

回帰分析モデル（暫定版）

前川眞一

20200423,30

目次

| | | |
|-------|------------|----|
| 1 | はじめに | 1 |
| 2 | モデル | 1 |
| 2.1 | 線型重回帰モデル | 2 |
| 2.1.1 | 重回帰分析の行列表記 | 3 |
| 2.1.2 | 最小2乗解 | 5 |
| 2.2 | バリエーション | 5 |
| 2.3 | デザイン行列 | 7 |
| 2.4 | セル平均による表現 | 10 |
| 2.5 | 統計的検定 | 11 |

1 はじめに

回帰分析とは、ある変数（群）の変動を別の変数（群）で説明したい場合に用いられる統計的方法である。説明される側の変数は、基準変数 (criterion variable)、従属変数 (dependent variable)、被説明変数、等の名称で呼ばれる。説明する側の変数には、説明変数 (regressor variable)、独立変数 (independent variable) 等の名称が用いられる。通常、基準変数にはローマ字の y, Y 等が用いられ、説明変数には x, X 等が用いられるが、それ以外の文字を使っても構わない。この表記を用いると、回帰分析は、「基準変数 (y) を説明変数 (x) で説明する」ための方法であり、これは「基準変数 (y) を説明変数 (x) に回帰する (regress y on x)」と表現されることもある。また、説明するという代わりに予測するという言葉を用い「基準変数 (y) を説明変数 (x) で予測する」という表現を用いる場合もある。

2 モデル

基準変数 (y) を説明変数 (x) で説明するということは、一般的には以下のように考える。なお、回帰分析が用いられる場合では、基準変数は一つで説明変数が多数（多変量）の場合が多いので、以下では、 y はスカラーであり x はベクトルであるとする。いま、 $(y_i, \mathbf{x}_i), i = 1, 2, \dots, N$, を個体 i が持つ変数 y と x の値とする。 y のうち x で説明される部分を $f(\mathbf{x}|\beta)$ として

$$y_i = f(\mathbf{x}_i|\beta) + e_i, \quad i = 1, 2, \dots, N \quad (1)$$

とする。ただし、 $f(\mathbf{x})$ は \mathbf{x} を与えたときにスカラー値を出力する関数で β はそのパラメタ（母数とも呼ばれる未知数）、また、 e_i はモデルとデータのズレである。 $f(\mathbf{x}_i)$ はモデルの予測値と

も呼ばれ \hat{y}_i と書かれることもある。この表現を用いると

$$y_i = \hat{y}_i + e_i \quad (2)$$

$$\text{where } \hat{y}_i = f(\mathbf{x}_i), \quad i = 1, 2, \dots, N, \quad (3)$$

である。

回帰分析モデルにおいては、 \mathbf{x} を与えた場合の y の予測値を与える関数の形により様々なモデルが考えられるが、最も簡単な場合は、以下に示す一つの説明変数 x を用いて y を説明する単回帰モデルと呼ばれるもの

$$\hat{y}_i = f(\mathbf{x}_i) = \beta_0 + \beta_1 x_i, \quad i = 1, 2, \dots, N, \quad (4)$$

である。これを、説明変数が q 個ある場合に拡張したモデル

$$\hat{y}_i = f(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^q x_{ij} \beta_j, \quad i = 1, 2, \dots, N, \quad (5)$$

を重回帰モデルと呼ぶ。切片のパラメタは、通常は説明変数の数に加えないため 0 という添字を用いるが、行列表記をする場合には、この表現は微妙である。当然であるが、単回帰モデルは重回帰モデルの特殊例 ($q = 1$) である。また、上記のモデルのように、パラメタに関して線形なモデルを線型モデルと呼び、そうでないもの、例えば

$$\hat{y}_i = \beta_0 + \log(x_i - \beta_1), \quad \text{or} \quad \hat{y}_i = \beta_0 + \frac{1}{1 + \exp(\beta_1 x_i - \beta_2)} \quad (6)$$

