（要約） SAS による多変量データの解析

前川眞一
東京大学出版会 1997

本書の目的は、所謂矩形のデータ行列の分析のために SAS に用意されているプロシジャを、線形モデルと最小２乗法の観点から統一的に表現し、各方法の関連を明確にすることである。

1　多変量解析法と線形モデル

所謂 $N \times p$ のデータ行列（N 個の個体の p 変量による測定値）を取り扱う多変量解析の方法のうち、利用度の高いものには、回帰分析 (proc reg, glm, transreg)、因子分析 (proc factor)、主成分分析 (proc princomp, prinqual)、クラスター分析 (proc fastclus, cluster, varclus)、冗長性分析 (proc transreg)、等があるが、これらの方法は線形モデルおよび最小２乗法の観点から統一的に説明する事が可能である。

Table 1: 多変量解析法の分類

		説明変数 $\boldsymbol{X}$ は観測されているか	
		観測されている （顕在変数）	観測されていない （潜在変数）
説明変数 $\boldsymbol{X}$ は連続か離散か	連続	回帰分析 冗長性分析 (判別分析) (ロジスティク回帰)	因子分析 主成分分析 (IRT) （カテゴリカルデータの因子分析） （数量化 III 類, CA）
	離散	分散分析 (対数線形モデル)	クラスター分析 (潜在クラスモデル)

表中、括弧の中は、基準変数 $\boldsymbol{Y}$ がカテゴリカルな場合、それ以外は、基準変数が連続の場合を表す。 また、説明変数が観測されている場合、その連続離散の区別はあまり重要ではない。

一般に p 変量の多変量線形モデルは

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}' + \boldsymbol{E} \tag{1}$$

という形で書き表されるが、ここに、$\boldsymbol{Y}$ は $N \times p$ のデータ行列、$\boldsymbol{X}$ は $N \times q$ の説明変数行列、$\boldsymbol{\beta}$ は $p \times q$ の回帰係数行列、 $\boldsymbol{E}$ は $N \times p$ の残差行列である。説明変数行列として観測不可能な潜在変数を含ばこれらの方法は以下のように分類することが出来る。なお、表中に括弧を付した方法は基準変数 $\boldsymbol{Y}$ が離散変数の場合の方法である。また、基準変数の尺度水準に応じた最適を考えると、$\boldsymbol{Y}$ が順序尺度の場合の取り扱いも可能となる。

1. 回帰分析：$\boldsymbol{X}$ が連続な顕在変数、
   $\boldsymbol{x_i}$ は個体 i の説明変数の値、$\boldsymbol{\beta_j}$ は回帰係数
2. 分散分析：$\boldsymbol{X}$ がカテゴリカルな顕在変数、
   $\boldsymbol{x_i}$ は個体 i がどの要因の組み合わせの下で観測されたかを示すインディケイタ（デザイン行列）、$\boldsymbol{\beta_j}$ は各要因の水準の効果
3. 因子分析：$\boldsymbol{X}$ が連続な潜在変数、
   $\boldsymbol{x_i}$ は個体 i の因子スコア、$\boldsymbol{\beta_j}$ は変数 j の因子パタン
4. クラスター分析：$\boldsymbol{X}$ がカテゴリカルな潜在変数、
   $\boldsymbol{x_i}$ は個体 i がどのクラスターに属するかを示すインディケイタ、$\boldsymbol{\beta_j}$ は、各クラスターにおける変数 j の平均値

という具合になる。なお、通常の回帰分析や分散分析の場合には、推定されるべきパラメタは $\boldsymbol{\beta_j}$, j=1, ... , p であるが、説明変数が潜在変数の場合には、$\boldsymbol{X}$ をも推定することになる。

2 データの最適変換を含む多変量線形モデル

2.1 回帰分析

実際にデータを収集する場合、間隔尺度水準以上の変数を用いることは必ずしも可能であるとは限らない。たとえば、ある商品が好まれる理由をその商品の持つさまざまな特性を用いて説明しようとする場合を考えてみる。この場合、基準変数 $\boldsymbol{Y}$ としてはｎ個の商品の好みの程度をｐ人に評定させたものを用いることになるが、その評定値は果たして間隔尺水準の測定値として取り扱ってよいものであろうか。もしも、この評定値が単に好みの大小関係のみを示すものであれば、それは順序尺度水準の測定値であると呼ばれる。従って、 $y_{1j} = 2 < y_{2j} = 5 < y_{3j} = 6, \ldots$ というデータは、その順序を保つように、$y_{1j} = 6 < y_{2j} = 7 < y_{3j} = 8, \ldots$ と変換してもそれが持つ情報はなんら失われたことにならない。しかし、これらの評定値を商品の特性 $\boldsymbol{X}$ に回帰する場合、どちらのデータを用いたかにより回帰係数は一般に異なったものとなる。また、商品の特性そのものも、順序尺度や名義尺度水準の評定値として得られているかもしれない。

