平成 30 年度　農研機構短期集合研修「数理統計（応用編）」
「一般化線型モデル」テキストから抜粋（2018 年 11 月 15 日）

一般化線形モデルと R_D 基準

農研機構農業環境変動研究センター　統計モデル解析ユニット　山村光司
http://cse.naro.affrc.go.jp/yamamura/index.html

1. 一般線形モデル

1-1. 一般線形モデルの推定問題

統計解析で用いられるモデルの多くは次のような行列の形で簡潔に記述できる。これが線形モデルあるいは一般線型モデルと呼ばれている。

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \qquad \mathbf{e} \sim N(0, \sigma^2\mathbf{I}) \tag{1}$$

b は推定すべきパラメーターのベクトル，**X** はデザイン行列と呼ばれる行列である。**e** は 誤差のベクトルであり，平均ゼロ，分散 σ^2 の正規分布にしたがうと仮定されている。単回帰分析，重回帰分析，１元配置分散分析，２元配置分散分析，共分散分析など，すべてこの形をしている。

（例）単回帰分析
n 個のデータに $y = a + bx$ という式をあてはめるとする。これはデータに次のモデルを想定することと同じである。

$$y_i = a + bx_i + e_i \quad (i = 1, 2, ..., n) \tag{2}$$

これは次のように書き表すことができる。

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{3}$$

今の場合は，デザイン行列とパラメーター行列は，それぞれ次式である。

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} a \\ b \end{bmatrix} \tag{4}$$

（例）重回帰分析
説明変数が k 個あり，n 個のデータを用いるとき，モデルは次式である。

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} + e_i \qquad (i = 1, 2, ..., n) \tag{5}$$

これは書き直すと

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{6}$$

この場合も，**X** の各列が線形独立ならば，後に述べる行列演算により何も問題なく推定検定が行なえる。

（例） 1 元配置分散分析

三つの水準を設けて，完全無作為法でそれぞれの処理を 2 回ずつ繰り返したとき，モデルは

$$y_{ij} = T_i + e_{ij} \qquad (i=1,2,3;\ j=1,2)$$

ここに，T_i は第 i 番目の処理(Treatment)の効果を示す。このモデルは次のように書かれることの方が多い。

$$y_{ij} = \mu + T_i + e_{ij} \qquad (i=1,2,3;\ j=1,2)$$

ここに μ は 3 処理全体の基準となる効果である。これは次のように書き表される。

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ T_1 \\ T_2 \\ T_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}$$

（例）二元配置分散分析

二つの要因 (A,B) のそれぞれに 2 水準を設け，それぞれの組み合わせで 2 回ずつ繰り返しを行なった場合，モデルは次のように書き表すことができる。

$$y_{ijk} = \mu + A_i + B_j + AB_{ij} + e_{ijk} \qquad (i=1,2;\ j=1,2;\ k=1,2) \tag{7}$$

ここに，μ は全体の基準となる効果，A_i は要因 A の第 i 番目の水準の効果，B_j は要因 B の第 j 水準の効果を示している。AB_{ij} は，要因 A の水準が i で要因 B の水準が j の場合の平均効果のうち，A_i と B_j のいずれの効果でも説明できない残りの部分（交互作用）を示している。このように定性的要因を 2 要因を含む分析は二元配置分散分析を呼ばれる。このモデルは書き直すと，

$$\begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ A_1 \\ A_2 \\ B_1 \\ B_2 \\ AB_{11} \\ AB_{12} \\ AB_{21} \\ AB_{22} \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ e_{121} \\ e_{122} \\ e_{211} \\ e_{212} \\ e_{221} \\ e_{222} \end{bmatrix} \tag{8}$$

基礎編の「回帰分析」の講義で学習したように，誤差に正規分布を仮定した場合には，最尤推定法は最小二乗法と同じになる。この最小二乗法は行列の演算により簡単に実行できることが知られている。ここでは幾何学的に考えてみよう。まず簡単のため，次のモデルを考える。

$$y_i = a + bx_i + e_i \qquad (i = 1, 2, 3) \tag{9}$$

行列で表現すれば，

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} \tag{10}$$

推定値のセット $\hat{y}_1, \hat{y}_2, \hat{y}_3$ に関しては

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \tag{11}$$

書き直すと

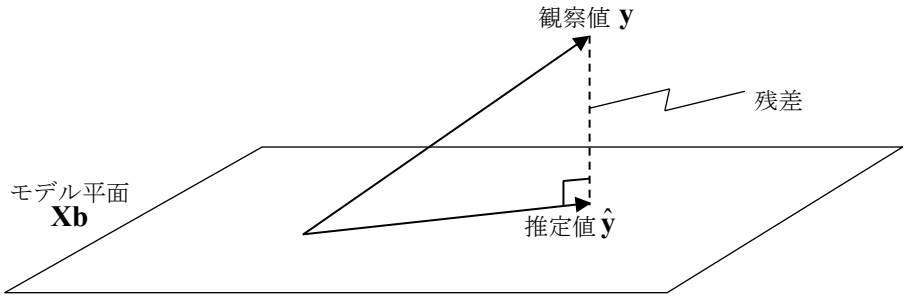
$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = a \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + b \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{12}$$

$\hat{y}_1, \hat{y}_2, \hat{y}_3$ のベクトルは二つのベクトルの定数倍の和で表現されていることから，座標 (0, 0, 0), (1, 1, 1), (x_1, x_2, x_3)の３点を通る平面上の値しかとりえない。ここで観測値の座標 (y_1, y_2, y_3) と推定値の座標 $(\hat{y}_1, \hat{y}_2, \hat{y}_3)$ の空間的な距離を考えてみよう。空間的距離は次の式で表現される。

$$\sqrt{(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2} \tag{13}$$

これは誤差の二乗和の平方根とまったく同じ式である。したがって，最小二乗問題は，観察値と推定値の距離を最小化する問題と同一になる。ピタゴラスの定理より，距離が最小になるのは推定値が観察値から平面に降ろした垂線の足に来たときである。