の様なモデルを非線型モデルと呼ぶ。

回帰分析モデルを用いてデータを分析するということは、未知のモデルのパラメタ β を何らかの統計的方法を用いて推定することに他ならないが、一般に、線型モデルには解析的な解があるが、非線形モデルのパラメタは何らかの数値解析法を用いて求めなければならない。

2.1 線型重回帰モデル

q 個の説明変数を用いた重回帰モデルを

$$y_i = \hat{y}_i + e_i, \quad i = 1, 2, \dots, N \quad (7)$$

$$\text{where } \hat{y}_i = \beta_0 + \sum_{j=1}^q x_{ij} \beta_j \quad (8)$$

とする。誤差項 e_i に関して分布を仮定する場合は、通常、正規分布

$$e_i \sim N(0, \sigma^2), \quad i = 1, 2, \dots, N, \quad iid \quad (9)$$

を仮定する場合が多い。

このモデルのパラメタは $q + 1 \times 1$ のベクトル β であるが、誤差項に正規分布を仮定した場合には誤差分散 σ^2 もパラメタとして取り扱う。

特に誤差項に分布を仮定しない場合でも、誤差項の期待値は 0

$$\mathcal{E}(e_i) = 0 \quad (10)$$

という仮定が暗黙におかれていると考えられる。したがって、

$$\mathcal{E}(y_i) = \mathcal{E}(\hat{y}_i + e_i) = \mathcal{E}(\hat{y}_i) + \mathcal{E}(e_i) = \hat{y}_i = f(\mathbf{x}_i) \quad (11)$$

である。なお、通常の回帰分析モデルでは、説明変数は確率変数としては取り扱わない。特に誤差項に分布を仮定しない場合、モデルの母数 β は最小2乗法を用いて推定される。最小2乗法は以下に示す残差の2乗和 RSS

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (12)$$

を最小とするような β を求める方法であるが、重回帰分析の場合、その一般解は、以下に示す行列を用いた表記を用いて導くのが便利である。

2.1.1 重回帰分析の行列表記

いま、 β_0 に対応する説明変数として、全ての個体に関して 1 という値を持つ変数 x_0 を考える。すなわち

$$x_{i0} = 1 \quad (13)$$

である。これを用いると、重回帰分析のモデルは

$$y_i = \sum_{j=0}^q x_{ij}\beta_j + \varepsilon_{ij} \quad (14)$$

と書ける。

したがって、回帰分析の行列表記は以下のようになる。

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (15)$$

$$= \hat{\mathbf{y}} + \boldsymbol{\varepsilon} \quad (16)$$

ただし、

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{y} \quad n \times 1$$

基準変数

$$\mathbf{X} \quad n \times (q+1)$$

説明変数 + $\mathbf{1}$

$$\boldsymbol{\beta} \quad (q+1) \times 1$$

回帰係数 (切片を含む)

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} \quad n \times 1$$

モデルの予測値

である。

これを図示すると、

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{bmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{pmatrix} \quad (17)$$

となる。

本稿では、特別に断らない限り、ベクトルは列ベクトルであるとする。したがって、 \mathbf{X} 行列の第 j 列を $n \times 1$ のベクトル $\mathbf{x}_{(j)}$ を用いて

$$\mathbf{x}'_{(j)} = (x_{1j}, x_{2j}, x_{3j}, \dots, x_{nj}) \quad (18)$$

もしくは

$$\mathbf{x}_{(j)} = (x_{1j}, x_{2j}, x_{3j}, \dots, x_{nj})' \quad (19)$$

と書けば \mathbf{X} は以下のようにベクトルの集合として表現することも出来る。

$$\mathbf{X} = (\mathbf{x}_{(0)}, \mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \dots, \mathbf{x}_{(q)}) = \begin{pmatrix} \mathbf{x}'_{(0)} \\ \mathbf{x}'_{(1)} \\ \mathbf{x}'_{(2)} \\ \vdots \\ \mathbf{x}'_{(q)} \end{pmatrix}' \quad (20)$$

