random walks")
10  ## 相関係数を計算
11  cor(rw1,rw2)
12  ## 相関関数を表示
13  ccf(rw1,rw2,lag=500,
14      main="cross correlation of two random walks")
15  ## 差分の相関関数を表示
16  ccf(diff(rw1),diff(rw2),lag=500,
```

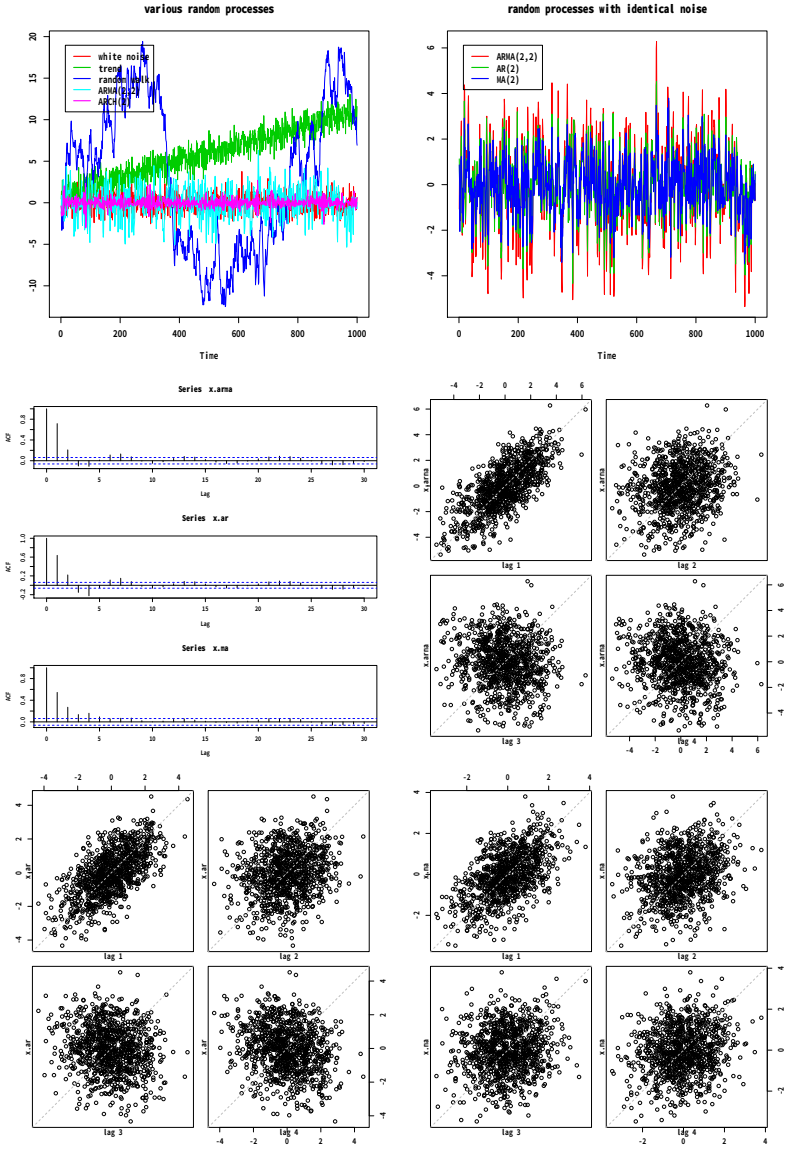


図 7.1: いろいろな確率過程.

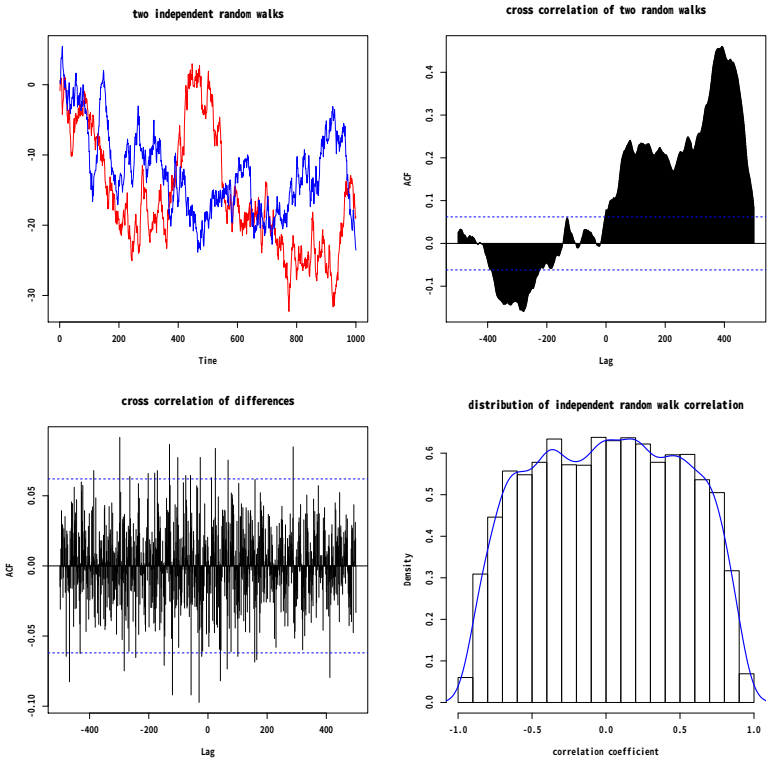


図 7.2: 2 つの独立なランダムウォーク (左上) とその相関関数 (右上)，差分の相関関数 (左下)，Monte Carlo 法により求めた相関係数の分布 (右下).