このような順序尺度もしくは名義尺度水準のデータを用いて回帰分析を行うための方法として、SAS/STAT の PROC TRANSREG があるが、そこでは通常の回帰分析のモデルを修正した、以下のようなモデルが用いられている。

$$\mathcal{T}_j(y_{ij}) = \mathcal{E}(\mathcal{T}_j(y_{ij})|\boldsymbol{x_i}) + e_{ij} \tag{2}$$

$$\begin{aligned} \mathcal{E}(\mathcal{T}_j(y_{ij})|\boldsymbol{x_i}) &= \mathcal{S}(\boldsymbol{x_i})'\boldsymbol{\beta_j} & (3)\\ &= \sum_{\ell=1}^{q} \mathcal{S}_\ell(x_{i\ell})\beta_{j\ell} & (4) \end{aligned}$$

ただし、$\mathcal{T}_j()$ は第 j 番目の予測変数を間隔尺度へ変換するための関数、また、$\mathcal{S}()$ は各説明変数を間隔尺度に変換するための関数 $\mathcal{S}_\ell()$, ℓ=1,2, ... , q, を元のデータの形に並べたもの

$$\mathcal{S}(\boldsymbol{x_i}) = [\mathcal{S}_1(x_{i1}), \mathcal{S}_2(x_{i2}), \ \ldots, \mathcal{S}_q(x_{iq})]' \tag{5}$$

である。Eq.(4) は、変換後の予測変量の $\boldsymbol{x_i}$ を与えたときの条件付き期待値が、変換後の $\boldsymbol{X}$ の線形関数で書き表されることを示している。たとえば、被験者ｋ以外の評定値は間隔尺度水準の測定値とみなせるが、被験者ｋの評定値はその対数をとることによって間隔尺度へ変換されること、また、商品の特性のうち第m番目以外のものは間隔尺度とみなせるが、第m番目の特性はそれを３乗することにより間隔尺度へ変換可能である、ということが前もって解っていれば、上記のモデルのパラメタは、$\mathcal{T}_j(y_{ij}) = y_{ij}, j = 1, 2, \ \ldots, k-1, k+1, \ \ldots, p$, $\mathcal{T}_k(y_{ik}) = \log(y_{ik})$、$\mathcal{S}_\ell(x_{i\ell}) = x_{i\ell}, \ell = 1, 2, \ \ldots, m-1, m+1, \ \ldots,$ q $\mathcal{S}_m(x_{im}) = {x_{im}}^3$、という変換を施したデータに通常の回帰分析を行えば得られる。ただ、この例のように、変換の形が前もって知られている場合は希であり、実際は、その変換の形も推定することが普通である。（$\mathcal{T}()$ や$\mathcal{S}()$ をの形を規定する変換のパラメタをも推定する。）

上記の回帰分析モデルの場合、最小２乗法を用いて推定を行うとすれば、最小２乗基準が

$$\begin{aligned}
\mathrm{RSS} &= \mathrm{tr}\left((\mathcal{T}(\boldsymbol{Y}) - \mathcal{S}(\boldsymbol{X})\boldsymbol{B}')'(\mathcal{T}(\boldsymbol{Y}) - \mathcal{S}(\boldsymbol{X})\boldsymbol{B}')\right) &(6)\\
&= \sum_{j=1}^{p}\sum_{i=1}^{n}(\mathcal{T}_j(y_{ij}) - \mathcal{S}(\boldsymbol{x_i})'\boldsymbol{\beta_j})^2 &(7)\\
&= \sum_{j=1}^{p}\mathrm{RSS}_j &(8)
\end{aligned}$$

と書けるため、説明変数の変換、$\mathcal{S}_\ell$, ℓ=1,2, ... ,q, が既知の場合、RSS を最小にするような変換 $\mathcal{T}_j()$ および回帰係数 $\boldsymbol{\beta_j}$ は、各予測変数ごとに、 RSS_j を最小とするように求めれば良いことが解る。また、同じ最小２乗基準が、第ℓ 番目の説明変数ベクトル $\boldsymbol{x}_{(\ell)}$ (n $\times$ 1 のベクトル) とそれを除く部分の変換 $\mathcal{S}(\boldsymbol{X}_{(\ell)})$ (n $\times$ (q-1) の行列) を用いて