また、 \mathbf{X} 行列の第 i 行を $(q+1) \times 1$ のベクトル \mathbf{x}_i の転置を用いて

$$\mathbf{x}'_i = (x_{i0}, x_{i1}, x_{i2}, \dots, x_{iq}) \quad (21)$$

もしくは

$$\mathbf{x}_i = (x_{i0}, x_{i1}, x_{i2}, \dots, x_{iq})' \quad (22)$$

と書けば \mathbf{X} は以下のようにベクトルの集合として表現することも出来る。

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)' \quad (23)$$

したがって、

$$\hat{y}_i = \sum_{j=0}^q x_{ij} \beta_j = \mathbf{x}'_i \boldsymbol{\beta} \quad (24)$$

であり、

$$\mathbf{y} = \sum_{j=0}^q \mathbf{x}_{(j)} \beta_j \quad (25)$$

である。

また、残差の2乗和は

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta}) (y_i - \mathbf{x}'_i \boldsymbol{\beta}) \quad (26)$$

$$= \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \quad (27)$$

$$= (\mathbf{y} - \hat{\mathbf{y}})' (\mathbf{y} - \hat{\mathbf{y}}) \quad (28)$$

$$= \text{tr}((\mathbf{y} - \hat{\mathbf{y}})' (\mathbf{y} - \hat{\mathbf{y}})) = \text{tr}((\mathbf{y} - \hat{\mathbf{y}}) (\mathbf{y} - \hat{\mathbf{y}})') \quad (29)$$

と書ける。

2.1.2 最小2乗解

残差の2乗和 RSS を回帰係数 β の要素で偏微分すると

$$\frac{\partial \text{RSS}}{\partial \beta_k} = 2 \sum_{i=1}^n (y_i - \hat{y}_i) \times \frac{\partial (y_i - \hat{y}_i)}{\partial \beta_k} \quad (30)$$

$$= 2 \sum_{i=1}^n (y_i - \hat{y}_i) \times \frac{\partial \sum_{j=0}^q x_{ij} \beta_j}{\partial \beta_k} \quad (31)$$

$$= 2 \sum_{i=1}^n (y_i - \hat{y}_i) \times x_{ik} \quad (32)$$

$$= 2 \sum_{i=1}^n (y_i - \sum_{j=0}^q x_{ij} \beta_j) \times x_{ik} \quad (33)$$

$$= 2 \sum_{i=1}^n y_i x_{ik} - 2 \sum_{i=1}^n \left(\sum_{j=0}^q x_{ij} \beta_j \right) x_{ik} \quad (34)$$

$$= 2 \sum_{i=1}^n y_i x_{ik} - 2 \sum_{i=1}^n \left(\sum_{j=0}^q x_{ij} x_{ik} \beta_j \right) \quad (35)$$

$$= 2 \sum_{i=1}^n y_i x_{ik} - 2 \sum_{j=0}^q \left(\sum_{i=1}^n x_{ij} x_{ik} \right) \beta_j \quad (36)$$

を得るが、上記の $q+1$ 個の式を 0 と置いた β の $q+1$ 個の要素に関する連立方程式が正規方程式となる。

今ここで $\sum_{i=1}^n y_i x_{ik}$ は $\mathbf{X}'\mathbf{y}$ の第 k 要素であり $\sum_{i=1}^n x_{ij} x_{ik}$ は $\mathbf{X}'\mathbf{X}$ の第 jk 要素であることを利用すれば、 β の $q+1$ 個の要素の関する連立方程式 $\frac{\partial \text{RSS}}{\partial \beta_k} = 0, k = 0, 1, 2, \dots, q$, は、まとめて

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta = \mathbf{0} \quad (37)$$