```
17      main="cross correlation of differences")
18
19  ## Monte Carlo 法により相関係数の分布を求める
20  ## (いくつものランダムウォークで相関係数を計算)
21  B <- 10000 # Monte Carlo 法の実験回数
22  rwcor <- double(B)
23  for (b in 1:B) {
24    rwcor[b] <- cor(cumsum(rnorm(T)),cumsum(rnorm(T)))
25  }
26  ## 相関関数の分布を表示
27  hist(rwcor,freq=FALSE,
28       main="distribution of independent random walk correlation",
29       xlab="correlation coefficient")
30  lines(density(rwcor),col="blue",lwd=2)
```

(tsa-ex2.r)

このように，非定常性な系列同士で相関などを考えてもあまり意味のない関係しか出てこないことがあるので，分析においては後述するように何らかの形で非定常な系列の中の定常な部分を切り出して解析を進める必要がある．こうした操作は定常化と呼ばれる．

7.3 時系列の定常化

7.3.1 差分による定常化

現実にある多くの時系列は非定常であるが，こうした時系列でもランダムウォークのような比較的単純な構造の非定常性である場合が多く，差分や対数差分を取ることによって定常な系列に (疑似的に) 変換することができる．

例えば時系列がランダムウォークの場合

$$X(t) = X(t-1) + \varepsilon(t)$$

と表されるが，確率過程 $\varepsilon(t)$ が定常であれば階差

$$Y(t) = X(t) - X(t-1) = \varepsilon(t)$$

が定常となることは明らかである．確率過程 $\varepsilon(t)$ がランダムウォークで表される非定常過程の場合にはさらにもう一度階差を取れば定常となることがわかる．このように必要であれば高階差分を行うことによって定常化が行われる場合もある．

また，現在の信号の大きさに比例して変動が加わる時系列

$$X(t) = X(t-1)(1 + \varepsilon(t))$$

を考えよう．確率過程 $\varepsilon(t)$ が定常な場合には，対数変換を行った上で階差を取ると

$$Y(t) = \log X(t) - \log X(t-1) = \log(1 + \varepsilon(t))$$

となり，$Y(t)$ は定常となることがわかる．ここで $\varepsilon(t)$ が十分小さいとすれば

$$Y(t) = \log(1 + \varepsilon(t)) \simeq \varepsilon(t)$$

と近似でき，対数差分によって確率過程 $\varepsilon(t)$ を取り出していると考えることができる．

例 7.5. 実際の金融データ (欧州・独株式指標) は経済成長とともに変化していく (図 7.3:左上) 典型的な非定常系列である．

```
### 差分・対数差分による定常化の例

## 欧州の市場データを利用する
data(EuStockMarkets)
dax <- EuStockMarkets[,1] # DAX のデータのみ取り出す
## データの表示
ts.plot(dax,col="blue",lwd=2,main="DAX stock index")
lag.plot(dax,9)
acf(dax,lag.max=250)

## 対数変換したデータの表示
ts.plot(log(dax), col="blue",lwd=2,
     main="DAX stock index (log transformed)")
lag.plot(log(dax),9)
acf(log(dax),lag.max=250)

```

```
17  ## 差分による定常化
18  ddax <- diff(dax)
19  ts.plot(ddax,col="red",main="differences of DAX")
20  lag.plot(ddax,9)
21  acf(ddax,lag.max=250)
22
23  ## 対数差分による定常化
24  dldax <- diff(log(dax))
25  ts.plot(dldax,col="red",main="differences of log(DAX)")
26  lag.plot(dldax,9)
27  acf(dldax,lag.max=250)
```

(tsa-ex3.r)

差分による定常化では，平均の変動分を取り除くことはできるが時間とともに大きくなる分散の変動を取り除くことはできない (図 7.4:左上の 3 つの図) が，対数変換したあとに差分をとることによって定常化されている (図 7.4:右下の 3 つの図) ことがわかる．これは各時刻の指標の大きさに比例するような変動が加わっているから (株価が金額の絶対値で変動するというより，株価の比率で変動すると考える方が自然) と考えることができる．

7.3.2 トレンドと季節成分の分解

時系列の中には，時間の発展とともに確定的に増減する変動を含んでいるものがあるが，こうした成分を一般にトレンドという．また，年間の日射量のように季節によって毎年ほぼ同じで，1 年周期で変動する成分が含まれているものがある．これを季節成分と呼ぶ．これらが組み合わされて

$$(\text{時系列}) = (\text{トレンド}) + (\text{季節成分}) + (\text{確率過程})$$

と表される現実の時系列は数多くある．

トレンドが時間の線形関数であれば差分によって，2 次関数であれば 2 階差分によって取り除くことができる．また季節成分は周期分ずらした時系列を平均化することによって推定し，これを引くことでほぼ取り除くことができる．解析においては，3 種類の成分に分解してから個々の成分の性質について考察し，分析を行う場合も多い．

例 7.6. 前例と同じ金融データおよび CO2 データをトレンド，季節成分，確率変動分に分解する．そのまま分解した場合と，対数変換して分解したものでは，結果が異なることに注意する．