$$\begin{aligned}
\mathrm{RSS} &= \mathrm{tr}\left(\mathcal{T}(\boldsymbol{Y}) - \mathcal{S}(\boldsymbol{X}_{(\ell)})\boldsymbol{B}_{(\ell)}{}' - \mathcal{S}_\ell(\boldsymbol{x}_{(\ell)})\boldsymbol{\beta}_{(\ell)}{}''[\mathcal{T}(\boldsymbol{Y}) - \mathcal{S}(\boldsymbol{X}_{(\ell)})\boldsymbol{B}_{(\ell)}{}' - \mathcal{S}_\ell(\boldsymbol{x}_{(\ell)})\boldsymbol{\beta}_{(\ell)}{}']\right) &(9)\\
&= [\mathcal{S}_\ell(\boldsymbol{x}_{(\ell)}) - (\mathcal{T}(\boldsymbol{Y}) - \mathcal{S}(\boldsymbol{X}_{(\ell)})\boldsymbol{B}_{(\ell)}{}')\boldsymbol{\beta}_{(\ell)}/(\boldsymbol{\beta}_{(\ell)}{}'\boldsymbol{\beta}_{(\ell)})]'\\
&\quad \times [\mathcal{S}_\ell(\boldsymbol{x}_{(\ell)}) - (\mathcal{T}(\boldsymbol{Y}) - \mathcal{S}(\boldsymbol{X}_{(\ell)})\boldsymbol{B}_{(\ell)}{}')\boldsymbol{\beta}_{(\ell)}/(\boldsymbol{\beta}_{(\ell)}{}'\boldsymbol{\beta}_{(\ell)})]\\
&\quad \times (\boldsymbol{\beta}_{(\ell)}{}'\boldsymbol{\beta}_{(\ell)}) \ + \ (\mathcal{S}_\ell(\boldsymbol{x}_{(\ell)}) \text{ を含まない部分}) &(10)
\end{aligned}$$

と書き表せるため、基準変数の変換 $\mathcal{T}(\boldsymbol{Y})$、回帰係数行列 $\boldsymbol{B}$ 、および $\mathcal{S}(\boldsymbol{X}_{(\ell)})$ が既知の場合には、第ℓ番目の説明変数の変換 $\mathcal{S}_\ell(\boldsymbol{x}_{(\ell)})$ は上記の２次形式を最小にするように求めればよいことが解る。

このように、モデルに含まれるパラメタをいくつかのサブセットに分轄し、パラメタの各サブセットごとに同じ最小２乗基準を最小化する様な推定を行うことを繰り返す方法は、交互最小２乗法（ALS）と呼ばれ、複雑なモデルのパラメタを推定する際の常套手段となっている。回帰分析モデルの場合は、データの変換が与えられた場合の回帰係数の推定は、

$$\widehat{\boldsymbol{B}}' = (\mathcal{S}(\boldsymbol{X})'\mathcal{S}(\boldsymbol{X}))^{-1}\mathcal{S}(\boldsymbol{X})'\mathcal{T}(\boldsymbol{Y}) \tag{11}$$

となることは明かである。また、Eq.(7) や Eq.(10) は、(u_i, v_i) を既知のデータ、$\mathcal{F}$ を推定すべき変換とすれば、

$$\mathrm{RSS} = \sum_{i=1}^{n}(\mathcal{F}(v_i) - u_i)^2 \ + \ (\mathcal{F} \text{ のパラメタを含まない部分}) \tag{12}$$

という形をしているため、$\mathcal{F}()$ に含まれるパラメタの推定は、$\boldsymbol{o}$u の $\boldsymbol{v}$ へ回帰する１変量の回帰分析の問題となる。特に、 PROC TRANSREG においては、データの変換として、スプライン関数が用いられる

ことが多いが、これは変換のパラメタに関して線形に書き表すことが出来るため、その推定は簡単である。また、$\mathcal{F}()$ として任意の単調変換を仮定すれば、Eq.(12) を最小にするようなものは、Kruskal の最小２乗単調変換として計算することが出来るし、名義尺度のデータは Fisher の方法を用いてカテゴリの平均値を計算すればよい。このように、変数の尺度水準を考慮しながら、その変数の持つ情報を損なわない形で最小２乗基準を最小にするようなデータの変換を求めることを、データの最適変換と呼ぶ。なお、スプライン関数を用いてデータを変換する場合は、単にその変換が連続で滑らかなものであることのみを仮定していることになる。

PROC TRANSREG は、きわめて一般的なプロシジャであるが、主に、説明変数として名義尺度の変数を用いたコンジョイント分析としてマーケッティングの分野で使用されることが多い。また、順序データの PREFMAP モデルによる分析にも利用することが出来る。

2.2 因子分析

上記の方法は、そのまま因子分析モデルに応用することが出来る。すなわち、データを変換したものが、潜在変数である因子スコアと線形に関係していることを仮定するモデル、

$$\mathcal{E}(\mathcal{T}_j(y_{ij})|\boldsymbol{x_i}) = \boldsymbol{x_i}'\boldsymbol{\beta_j} \tag{13}$$