と書ける。従って、解は

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \quad (38)$$

となる。

2.2 バリエーション

1. 単回帰分析

基準変数 1 個、説明変数 1 個

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta \mathbf{x} + \varepsilon \quad (39)$$

2. 重回帰分析

基準変数 1 個、説明変数 q 個

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \quad (40)$$

3. 多変量回帰分析

基準変数 p 個、説明変数 q 個

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}' + \boldsymbol{\varepsilon}, \quad \boldsymbol{\beta} = (p \times q) \quad (41)$$

$$\widehat{\boldsymbol{\beta}}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (42)$$

4. 冗長性分析

$$\mathbf{Y} = \mathbf{1}\boldsymbol{\mu}' + \mathbf{X}\mathbf{A}\mathbf{B}' + \boldsymbol{\varepsilon}, \quad \mathbf{A} = (q \times r), \quad \mathbf{B} = (p \times r) \quad r < p, q \quad (43)$$

5. 多項式回帰分析

$$\widehat{y}_i = \sum_{m=0}^M \beta_m x_i^m \quad (44)$$

$$\mathbf{Y} = \mathbf{G}^{(M)}(\mathbf{x})\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbf{G}^{(M)}(\mathbf{x}) = (n \times M + 1) \quad (45)$$

$$\mathbf{G}(\mathbf{x}) = [\mathbf{g}_m(\mathbf{x})], \quad m = 0, 1, \dots, M \quad (46)$$

$$\mathbf{g}_m(\mathbf{x}) = [x_i^m] \quad (47)$$

6. スプライン回帰分析 (多項式 + 切断冪関数)

$$\widehat{y}_i = \sum_{m=0}^M \beta_m x_i^m + \sum_{\ell=1}^K \beta_\ell (x_i - c_\ell)_+^M \quad (48)$$

$$(x_i - c_\ell)_+^M = \begin{cases} 0 & (x_i < c_\ell) \\ (x_i - c_\ell)^M & (x_i \geq c_\ell) \end{cases} \quad (49)$$

$$\mathbf{Y} = \mathbf{G}^{(K,M)}(\mathbf{x})\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \mathbf{G}^{(K,M)}(\mathbf{x}) = (n \times M + K + 1) \quad (50)$$

ただし、 K は節点の数、 c_ℓ は節点

7. ニューラルネットワーク

$$y_i = \beta_0 + \sum_{j=1}^Q \psi(a_j(x_i - b_j))\beta_j + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (51)$$

ただし、

$$\psi(x) = \frac{1}{(1 + \exp(-x))} \quad (52)$$

は係数 a_j, b_j を持つロジスティック関数である。従って

$$\mathbf{y} = \mathbf{G}(\mathbf{x})\boldsymbol{\beta} + \varepsilon \quad (53)$$

の形で書くと

$$\mathbf{G}(\mathbf{x}) = [\mathbf{1}, \psi(a_1(\mathbf{x} - b_1)), \psi(a_2(\mathbf{x} - b_2)), \dots, \psi(a_Q(\mathbf{x} - b_Q))] \quad (54)$$

となる。

スプライン回帰や多項式と似ている点は、両者とも、本来の説明変数を G という形に展開し、それらの線形結合で \mathbf{y} を予測しようと試みることである。

両者の違いは、スプラインの場合は、展開に際して用いられる接点 $c_\ell, \ell = 1, 2, \dots, K$ は定数であるのに対し、ニューラルネットワークにおいては、その係数 $(a_j, b_j), j = 1, 2, \dots, Q$ はパラメタとして推定されるものである点である。

また、多項式回帰の場合、中間層を増やすことは、結局、最初に展開された説明変数の線形回帰を作ることになり、意味がないが、スプラインや、特にニューラルネットワークの場合は、意味がある。