```
1  ### 時系列の分解 (トレンド・季節成分・確率変動)
2
3  ## Daily Closing Prices of Major European Stock Indices, 1991
 -1998
4  data(EuStockMarkets)
5  dax <- EuStockMarkets[,1] # DAX のデータのみ取り出す
6  ts.plot(dax,col="darkgreen",lwd=2,main="DAX stock index") # デー
 タの表示
7
```

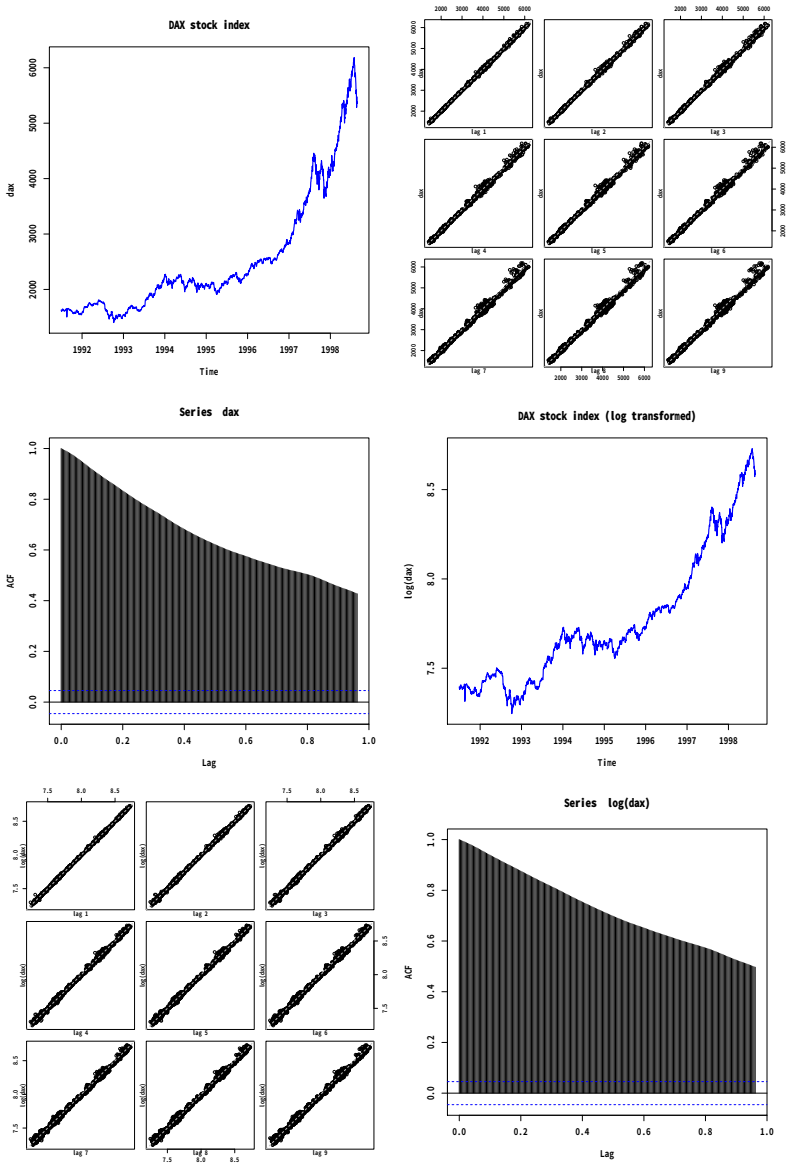


図 7.3: 差分・対数差分による非定常系列の定常化: DAX(独株式指標) データ，右上:対数変換したデータ.

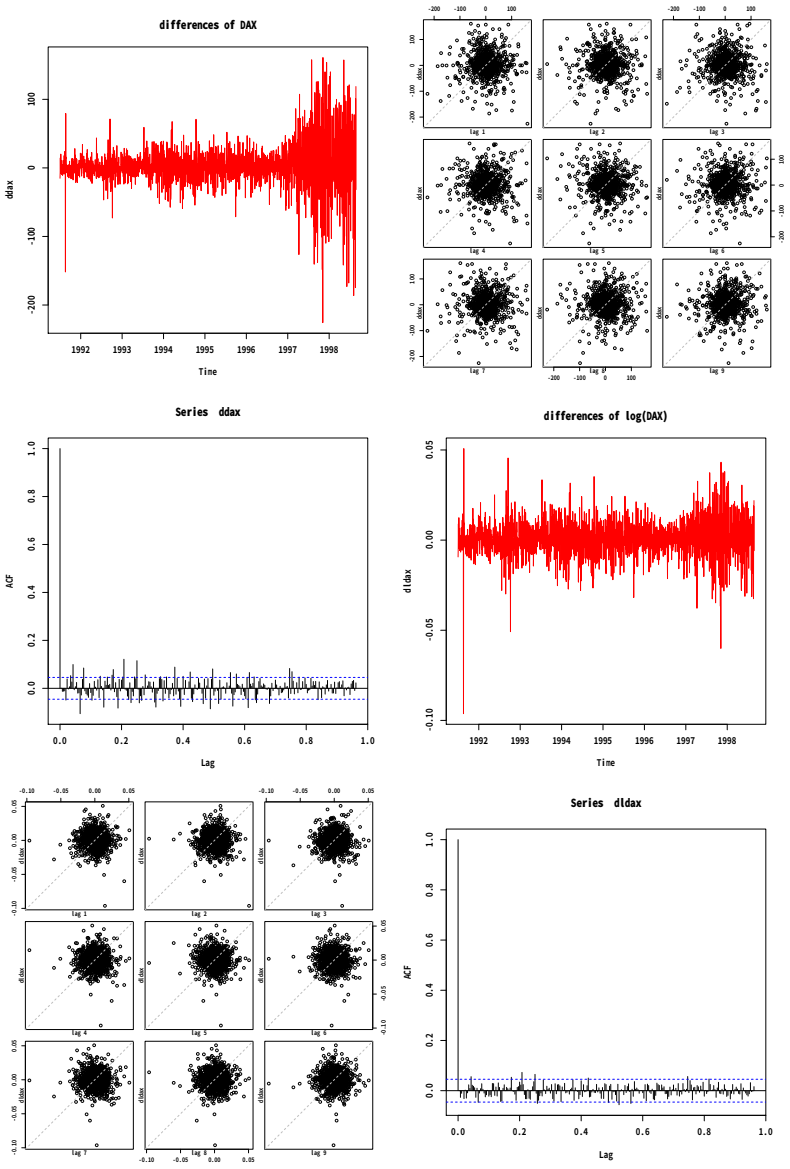


図 7.4: 差分による定常化，対数差分による定常化.