である。これは、まとめて書き表せば

$$\mathcal{E}(\mathcal{T}(\boldsymbol{Y})|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{B}' \tag{14}$$

となるが、$\boldsymbol{X}$ は潜在変数であるため、その変換 $\mathcal{S}()$ をモデルにいれる必要はない。回帰分析と同様に ALS による最小２乗推定を行うことにすれば、最小２乗基準は以下のように書き表される。

$$\begin{aligned} \mathrm{RSS} &= \mathrm{tr}\left((\mathcal{T}(\boldsymbol{Y}) - \boldsymbol{X}\boldsymbol{B}')'(\mathcal{T}(\boldsymbol{Y}) - \boldsymbol{X}\boldsymbol{B}')\right) && (15)\\ &= \sum_{j=1}^{p}\sum_{i=1}^{n}(\mathcal{T}_j(y_{ij}) - \boldsymbol{x_i}'\boldsymbol{\beta_j})^2 && (16)\\ &= \sum_{j=1}^{p}\mathrm{RSS}_j && (17) \end{aligned}$$

したがって、変換を既知とした場合の因子スコア $\boldsymbol{X}$ および因子パタン $\boldsymbol{B}$ の推定値は、$\mathcal{T}(\boldsymbol{Y})$ をあたかもデータ行列のように取り扱った因子分析の解（いわゆる主成分分解および主成分スコア）

$$\widehat{\boldsymbol{B}} = \boldsymbol{P}\boldsymbol{\Lambda}^{\frac{1}{2}} \tag{18}$$

ただし、$\boldsymbol{\Lambda}$ と $\boldsymbol{P}$は　$\frac{1}{n}\mathcal{T}(\boldsymbol{Y})'\mathcal{T}(\boldsymbol{Y})$　の最大ｑ個の固有値と固有ベクトル行列、

$$\boldsymbol{X} = \mathcal{T}(\boldsymbol{Y})\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1} = \mathcal{T}(\boldsymbol{Y})\boldsymbol{P}\boldsymbol{\Lambda}^{-\frac{1}{2}} \tag{19}$$

として与えられる。また、データの最適変換は、各変数毎に、Eq.(16) を最小とするように回帰分析の場合と全く同様にして求めればよい。この、データの最適変換を含む因子分析モデルによる分析は、SAS の PROC PRINQUAL を用いて行うことが出来る。データの最適変換という過程が含まれるため、通常の因子分析を行う場合よりも少ない因子の数でデータを説明することが可能となる。

3 類似度データのモデル

他方、多変量解析には、上記のようなプロフィールデータの分析ではなく、個体相互間の関係を取り扱う方法も存在する。たとえば、個体 i と個体 j の類似度の測定値を、o_{ij}, i,j=1, . . . ,n　とすれば、その期

待値の構造として個体のｑ次元の尺度値（潜在変数による個体の測定値） $\boldsymbol{x_i}$,i=1,2, . . . ,n、から計算された距離を当てはめるモデル

$$\mathcal{E}(\mathcal{T}(o_{ij})|\boldsymbol{x_i},\boldsymbol{x_{(j)}}) = \sqrt{(\boldsymbol{x_i}-\boldsymbol{x_{(j)}})'(\boldsymbol{x_i}-\boldsymbol{x_{(j)}})} \tag{20}$$

を考えることができる。この場合、もはやモデルはパラメタに関して線形とはならないが、 ALS を用いれば、

$$\mathrm{RSS} = \sum_{i>j}\left(\mathcal{T}(o_{ij}) - \sqrt{(\boldsymbol{x_i}-\boldsymbol{x_{(j)}})'(\boldsymbol{x_i}-\boldsymbol{x_{(j)}})}\right)^2 \tag{21}$$

を最小とするデータの変換は前節と同様にして求められることが解る。また、変換が既知の場合の尺度値の推定は、例えば、 Gauss-Newton 法を用いて数値的に行うことができる。 SAS の PROC MDS はこのモデルの下での解析を行うためのもので、類似度が多数の個人から得られている場合には、個人差を各尺度値にかかる重みの大小という形でモデルに組み込んだ、いわゆる重み付きユークリッドモデル（個人差 MDS モデル）

$$\mathcal{E}(\mathcal{T}_k(o_{ijk})|\boldsymbol{x_i},\boldsymbol{x_{(j)}},\boldsymbol{o}W_k) = \sqrt{(\boldsymbol{x_i}-\boldsymbol{x_{(j)}})'\boldsymbol{o}W_k(\boldsymbol{x_i}-\boldsymbol{x_{(j)}})} \tag{22}$$