2.3 デザイン行列

水準数が A の 1 元配置の分散分析モデルは、各水準での観測値数を $n_j, j = 1, 2, \dots, A$ とすると、

$$y_{ij} = \hat{y}_{ij} + \varepsilon_{ij}, \quad \hat{y}_{ij} = \mu + \alpha_j, \quad i = 1, 2, \dots, n_j, \quad j = 1, 2, \dots, A \quad (55)$$

である。これを線形モデル $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$ の形で書くと

$$\begin{pmatrix} \hat{y}_{11} \\ \hat{y}_{21} \\ \vdots \\ \hat{y}_{n_1 1} \\ \hat{y}_{12} \\ \hat{y}_{22} \\ \vdots \\ \hat{y}_{1A} \\ \vdots \\ \hat{y}_{n_A A} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_A \end{pmatrix} \quad (56)$$

となる。なお、このデザイン行列は、その第 2 列目から最後までを加えると第 1 列目の単位ベクトルとなるが、このように、行列のある列が、残りの列の線形結合として表されるような場合にその行列の各列は 1 次独立ではない（1 次従属である）と言う。その場合、 $(\mathbf{X}'\mathbf{X})^{-1}$ は存在しない。従って、パラメタ $\boldsymbol{\beta}$ の推定値は一意には定まらない。このデザイン行列のように、各行に一つだけ 1 があり他の要素は全て 0 という行列をダミー行列と呼ぶことがある。

通常の制約条件 $\sum_{j=1}^A \alpha_j = 0$ を入れると、 $\alpha_A = -\sum_{j=1}^{A-1} \alpha_j$ であるから、同じモデルを

$$\begin{pmatrix} \hat{y}_{11} \\ \hat{y}_{21} \\ \vdots \\ \hat{y}_{n_1 1} \\ \hat{y}_{12} \\ \hat{y}_{22} \\ \vdots \\ \hat{y}_{1A} \\ \vdots \\ \hat{y}_{n_{AA}} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & -1 & -1 & \dots & -1 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & -1 & -1 & \dots & -1 \end{bmatrix} \times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{A-1} \end{pmatrix} \quad (57)$$

と書くこともできる。このデザイン行列の各列は 1 次独立である。

2元配置の場合のモデルは

$$y_{ijk} = \hat{y}_{ijk} + \varepsilon_{ijk}, \quad \hat{y}_{ijk} = \mu + \alpha_j + \beta_k, \quad i = 1, 2, \dots, n_{ij}, \quad j = 1, 2, \dots, A \quad k = 1, 2, \dots, B \quad (58)$$

であるが、全ての n_{ij} を 1 とした場合のデザイン行列は

$$\begin{pmatrix} \hat{y}_{111} \\ \hat{y}_{112} \\ \vdots \\ \hat{y}_{11B} \\ \hat{y}_{121} \\ \hat{y}_{122} \\ \vdots \\ \hat{y}_{1A1} \\ \vdots \\ \hat{y}_{1AB} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_A \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_B \end{pmatrix} \quad (59)$$

となる。このデザイン行列の場合、要因 A に対応する列の和、並びに要因 B に対応する列の和がそれぞれ単位ベクトルになり、二つの線形従属関係がある。従って、パラメタの推定値は一意には定まらない。なお、このデザイン行列を $X = \{\mathbf{1}, \mathbf{A}, \mathbf{B}\}$ と分割し、すべてのデータの分だけ生成すると、二つの要因の水準間の同時度数行列 F は

$$F = AB' \quad (60)$$

で与えられる。

通常の制約条件 $\sum_{j=1}^A \alpha_j = 0, \sum_{k=1}^B \beta_k = 0$ を入れると、 $\beta_B = -\sum_{k=1}^{B-1} \beta_k$ であるから、同