```
 8  ## 時系列の分解
 9  dax.dc <- decompose(dax)
10  ## summary(dax.dc) # 分解したデータ構造の確認する場合
11  ldax.dc <- decompose(log(dax)) # 対数変換をして表示
12  ## summary(ldax.dc)
13
14  ## 分解したトレンド成分の比較
15  ts.plot(dax.dc$trend,exp(ldax.dc$trend),col=c("blue","red"),lwd=2,
16          main="decomposed trends")
17  legend("topleft",inset=.05,
18         legend=c("trend of DAX","trend of log(DAX)"),
19         col=c("blue","red"),lty=1,lwd=2)
20
21  ## 分解したデータの表示
22  plot(dax.dc,col="darkgreen")
23  plot(ldax.dc,col="darkgreen")
24
25  ## Mauna Loa Atmospheric CO2 Concentration
26  data(co2)
27  ts.plot(co2,col="darkgreen",lwd=2,
28          main="Mauna Loa Atmospheric CO2 Concentration") # デー
  タの表示
29
30  ## 分解したデータの表示（別の方法）
31  plot(stl(co2,"periodic"),col="darkgreen")
32
33
```

(tsa-ex4.r)

7.4 確率過程の性質の検定

定常性などの時系列の性質を確認するには，視覚化と検定が重要である．各系列がどのような性質を持っているかを調べるために，以下のようないくつかの検定方法が提案されている．

7.4.1 単位根検定

時系列モデルとして $AR(1)$ を仮定したときに単位根 $(a_1 = 1)$ を持つかどうかを検定する方法を単位根検定という．単位根はランダムウォークに対応するので，ランダムウォーク様の非定常性を持つかどうかを検定していることになる．実際のデータにおいては $AR(1)$ という仮定は強すぎることが多いため，$AR(p)$ を仮定して $a_1 = 1$ を検定する

- Phillips-Perron 検定 (PP test)
- Augmented Dickey-Fuller 検定 (ADF test)

などが用いられる．

7.4.2 独立性の検定

時系列の各時刻での値が独立性を持つかどうか検定する方法として

- Box-Pierce 検定

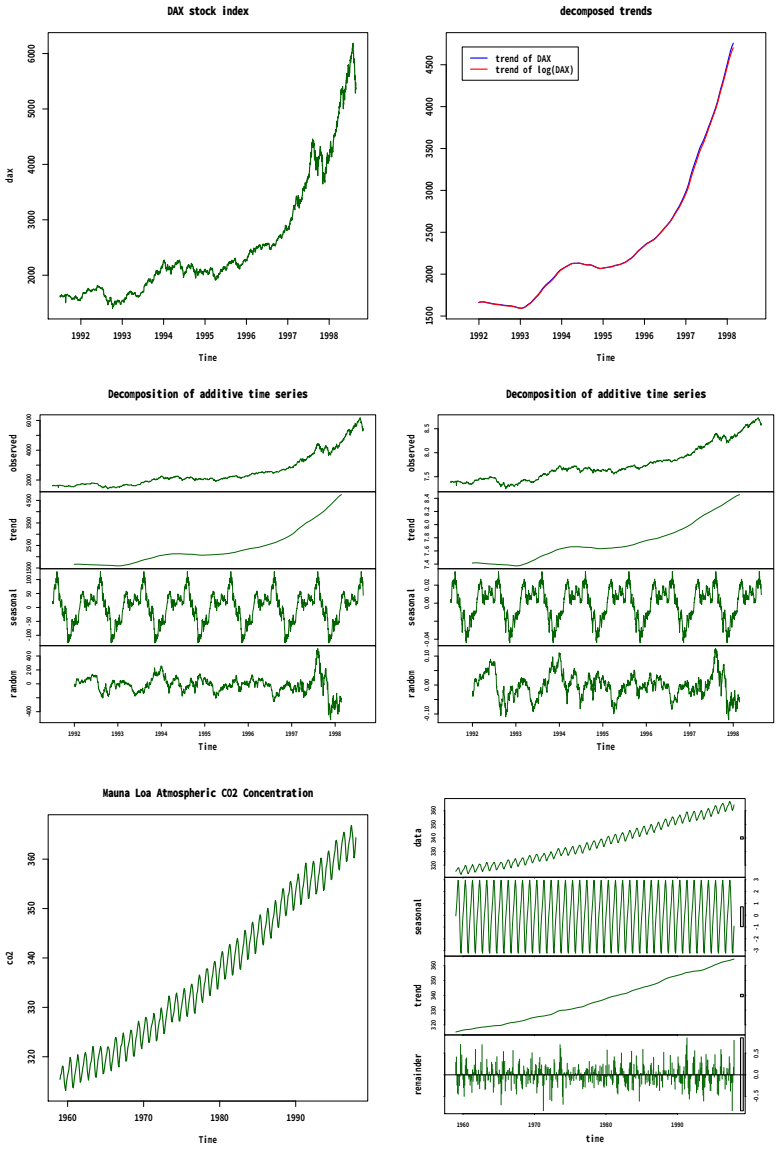


図 7.5: 非定常系列の分解．左上:DAX(独株式指標) データ，右上:単純な分解と対数変換後の分解によるトレンド成分の比較，左中:単純な分解結果，右中:対数変換後の分解結果，左下:1959 年から 1997 年までの月毎の CO_2 データ，右下:分解結果．

- Box-Ljung 検定

などがある．ARMA モデルを当て嵌めて推定した雑音の系列の白色性を確認するなどに用いる．

7.4.3 独立性の検定

時系列が正規性を持つかどうか検定する方法として

- Jarque-Bera 検定

がある．

7.4.4 定常性の検定

時系列が定常性を持つかどうか検定する方法として

- Kwiatkowski-Phillips-Schnidt-Shin 検定 (KPSS test)

がある．単位根検定とは異なり，帰無仮説は定常な系列であることである．またトレンドありの定常性 (平均は変動するが，分散が一定) の検定も行うことができる．

例 7.7. パッケージ `tseries` を用いると，確率過程の基本的な性質を調べるこれらの検定を行うことができる．

```
### 確率過程の性質の検定

## パッケージの読み込み
require(tseries)

## いくつかの確率過程を用意
T <- 200 # 時系列の長さ
wn <- ts(rnorm(T)) # 白色雑音 (正規分布)
wn2 <- ts(c(rnorm(T/2),runif(T/2))) # 正規分布から一様分布に切り替わる
wn3 <- ts(runif(T)) # 白色雑音 (一様分布)
rw <- ts(cumsum(rnorm(T))) # ランダムウォーク
tr <- 0.1*(1:T)+rnorm(T) # トレンドのある過程
ts.plot(wn,wn2,wn3,rw,tr,col=2:6)
legend("topleft",inset=.05,
       legend=c("white (Gauss)","white (switch)","white (unif)",
         "random walk","trend"),col=2:6,lty=1,lwd=2)

## 単位根検定 (ランダムウォークの検定)
## Phillips-Perron 検定
pp.test(wn) # 単位根を持たない (ランダムウォークではない)
pp.test(rw) # 単位根を持つ (ランダムウォーク)
## Augmented Dickey-Fuller 検定
adf.test(wn) # 単位根を持たない (ランダムウォークではない)
adf.test(rw) # 単位根を持つ (ランダムウォーク)