による分析も行うことができる。

4　最小２乗法によるパラメタの推定

データとモデルの値との残差の２乗和、

$$\mathrm{RSS} \quad = \quad \sum_{j=1}^{p}\sum_{i=1}^{N} {e_{ij}}^2 = \sum_{j=1}^{p}\sum_{i=1}^{N}(y_{ij}-\boldsymbol{x_i}\boldsymbol{\beta_j})^2 \tag{23}$$

で定義される最小２乗基準 RSS を最小とするようにモデルのパラメタ、$\boldsymbol{\beta}$ を推定する方法を最小２乗法と呼び、推定されたパラメタの値を $\widehat{\boldsymbol{\beta}}$ と記す。

残差の２乗和 RSS をモデルのパラメタの関数と見た場合、その下限はゼロであるから必ず最小値が存在し、その点では RSS のすべてのパラメタに関する偏微分がゼロになっているはずである。従って、最小２乗推定値を求めるためには、正規方程式

$$\frac{\partial\,\mathrm{RSS}}{\partial\,\beta_{j\ell}} \quad = \quad 0, \quad j=1,2,\ldots,p, \ell=1,2,\ldots,q \tag{24}$$

を解けばよい。また、モデルが潜在変数を含む場合には $\boldsymbol{X}$ をもモデルのパラメタとして取り扱い

$$\frac{\partial\,\mathrm{RSS}}{\partial\,x_{i\ell}} \quad = \quad 0, \quad i=1,2,\ldots,N, \ell=1,2,\ldots,q \tag{25}$$

を満たす $\boldsymbol{X}$ の推定値 $\widehat{\boldsymbol{X}}$ をも同時に求めることになる。ただ、因子分析モデルの場合、$\boldsymbol{X}$ と $\boldsymbol{\beta}$ の間に乗算的不定性が存在するため、一意な解を求めるためには何らかの制約を課す必要が生じる。また、クラスター分析モデルの場合には、$\boldsymbol{X}$ の各行は一つの１と $q-1$ 個のゼロからなるという制約を付ける。

最後に、データの最適変換を含むモデルの場合には、最適変換のパラメタを $\boldsymbol{\Gamma}$,$\boldsymbol{\Lambda}$ とすれば、RSS は

$$\mathrm{RSS} = \sum_{j=1}^{p}\sum_{i=1}^{N}(\mathcal{T}_j(y_{ij}|\boldsymbol{\gamma}_j) \;-\; \mathcal{S}(\boldsymbol{x_i}|\boldsymbol{\Lambda})\boldsymbol{\beta_j})^2 \tag{26}$$

と書き表されるが、この RSS を最小とするように最適変換のパラメタとモデルのパラメタを同時に推定することになる。ただ、一般に、基準変数の最適変換を含む場合には、$\left\|\mathcal{T}_j(\boldsymbol{Y}|\boldsymbol{\gamma}_j)\right\|^2$ を小さくすることにより RSS は限りなくゼロに近づくため、RSS を $\left\|\mathcal{T}(\boldsymbol{Y}|\boldsymbol{\Gamma})\right\|^2$ で基準化した、規準化最小２乗基準

$$\mathrm{RSS}_1 = \sum_{j=1}^{p}\frac{\sum_{i=1}^{N}(\mathcal{T}_j(y_{ij}|\boldsymbol{\gamma}_j) \;-\; \mathcal{S}(\boldsymbol{x_i}|\boldsymbol{\Lambda})\boldsymbol{\beta_j})^2}{var\left(\mathcal{T}_j(y_{ij}|\boldsymbol{\gamma}_j)\right)} \tag{27}$$

を最小化することが多い。すなわち、正規方程式

$$\frac{\partial\,\mathrm{RSS}_1}{\partial\,\beta_{j\ell}} = 0, \quad j=1,2,\ldots,p, \ell=1,2,\ldots,q \tag{28}$$

$$(\frac{\partial\,\mathrm{RSS}_1}{\partial\,x_{i\ell}} = 0, \quad i=1,2,\ldots,N, \ell=1,2,\ldots,q) \tag{29}$$

$$\frac{\partial\,\mathrm{RSS}_1}{\partial\,\boldsymbol{\gamma}_j} = 0, \quad j=1,2,\ldots,p \quad \text{and} \quad \frac{\partial\,\mathrm{RSS}_1}{\partial\,\boldsymbol{\lambda}_\ell} = 0, \quad \ell=1,2,\ldots,q \tag{30}$$

$$\tag{31}$$

が解が求める最小２乗推定値となる。この連立方程式が解析的に解けない場合には交互最小２乗法（ ALS ）を用いて数値的に解を求めることになる。

5　モデルの適合度

モデルの最小２乗推定値を最小２乗法でパラメタを推定する場合、データへのモデルの適合度を表す指標として、分散説明率、 R^2 を考えると都合がよい。いま、モデルの最小２乗推定値を

$$\hat{\boldsymbol{y}}_{(j)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j \quad \text{または} \quad \widehat{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}_j \tag{32}$$