じモデルを

$$\begin{pmatrix} \hat{y}_{111} \\ \hat{y}_{112} \\ \vdots \\ \hat{y}_{11B} \\ \hat{y}_{121} \\ \hat{y}_{122} \\ \vdots \\ \hat{y}_{1A1} \\ \vdots \\ \hat{y}_{1AB} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 0 & \dots & 0 & -1 & -1 & \dots & -1 \\ 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & -1 & -1 & \dots & -1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & -1 & -1 & \dots & -1 & -1 & -1 & \dots & -1 \end{bmatrix} \times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_A \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{B-1} \end{pmatrix} \quad (61)$$

と書くこともできる。このデザイン行列の各列は1次独立である。

なお、モデルに交互作用が組み込まれている場合には、交互作用のパラメタ $\alpha\beta_{jk}$ に対応するデザイン行列の列を、 α_j と β_k に対応する X の2つの列 $\mathbf{x}_{(1+j)}$, $\mathbf{x}_{(1+A+k)}$ の要素毎の積を要素として持つベクトルとして定義していけばよい。

例えば、 $A = 2, B = 3$ の場合には以下ようになる。

$$\mathbf{y} = \begin{pmatrix} y \\ y_{111} \\ y_{211} \\ \vdots \\ y_{N_{11}11} \\ y_{112} \\ y_{212} \\ \vdots \\ y_{N_{12}12} \\ y_{113} \\ y_{213} \\ \vdots \\ y_{N_{13}13} \\ y_{121} \\ y_{221} \\ \vdots \\ y_{N_{22}21} \\ y_{122} \\ y_{222} \\ \vdots \\ y_{N_{22}22} \\ y_{123} \\ y_{223} \\ \vdots \\ y_{N_{23}23} \end{pmatrix} \quad \mathbf{X} = \begin{bmatrix} \mu & \alpha_1 & \beta_1 & \beta_2 & \alpha\beta_{11} & \alpha\beta_{12} \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \alpha\beta_{11} \\ \alpha\beta_{12} \end{pmatrix} \quad (62)$$

ただし、たとえば y_{113} の要素のところ入っている -1 は、本来は $y_{113} = \mu + \alpha_1 + \beta_3 + \alpha\beta_{13} + e_{113}$ であるが、そこに、制約条件を使って $\beta_3 = -\beta_1 - \beta_2$ を代入したためである。同様にまた、

y_{121} の要素のところ入っている -1 は、 $\alpha_2 = -\alpha_1$ を代入したためである。また、交互作用 $\alpha\beta_{ij}$ に対応するデザイン行列の列は α_1 と $\beta_j, j = 1, 2$ に対応するデザイン要列の列の積として定義されている。

2.4 セル平均による表現

水準数が A の 1 元配置の分散分析モデルは、各水準での観測値数を $n_j, j = 1, 2, \dots, A$ とすると、

$$y_{ij} = \hat{y}_{ij} + \varepsilon_{ij}, \quad \hat{y}_{ij} = \mu_j, \quad i = 1, 2, \dots, n_j, \quad j = 1, 2, \dots, A \quad (63)$$

である。これを線形モデル $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$ の形で書くと

$$\begin{pmatrix} \hat{y}_{11} \\ \hat{y}_{21} \\ \vdots \\ \hat{y}_{n_1 1} \\ \hat{y}_{12} \\ \hat{y}_{22} \\ \vdots \\ \hat{y}_{1A} \\ \vdots \\ \hat{y}_{n_A A} \end{pmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_A \end{pmatrix} \quad (64)$$

となる。

ここで μ_j は、第 j 群で観測される $y_{ij}, i = 1, 2, \dots, n_j$ の母平均である。(これをセル平均と呼ぶことがある。) この母平均を、前のページでは $\mu_j = \mu + \alpha_j$ と書いている。これと、zero-sum の制約条件を考慮すると