## Box-Pierce/Box-Ljung 検定 (独立性の検定)
Box.test(wn,lag=1) # 独立性を受容 (各時刻で独立)
Box.test(rw,type="Ljung") # 独立性を棄却 (独立でない)

## Jarque-Bera 検定 (正規性の検定)
jarque.bera.test(wn) # 正規性を受容 (正規ノイズ)
```

```
32  jarque.bera.test(wn3) # 正規性を棄却（一様ノイズ）
33  jarque.bera.test(rw) # 正規性を受容（正規ノイズだが非定常）
34
35  ## Kwiatkowski-Phillips-Schnidt-Shin 検定（定常性の検定）
36  kpss.test(wn) # 定常性を受容（定常過程）
37  kpss.test(wn2) # 定常性を棄却（定常過程）
38  kpss.test(rw) # 定常性を棄却（分散が変動する非定常過程）
39  kpss.test(tr) # 定常性を棄却（平均が変動する非定常過程）
40  kpss.test(tr,null="Trend") # トレンドを除けば定常（帰無仮説がトレンドあり）
```

(tsa-ex5.r)

7.5 解析の事例

7.5.1 AR モデルの適用例

自己回帰モデル $AR(p)$

$$X(t) = a(1)X(t-1) + \cdots + a(p)x(t-p) + \varepsilon(t),$$

を推定するためには，パッケージ stats の中に関数 ar が用意されている．ここではデータ EuStockMarkets — Daily Closing Prices of Major European Stock Indices, 1991–1998 — の中の DAX データ (独株式指標) を例として解析と予測を行う．元のままのデータのみならず，対数を取る，差分を取るなどさまざまな前処理を行った例を示す．

```
1   ### AR モデルの適用例
2   ### - Daily Closing Prices of Major European Stock Indices, 1991-1998
3
4   require(tseries)
5
6   ## データの整理
7   ldax <- log(EuStockMarkets[,"DAX"]) # DAX データを取り出す
8   ld <- window(ldax, # 過去のデータを切り出す
9                start=c(1993,1),end=c(1997,260))
10  lp <- window(ldax, # 未来のデータを切り出す
11               start=c(1998,1),end=c(1998,260),extend=T)
12  ## 2 つのデータを並べて表示する
13  seqplot.ts(ld,lp,colx="blue",coly="red",main="DAX")
14
15  ## AR モデルの作成例
16  (ld.ar <- ar(ld,method="ols"))
17  attributes(ld.ar) # モデルの持っている属性の表示
18  ld.ar$order # 推定された次数
19  ld.ar$ar # AR 係数
20  ld.ar$aic # 次数選択に用いられた AIC の値
21  ## AIC の値をグラフとして表示
22  plot(as.table(ld.ar$aic),type="l",col="blue")
23
24  ## 推定されたモデルを用いて予測を行う
25  p1 <- predict(ld.ar,ld,n.ahead=length(lp))
```

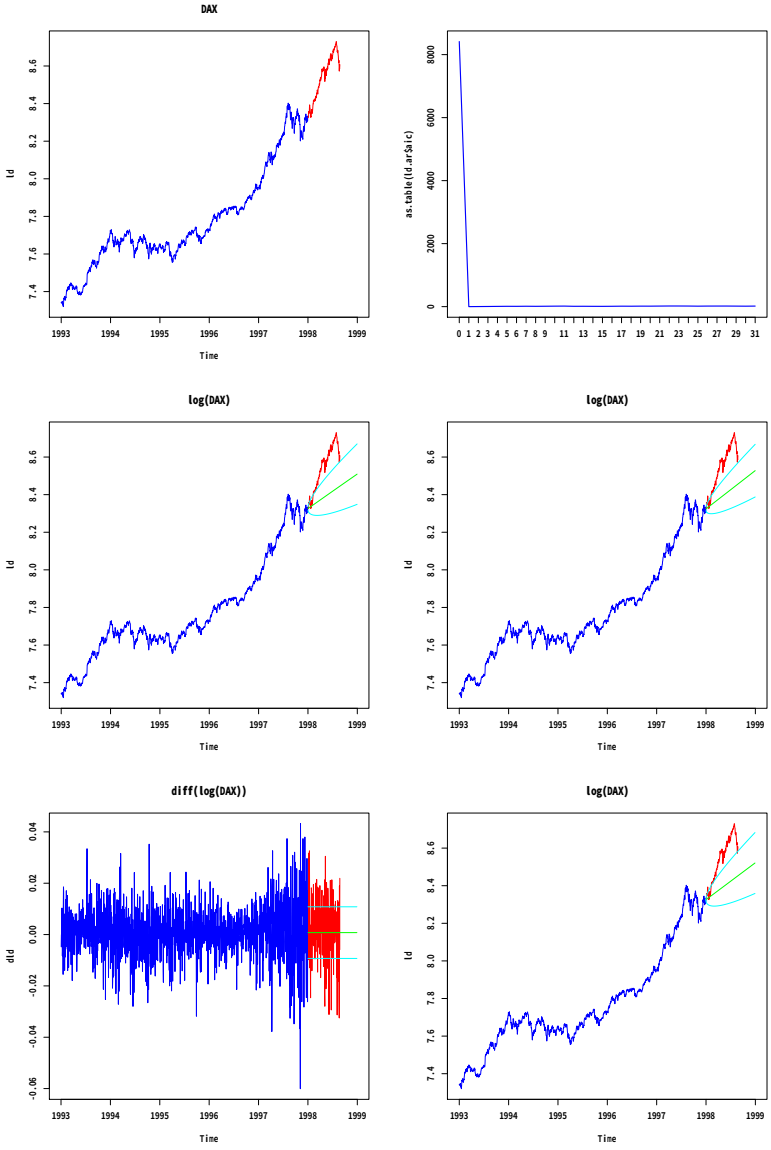


図 7.6: AR モデルによる解析の例 (その 1).

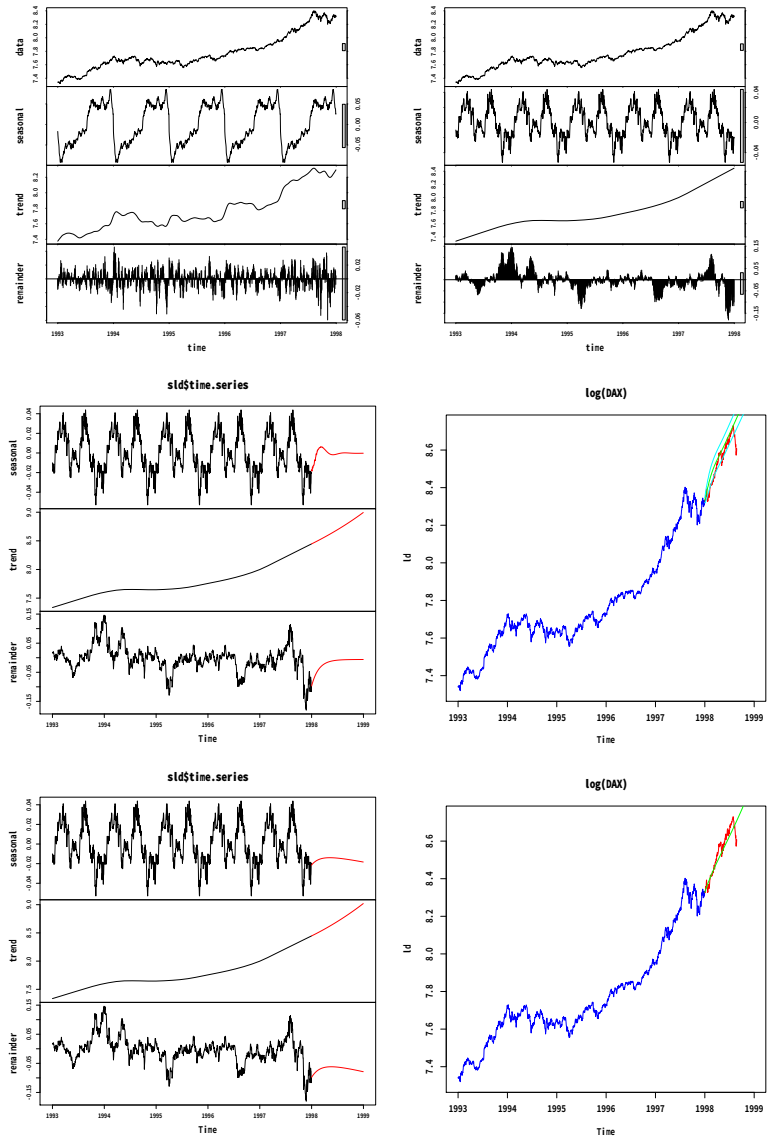


図 7.7: AR モデルによる解析の例 (その 2).