で表す。ただし、$\hat{\boldsymbol{y}}_{(j)}$ は第 j 番目の変数のモデルの最小２乗推定値を要素として持つ $\mathrm{N} \times 1$ のベクトルである。この記法および列中心化行列 $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{N}\mathbf{1}\mathbf{1}'$ を用いれば、第 j 番目の変数の分散説明率 R_j^2 は

$$\mathrm{R}_j^2 = \frac{var\left(\hat{\boldsymbol{y}}_{(j)}\right)}{var\left(\boldsymbol{y_{(j)}}\right)} = \frac{\hat{\boldsymbol{y}}_{(j)}\boldsymbol{J}\hat{\boldsymbol{y}}_{(j)}}{\boldsymbol{y_{(j)}}\boldsymbol{J}\boldsymbol{y_{(j)}}} \tag{33}$$

と定義されるが、これは、モデルの推定値の分散のデータとして得られる分散に対する比に他ならない。すなわち、R_j^2 が 1 の場合には変数 j の分散の全てがモデルによって説明されたことになる。また、この値は、$\boldsymbol{y_{(j)}}$ と $\hat{\boldsymbol{y}}_{(j)}$ の間の相関係数の２乗

$$\mathrm{R}_j^2 = \left(corr\left(\boldsymbol{y_{(j)}}, \hat{\boldsymbol{y}}_{(j)}\right)\right)^2 \tag{34}$$

となっている。

基準変数が p 個ある場合は全体の分散説明率をふた通りに定義することが出来る。一つは

$$\mathrm{R}_T^2 = \frac{\sum_{j=1}^{p} var\left(\hat{\boldsymbol{y}}_{(j)}\right)}{\sum_{j=1}^{p} var\left(\boldsymbol{y_{(j)}}\right)} = \frac{\sum_{j=1}^{p} var\left(\boldsymbol{y_{(j)}}\right)\mathrm{R}_j^2}{\sum_{j=1}^{p} var\left(\boldsymbol{y_{(j)}}\right)} = \sum_{j=1}^{p} \frac{var\left(\boldsymbol{y_{(j)}}\right)}{\sum_{k=1}^{p} var\left(\boldsymbol{y}_{(k)}\right)}\mathrm{R}_j^2 \tag{35}$$

で定義される、全分散説名率 R_T^2 であり、もう一つは

$$\mathrm{R}_A^2 = \sum_{j=1}^{p} \mathrm{R}_j^2/p = \sum_{j=1}^{p} \frac{var\left(\hat{\boldsymbol{y}}_{(j)}\right)}{var\left(\boldsymbol{y_{(j)}}\right)}/p \tag{36}$$

で定義される平均分散説明率 (Average r-squared R_A^2 である。

一般に、パラメタを最小２乗法を用いて推定した場合、その値は分散説明率を最大とするようなものになっていることがいえるが、データの最適変換がなくモデルが潜在変数を含まない場合には、RSS を最小とするパラメタを求めることにより R_T^2 と R_A^2 の両方が最大化され、どちらかを含む場合には、RSS_1 の最小化により R_A^2 が最大化されることになる。

6　説明変数の寄与率

ひとつの基準変数を q 個の説明変数で説明する際に、各説明変数がどの程度貢献しているかの測度を説明変数の寄与率と呼ぶ。寄与率の定義には様々なものが考えられるが、モデルから $\boldsymbol{x}_{(k)}$ を除いたときの R^2 の減少量を持って寄与率の測度とすることが考えられる。この場合、第 k 番目の説明変数 $(\boldsymbol{x}_{(k)})$ の寄与率 C_k を求めるためには、説明変数及びパラメタをを

$$\boldsymbol{X} = [\boldsymbol{X}_{(K)} \quad \boldsymbol{x}_{(k)}] \qquad \text{and} \qquad \boldsymbol{\beta} = [\boldsymbol{\beta}_K, \beta_k] \tag{37}$$

ただし、$\boldsymbol{X}_{(K)}$ は全説明変数から $\boldsymbol{x}_{(k)}$ を除いた $N \times \ q-1$ の行列、と分割し、モデル

$$\boldsymbol{y} = \boldsymbol{X}_{(K)}\boldsymbol{\beta}'_K \tag{38}$$

で計算される R^2 を R_1^2 、モデル

$$\boldsymbol{y} = \boldsymbol{X}_{(K)}\boldsymbol{\beta}_K + \boldsymbol{x}_{(k)}\beta_k \tag{39}$$