$$\mu = \frac{1}{A} \sum_{j=1}^A \mu_j \quad (\text{母平均の平均}) \quad (65)$$

$$\alpha_j = \mu_j - \mu \quad (\text{母平均の } \mu \text{ からのズレ}) \quad (66)$$

という関係がある。

2 要因モデルの場合には、交互作用の効果に関しても

$$\sum_{j=1}^A \alpha\beta_{jk} = \sum_{k=1}^B \alpha\beta_{jk} = 0 \quad (67)$$

を仮定すると、

$$\mu = \frac{1}{AB} \sum_{j=1}^A \sum_{k=1}^B \mu_{jk} \quad (\text{母平均の平均}) \quad (68)$$

$$\alpha_j = \mu_{j.} - \mu \quad (\text{行平均の } \mu \text{ からのズレ}) \quad (69)$$

$$\beta_k = \mu_{.k} - \mu \quad (\text{列平均の } \mu \text{ からのズレ}) \quad (70)$$

$$\alpha\beta_{jk} = \mu_{jk} - (\mu + \alpha_j + \beta_k) \quad (\text{加算モデルからのズレ}) \quad (71)$$

となる。ただし、

$$\mu_j = \frac{1}{B} \sum_{k=1}^B \mu_{jk}, \quad (\text{行平均}) \quad \mu_{\cdot k} = \frac{1}{A} \sum_{j=1}^A \mu_{jk} \quad (\text{列平均}) \quad (72)$$

である。

2.5 統計的検定

統計的仮定：誤差項は正規分布（平均は 0、分散は共通）

$$y_i = \beta_0 + \sum_{j=1}^q x_{ij} \beta_j + \varepsilon_{ij} = \sum_{j=0}^q x_{ij} \beta_j, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad (73)$$

行列で書くと

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (74)$$

このモデルに関して、以下の仮説検定を考える。

$$H_0 : \beta_{(1)} = \beta_{(2)} = \dots = \beta_{(r)} = 0 \quad (75)$$

$$H_1 : \quad \text{上記の制約がない} \quad (76)$$

このとき、帰無仮説の下で

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1) / r}{\text{RSS}_1 / (N - q - 1)} \quad (77)$$

が第 1 自由度 r 、第 2 自由度 $N - q - 1$ の F 分布をすることが分かっている。ただし、 RSS_k は、仮説 H_k の元でパラメタを推定した際の残差の 2 乗和

$$\text{RSS}_k = \sum_{i=1}^N (y_i - \widehat{y}_{(k)_i})^2 \quad (78)$$

である。（ RSS_0 は H_0 で 0 と指定された回帰係数に対応する説明変数を \mathbf{X} 行列から除いて回帰係数を推定した時の残差の 2 乗和。その時の予測値が $\widehat{y}_{(k)_i}$ ）

これは H_0 が切片を除くすべての回帰係数が 0 という $r=q$ 個の制約の場合

$$F = \frac{\sum_{i=1}^N (\widehat{y}_i - \bar{y})^2 / q}{\sum_{i=1}^N (y_i - \widehat{y}_i)^2 / (N - q - 1)} \quad (79)$$

となる。ただし \widehat{y}_i は H_1 の下での予測値である。

また、一つの回帰係数 $\beta_{(1)}$ のみが 0 という $r=1$ 個の制約の場合も同様の F 検定（第 1 自由度 1、第 2 自由度 $N - q - 1$ ）が使えるが、通常は、それと同値の t 検定（自由度 $N - q - 1$ ）を用いることが多い。

$\text{RSS}_1 = \sum_{i=1}^N (y_i - \widehat{y}_i)^2$ を誤差平方和 (error SS)、 $\text{RSS}_0 - \text{RSS}_1$ を回帰平方和 (model SS)、 $\sum_{i=1}^N (y_i - \bar{y})^2$ を全平方和 (total SS) と呼ぶことがある。また、

$$R^2 = \frac{\sum_{i=1}^N (\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} = \frac{\text{var}(\widehat{\mathbf{y}})}{\text{var}(\mathbf{y})} = 1 - \frac{\sum_{i=1}^N (y_i - \widehat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (80)$$

を決定係数と呼ぶ。