```
26  seqplot.ts(ld,lp,colx="blue", coly="red",main="log(DAX)")
27  lines(p1$pred,col="green") # 予測
28  lines(p1$pred+p1$se,col="cyan") # 予測の誤差 (1 sigma)
29  lines(p1$pred-p1$se,col="cyan")
30
31  ## 次数を指定してモデルを作る
32  (ld.ar10 <- ar(ld,method="ols",aic=F,order.max=10))
33  ld.ar10$ar # AR 係数
34  p2 <- predict(ld.ar10,ld,n.ahead=length(lp))
35  seqplot.ts(ld,lp,colx="blue", coly="red",main="log(DAX)")
36  lines(p2$pred,col="green") # 予測
37  lines(p2$pred+p2$se,col="cyan") # 予測の誤差 (1 sigma)
38  lines(p2$pred-p2$se,col="cyan")
39
40  ## 差分を取って定常化してモデルを作る
41  dld <- diff(ld)
42  dlp <- diff(lp)
43  (dld.ar <- ar(dld,method="mle"))
44  dld.ar$order # 選択された次数の確認
45  dld.ar$ar # AR 係数の表示
46  p3 <- predict(dld.ar,dld,n.ahead=length(dlp))
47  seqplot.ts(dld,dlp,colx="blue", coly="red",main="diff(log(DAX))")
48  lines(p3$pred,col="green") # 予測
49  lines(p3$pred+p3$se,col="cyan") # 予測の誤差 (1 sigma)
50  lines(p3$pred-p3$se,col="cyan")
51
52  ## 差分を元に戻して表示する
53  seqplot.ts(ld,lp,colx="blue", coly="red",main="log(DAX)")
54  p3pr <- diffinv(p3$pred,xi=ld[length(ld)]) # 差分を逆算
  する
55  p3se <- sqrt(diffinv(p3$se^2)) # 分散を累積する
56  lines(p3pr,col="green") # 予測
57  lines(p3pr+p3se,col="cyan") # 予測の誤差 (1 sigma)
58  lines(p3pr-p3se,col="cyan")
59  ## 時系列の分解 (stl() と decompose() がある)
60  plot(stl(ld,s.window=130, # 季節変動の長さ (半年) cf. 1 年
  =260 日
61            t.window=40, # トレンド推定の窓の長さ (8 週)
62            t.jump=1))
63  plot(sld <- stl(ld,"per",robust=T)) # 季節変動を 1 年周期
  とする分解
64  summary(sld)
65  sld$time.series[1:10,] # 最初の 10 日分のデータを表示
66  (se.ar <- ar(sld$time.series[,"seasonal"],method="ols"))
67  (tr.ar <- ar(sld$time.series[,"trend"],method="ols"))
68  (re.ar <- ar(sld$time.series[,"remainder"],method="ols"))
69  p4se <- predict(se.ar,sld$time.series[,"seasonal"],n.ahead=length(lp))
70  p4tr <- predict(tr.ar,sld$time.series[,"trend"],n.ahead=length(lp))
71  p4re <- predict(re.ar,sld$time.series[,"remainder"],n.ahead=length(lp))
72
73  ## 分解結果に予測を継いで表示
74  seqplot.ts(sld$time.series,ts.union(p4se$pred,p4tr$pred,p4re$pred))
75
76  ## 各予測結果を統合して表示
77  seqplot.ts(ld,lp,colx="blue", coly="red",main="log(DAX)")
78  p4 <- p4se$pred+p4tr$pred+p4re$pred
79  p4e <- sqrt(p4se$se^2+p4tr$se^2+p4re$se^2)
80  lines(p4,col="green") # 予測
81  lines(p4+p4e,col="cyan") # 予測の誤差 (1 sigma)
82  lines(p4-p4e,col="cyan")
83
84  ## 多次元確率過程モデルによる各成分の予測
```

```
msld.ar <- ar(sld$time.series,method="ols",order.max=5)
p5 <- predict(msld.ar,sld$time.series,
              n.ahead=length(lp),se.fit=F) # 不安定
seqplot.ts(sld$time.series,p5)