で計算される R^2 を R_2^2とすれば、

$$\begin{aligned} C_k &= \mathrm{R}_2^2 - \mathrm{R}_1^2 && (40)\\ &= \frac{\boldsymbol{y}'(\boldsymbol{P}_X\boldsymbol{J}\boldsymbol{P}_X - \boldsymbol{P}_K\boldsymbol{J}\boldsymbol{P}_K)\boldsymbol{y}}{N\ var\,(\boldsymbol{y})} && (41)\\ &= \left(corr\left(\boldsymbol{y}, \boldsymbol{Q}_K\boldsymbol{x}_{(k)}y\right)\right)^2 && (42) \end{aligned}$$

を計算すればよい。ただし、$\boldsymbol{P}_X$ および $\boldsymbol{P}_K$ は $\boldsymbol{X}$ と $\boldsymbol{X}_{(K)}$ の列空間への射影行列また、 $\boldsymbol{Q}_K = \boldsymbol{I} - \boldsymbol{P}_K$ である。なお、この $\boldsymbol{Q}_K\boldsymbol{x}_{(k)}$ は、$\boldsymbol{x}_{(k)}$ を $\boldsymbol{X}_{(K)}$ の線形モデルで説明した場合の最小２乗残差となっているが、上記の寄与率の定義は、$\boldsymbol{x}_{(k)}$ から他の説明変数で説明でる部分を除いたもの（残差）と基準変数との相関係数の２乗となっている。その意味で、寄与率は片側偏相関係数の２乗 (squared semipartial correlation coefficient) もしくは、部分相関係数の２乗 (squared part correlation coefficient) と呼ばれることがある。なお、Eq.(42) の寄与率の分子は、 いわゆる調整済み平方和、ＳＡＳ の type II SS（または type III SS ）であるが、このように定義された寄与率は加算性を持たない。すなわち、全ての寄与率の和に $var\,(\widehat{\boldsymbol{e}})/var\,(\boldsymbol{v})$ を加えたものが 1 とはならない。加算性を持つ寄与率は ＳＡＳ の type I SS をもとにして定義する事が出来るが、この場合には順序依存性という欠点が生じる。すなわち、同じ変数が、説明変数行列の何番目に位置するかによって、その寄与率が異なってくる。

また、$\boldsymbol{x}_{(k)}$ が $\boldsymbol{X}_{(K)}$ の全ての列と直交する場合は $\boldsymbol{Q}_K\boldsymbol{x}_{(k)} = \boldsymbol{x}_{(k)}$ となるため、

$$C_k = \left(corr\left(\boldsymbol{y}, \boldsymbol{x}_{(k)}\right)\right)^2 \tag{43}$$

となり、単純な寄与率と一致する。

基準変数が複数個ある場合の説明変数の寄与率は、それを説明変数を除いたときの、R_T^2 の減少量と定義するか R_A^2 の減少量と定義するかで異なってくるが、 R_T^2 を用いれば、

$$C_{Tk} = \frac{\sum_{j=1}^{p}\left(var\left(\boldsymbol{y_{(j)}}\right) \times \ \left(corr\left(\boldsymbol{y_{(j)}}, \boldsymbol{Q}_K\boldsymbol{x}_{(k)}\right)\right)^2\right)}{\sum_{j=1}^{p} var\left(\boldsymbol{y_{(j)}}\right)} \tag{44}$$

となり、 R_A^2 を用いれば、

$$C_{Ak} = \frac{1}{p}\sum_{j=1}^{p}\left(corr\left(\boldsymbol{y_{(j)}}, \boldsymbol{Q}_K\boldsymbol{x}_{(k)}\right)\right)^2 \tag{45}$$

となる。従って、基準変数の分散が同一の場合には $C_{Tk} = C_{Ak}$ であり、説明変数の各列が直交している場合には、

$$C_{Tk} = \frac{\sum_{j=1}^{p} \left(var\left(\boldsymbol{y_{(j)}}\right) \times \left(corr\left(\boldsymbol{y_{(j)}}, \boldsymbol{x}_{(k)}\right)\right)^2 \right)}{\sum_{j=1}^{p} var\left(\boldsymbol{y_{(j)}}\right)} \text{ and } C_{Ak} = \frac{1}{p} \sum_{j=1}^{p} \left(corr\left(\boldsymbol{y_{(j)}}, \boldsymbol{x}_{(k)}\right)\right)^2 \quad (46)$$

となる。

7 統計的推測について

デザイン行列 $\boldsymbol{X}$ をパラメタとして取り扱うモデル（因子分析やクラスター分析）には，観測値数を増やすとそれだけパラメタ数も増えるという欠点がある．この問題を解決するためには，デザイン行列の各行が同一分布からのランダムサンプルであることを仮定した後にそれらを周辺化した $\boldsymbol{\beta}$ のみの周辺尤度を最大とするような解を求めればよい．因子分析の場合は通常用いられている最尤解がこれにあたり，クラスター分析の場合には，混合正規分布のパラメタの推定問題となる．

8 おわりに

本稿で紹介したデータ解析の方法は、特に、因子分析モデルや MDS モデルは、分析の結果として得られるパラメタの値の大小を議論することよりも、それを基にして個体と変数の関係に視覚的に表現を与えることを主眼とした方法である。その意味でこれらの方法は探索的な方法であると言うことができる。また、推定されたデータ変換の意味を考える為にはそのプロットは必須である。

訂正表

前川眞一　　970819,1016,010329,031114

第２刷の修正は頁番号に * を付けて示した。

第 2 章

1. p22, 表 2.6: 変数 X の相関行列も掲示されているがそれは無視。
2. p23, 表 2.7: 変数 X を含む相関係数も掲示されているがそれは無視。
3. p25, l6: データセットの転置の説明は p29 を参照のこと。
4. p32, lb8: var x1–y8 とすると全ての相関係数を計算する。

第 3 章

1. p63, l11: $\boldsymbol{x_i} \leftarrow \boldsymbol{b}$
2. p63, lb4: $\boldsymbol{x_i} \leftarrow \boldsymbol{b}$
3. p70, l2: M 次Ｐスプラインで近似する
4. p76, l9: $\boldsymbol{H}^{(p)}(\boldsymbol{X}) = [\boldsymbol{g}_2^+(\boldsymbol{X}) \quad -2\boldsymbol{X}]$
5. p110, l9: &NFACT が未定義となっている。例えば、%let nfact=2 とする。
6. p110, lb9: &NFACT が未定義となっている。例えば、%let nfact=2 とする。
7. p119, 図 3.13(a): 非階層的方法の個体 4 と 5 を横線でつなぐ。
8. p125, l2: その解の R^2 の値は、その直線より下方に位置するクラスターを定義する横線の中で最も高いものの縦軸の座標として示される。

第 4 章

1. p150, l6:

$$\mathrm{RSS} = \mathrm{tr}\left[(\boldsymbol{V} - \boldsymbol{U}\hat{\boldsymbol{A}}\hat{\boldsymbol{B}}')'(\boldsymbol{V} - \boldsymbol{U}\hat{\boldsymbol{A}}\hat{\boldsymbol{B}}')\right]$$

2. p150, l13:

$$\mathrm{Rss}_w = \left\|(\boldsymbol{V} - \boldsymbol{U}\boldsymbol{A}\boldsymbol{B}')\boldsymbol{W}^*\right\|^2$$

3. p156, 図 4.1: 縦軸は $\boldsymbol{v}$ と $\hat{\boldsymbol{v}}$ を、横軸は $\boldsymbol{u}$ を表す。
4. p156, 図 4.2: 縦軸は $\boldsymbol{v}$ と $\hat{\boldsymbol{v}}$ を、横軸は $\boldsymbol{u}$ を表す。
5. p175, lb9

$$= -2\boldsymbol{U}'\boldsymbol{v} + 2\boldsymbol{U}'\boldsymbol{U}\boldsymbol{x_i} - 2\lambda\boldsymbol{U}'\boldsymbol{U}\boldsymbol{x_i}$$

第 5 章

1. p184* 式 (5.2) 三角不等式の訂正

$$d_{ij} + d_{jk} \geq d_{ik} \tag{47}$$

2. p186, l10: INDSCAL Model)
3. p204, 表 5.2（中）: 非類似度データ行列

第 6 章

1. p205, lb10: 回帰係数と定数項を $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}$ として、

付録 A

1. p213, lb11-10: 「この場合、 $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$ となる。」を削除。

付録 C

1. p231, l21: print ステートメントに obs= データセットオプションを付けると ….（例は p33 に示した。）
2. p245, lb2: abc=”MACV”; bcd=’hello’; xmean=12.6;
3. p246, lb12: out* で始まる行は削除。
4. p248, print1 を用いて n 個のオブザベーションのみをプリントしたい場合には
   print1 datasetname(OBS=n)
   とする。
5. p254, l1: ぜならば、マクロ変数を利用した場合は、
6. p254, lb2: ル積は次節の例で示すように
7. p256, l3-6: プログラムの最後に ; を付ける。
8. p262, lb16: 列 T, および RSS を変数の数と因子の数の積で除したものの平方根 RMSE が返される。

参考文献

1. p273: Crawford, C. B. (1975)

索引

1. p286: proc iml 7, 32, 92, 160, 162, 167, 180, 190, 196, 250