## 予測結果を統合して表示する
seqplot.ts(ld,lp,colx="blue", coly="red",main="log(DAX)")
p6 <- ts(rowSums(p5),start=start(p5),end=end(p5),
         frequency=frequency(p5))
lines(p6,col="green") # 予測
```

(tsa1.r)

7.6 トレンドと季節成分を含むデータの分析例

差分を d 回取って定常になる過程を次数 d の和分過程 (integrated process) と言うが，d 回差分が $ARMA(p,q)$ となる過程が自己回帰和分移動平均モデル $ARIMA(p,d,q)$ である．パッケージ `stats` には関数 `ar` のほかに関数 `arima` が用意されており，ARMA モデルを含むより一般的な ARIMA モデルの推定を行うことができる．

季節成分を考慮した ARIMA モデルは SARIMA モデルと呼ばれることもあるが，`arima` の `seasonal` オプションを用いることによって推定することができる．標準データセットに含まれる `AirPassengers` – Monthly Airline Passenger Numbers 1949-1960 – に対して

田中勝人，現代時系列分析，岩波書店 (2006)

の「ARIMA モデルと SARIMA モデル」の項の解析の流れに沿って行うと以下のようになる．

```
### ARIMA モデルによる AirPassengers データの予測

## AirPassengers データの読み込み
data(AirPassengers)
tsp(AirPassengers) # データの時間に関する情報を表示（月ごとのデータ）
plot(AirPassengers,col="blue") # データの表示
plot(log(AirPassengers),col="blue") # 対数変換データの表示
## 対数変換により分散変動が安定化していることがわかる

## 以下では対数変換したデータを扱う
ap <- log(AirPassengers) # 省略名 (ap) に代入しておく

## まずトレンドについて考察する
kpss.test(ap) # 定常か？-> 棄却される (当然定常でない)
pp.test(ap) # ランダムウォーク的な非定常性か？-> 棄却される
kpss.test(ap,null="Trend") # トレンドを除けば定常か？-> 棄却されない
## 非定常性の大きな要素はトレンドらしいことが予想される
plot(diff(ap),col="blue")
## 一回階差を取ることによりトレンドは除去されているように見える
kpss.test(diff(ap)) # 定常性の検定 -> 棄却されない
```

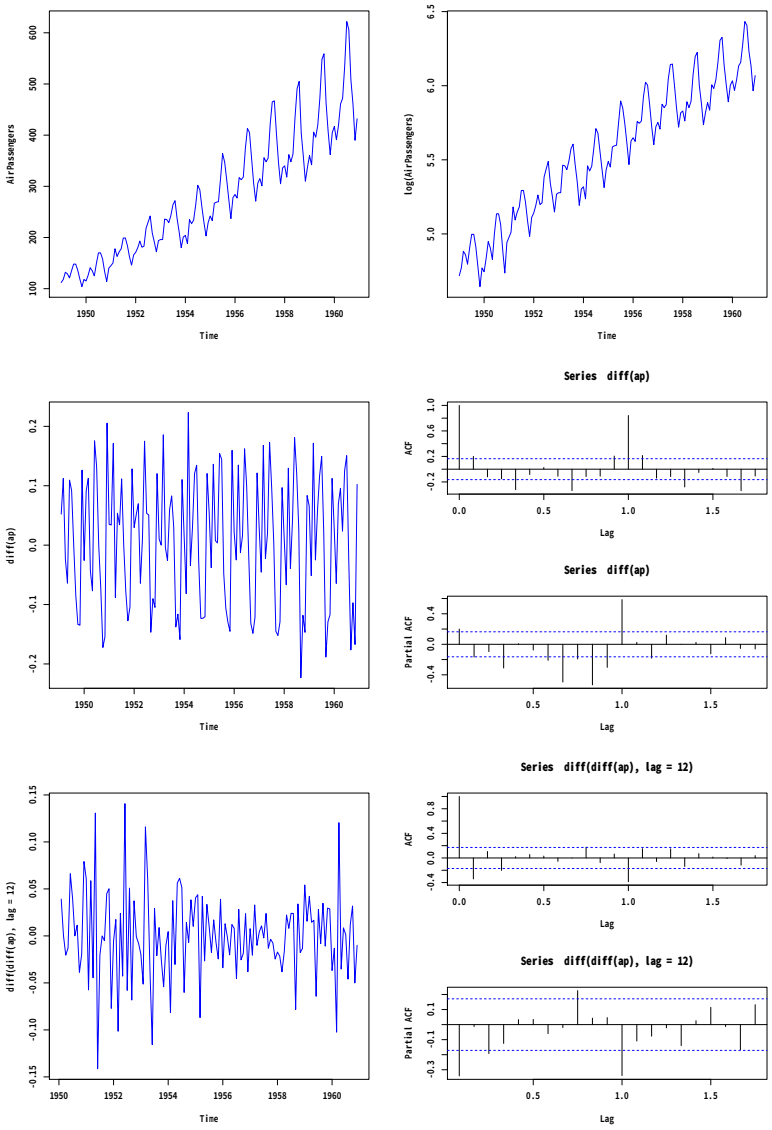


図 7.8: ARIMA モデルによる解析の例 (その 1).

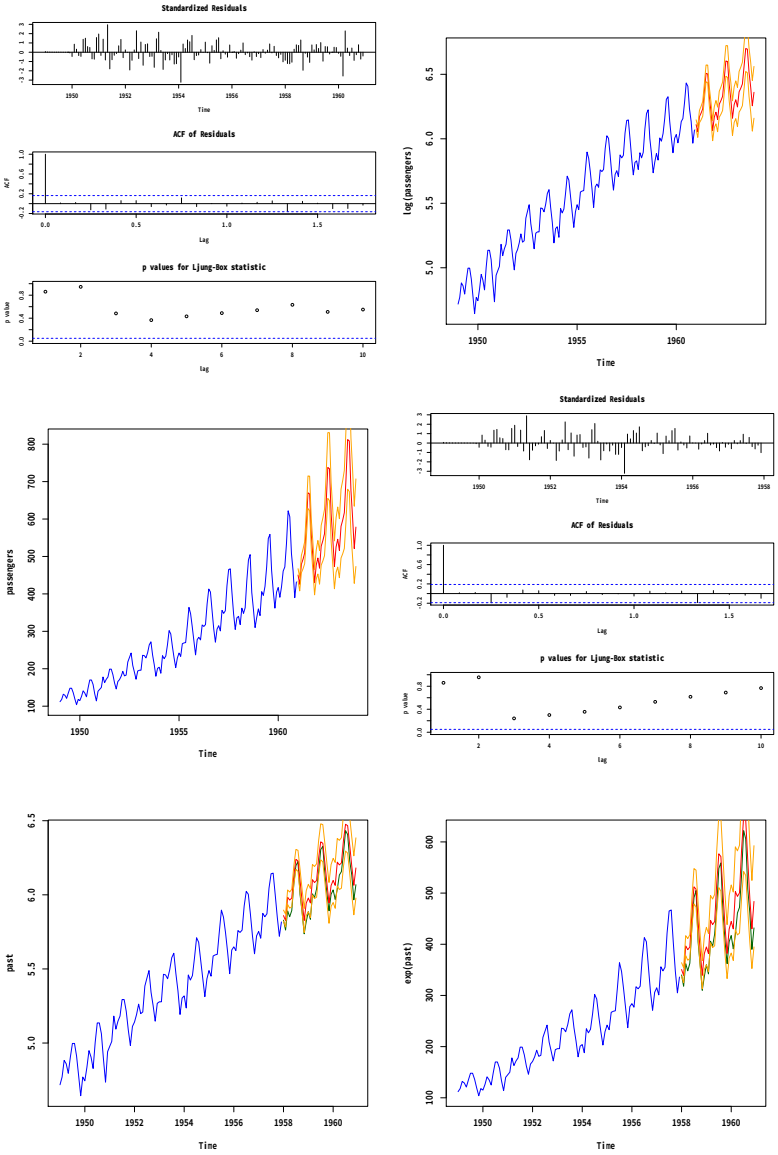


図 7.9: ARIMA モデルによる解析の例 (その 2).

```
21  ## 自己相関関数と偏自己相関関数を調べてみる
22  opar <- par(mfrow=c(2,1))
23  acf(diff(ap))
24  pacf(diff(ap))
25  par(opar)
26  ## lag=1(1 年) ごとに強い相関があることがわかる -> 季節成分
  の存在を示唆
27
28  ## 季節成分について考察する
29  ## 12 ヶ月で階差を取って同様に調べてみる
30  plot(diff(diff(ap),lag=12),col="blue")
31  opar <- par(mfrow=c(2,1))
32  acf(diff(diff(ap),lag=12))
33  pacf(diff(diff(ap),lag=12))
34  par(opar)
35  ## lag=1/12,3/12,1 (1 ヶ月, 3 ヶ月, 1 年) に相関が残っている
36  jarque.bera.test(diff(diff(ap),lag=12)) # 正規か？-> 棄
  却されない
37  ## 1 ヶ月と 12 ヶ月で階差を取った時系列は
38  ## 1 ヶ月と 3 ヶ月と 12 ヶ月に相関のある正規分布の系列と考える
  ことができる
39
40  ## ARIMA モデルの作成
41  ## diff(diff(ap),lag=12) =
42  ##   MA(12(た   だ   し   係   数   は{1,3,12}が   重
  要)), ARMA(1,12{1,3,12})
43  ## あたりを考える必要がありそう
44  ## 季節成分による ARMA 項の指定は seasonal オプションを用いる
45  ## 例えば seasonal=list(order=c(0,1,2),period=12) で
46  ## 差分= e(t) + b(12)*e(t-12) + b(24)*e(t-24) のモデルを
  当てはめる
47  ## なお, order と seasonal/order で指定する差分によって
48  ## 1 ヶ月階差と 12 ヶ月階差を取ることに注意する
49  arima(ap,order=c(0,1,1),seasonal=list(order=c(0,1,0),period=12))
50  arima(ap,order=c(0,1,2),seasonal=list(order=c(0,1,0),period=12))
51  arima(ap,order=c(0,1,3),seasonal=list(order=c(0,1,0),period=12))
52  arima(ap,order=c(0,1,1),seasonal=list(order=c(0,1,1),period=12))
53  arima(ap,order=c(0,1,2),seasonal=list(order=c(0,1,1),period=12))
54  arima(ap,order=c(0,1,3),seasonal=list(order=c(0,1,1),period=12))
55  arima(ap,order=c(1,1,0),seasonal=list(order=c(0,1,1),period=12))
56  arima(ap,order=c(1,1,1),seasonal=list(order=c(0,1,1),period=12))
57
58  ## これらの中で AIC 最小のモデルで予測を行う
59  model <- arima(ap,order=c(0,1,1),
60                 seasonal=list(order=c(0,1,1),period=12))
61  tsdiag(model) # 念のためモデルの診断図を見ておく
62  tmp <- predict(model,n.ahead=36)
63  pr <- tmp$pred # 予測値
64  se <- tmp$se # 予測誤差の標準偏差
65  seqplot.ts(ap,pr,colx="blue",coly="red",
66             ylab="log(passengers)")
67  lines(pr+se,col="orange");lines(pr-se,col="orange")
68  ## もとのデータの空間に戻してみる
69  seqplot.ts(exp(ap),exp(pr),colx="blue",coly="red",ylab="passengers")
70  lines(exp(pr+se),col="orange");lines(exp(pr-
  se),col="orange")
71
72  ## 実際の解析では上のような予測を行う訳であるが
73  ## 予測の可否を評価するために, データを分割して処理してみる
74  past <- window(ap,end=c(1957,12))
75  future <- window(ap,start=c(1958,1))
```

```
76   model <- arima(past,order=c(0,1,1),
77                  seasonal=list(order=c(0,1,1),period=12))
78   tsdiag(model) # 念のためモデルの診断図を見ておく
79   tmp <- predict(model,n.ahead=length(future))
80   pr <- tmp$pred # 予測値
81   se <- tmp$se # 予測誤差の標準偏差
82   seqplot.ts(past,future,colx="blue",coly="darkgreen")
83   lines(pr,col="red")
84   lines(pr+se,col="orange");lines(pr-se,col="orange")
85   ## もとのデータの空間に戻してみる
86   seqplot.ts(exp(past),exp(future),colx="blue",coly="darkgreen")
87   lines(exp(pr),col="red")
88   lines(exp(pr+se),col="orange");lines(exp(pr-
  se),col="orange")
```

(tsa5.r)

時系列解析に関する成書は数多く出ているので，様々な事例を参考にしながら実際の解析について習熟するのが望ましいと思われる．

索引