観察値の数が 3 個より多くなっても次元数が増すだけで上と同じことがいえる。デザイン行列が複雑になっても同じである。一般に仮定より，モデル$\mathbf{y} = \mathbf{Xb}$において，その推定値$\hat{\mathbf{y}}$は多次元平面$\mathbf{Xb}$に貼り付いている。観察値$\mathbf{y}$はその平面から少し離れたところに普通位置しており，この両者の差$\mathbf{y} - \hat{\mathbf{y}}$が残差である。残差平方和を最小にするとは，この残差ベクトルの長さを最



小にすることである。ベクトルの長さが最小になるのは，$\hat{\mathbf{y}}$がちょうど$\mathbf{y}$から多次元平面 $\mathbf{Xb}$に降ろした垂線の足のところにきたときである。観察値から下ろした垂線の足は次の式によって与えられる。

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{14}$$

ここに −1 は逆行列，プライム（′）は転置行列を表す。$\mathbf{y}$の予測値 $\hat{\mathbf{y}}$は$\mathbf{y}$に行列 $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ をかければ得られる。行列 $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ は平面$\mathbf{Xb}$に観察値$\mathbf{y}$の陰を落とす操作を行う行列であるから，「射影行列 (projection matrix) 」とよばれる。あるいは，この行列は$\mathbf{y}$から$\mathbf{y}$の推定値$\hat{\mathbf{y}}$を作る行列でもあるから，ハット (ˆ) を付ける行列ということで，「ハット行列 (hat matrix)」と呼ばれることもある。また，$\mathbf{b}$ の最尤推定値は

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{15}$$

で与えられる。「等分散正規分布誤差を仮定する線形モデル」の最尤推定問題は一言で言えば上の行列演算を行なうことである。また，この $\hat{\mathbf{b}}$ は次の分散共分散行列をもつ多変量正規分布に従う。

$$\mathbf{V} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{16}$$

（注）式の証明

$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$ と書けるので，$\mathbf{y} = \mathbf{X}\hat{\mathbf{b}} + (\mathbf{y} - \hat{\mathbf{y}})$ である。この両辺に $\mathbf{X}'$ をかけると，

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\hat{\mathbf{b}} + \mathbf{X}'(\mathbf{y} - \hat{\mathbf{y}})$$

$\mathbf{y} - \hat{\mathbf{y}}$ は平面へ降ろした垂線なので，直交しており，$\mathbf{X}'(\mathbf{y} - \hat{\mathbf{y}}) = 0$ である。したがって，両辺に $(\mathbf{X}'\mathbf{X})^{-1}$ をかければ

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

また，これを $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$ に代入すれば，

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

一方，

$$\begin{aligned}\mathbf{V} &= E\left[(\hat{\mathbf{b}}-\mathbf{b})(\hat{\mathbf{b}}-\mathbf{b})'\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\left[(\mathbf{y}-\mathbf{X}\mathbf{b})(\mathbf{y}-\mathbf{X}\mathbf{b})'\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

1-2. 線形制約の問題

最小二乗法は行列計算により簡単に実行することができると述べてきたが，実はそこには「デザイン行列が特異でないかぎり」という注釈が必要であった。分散分析の場合は，この注釈が満たされないのが普通である。また，回帰分析の場合でも「多重共線性」という性質がある場合にはこの注釈は満たされない。まず，もっとも簡単な１元配置分散分析の場合を考えてみよう。先ほどの例の場合，次のように書き表すことができた。

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ T_1 \\ T_2 \\ T_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix} \tag{17}$$

このデザイン行列の２列目，３列目，４列目をたすと１列目になり，線形独立ではない。この行列のランクは３であり，ランクの数が列の数よりも少ない。この場合，15式に代入しても $\mathbf{X}'\mathbf{X}$ の逆行列がひとつに決まらないので，このままでは推定ができない。（これをフルランクでない行列，とか，特異行列とかいう。）そこで，ふつう何かの制約条件を設けてデザイン行列の列を線形独立にしてから行列演算を行う。どのような制約をかけても計算できるが，まず最初に次の制約を考える。

$$T_1 + T_2 + T_3 = 0 \quad \text{あるいは} \quad T_3 = -(T_1 + T_2) \tag{18}$$

これは「ゼロ和制約」と呼ばれている。T_3 に関する行は下の２行であり，これが $-T_1 - T_2$ となるように工夫して行列を書き直すと次のようになることがわかるであろう。

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ T_1 \\ T_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix} \tag{19}$$

列が線形独立になったので，あとは先の行列演算でパラメーターの推定検定が行なえる。三つの処理の真の値はそれぞれ $\mu + T_1$, $\mu + T_2$, $\mu + T_3$ である。今のゼロ和制約のもとでは $[(\mu + T_1) + (\mu + T_2) + (\mu + T_3)]/3 = \mu$ であるから，このゼロ和制約条件とは，三つの処理の真の値を平均すると μ となるという条件である。つまり，「三つの処理の真の値の平均値を μ と定義する」という定義を採用しているわけである。なお，制約条件としては，「ゼロ和制約」の他に，「端点制約」というものもよく使われる。これは，パラメーターの一つをゼロとおくことにより（例えば，$T_1 = 0$ とおくことにより），デザイン行列の列を線形独立にしようというやり方である。$T_1 = 0$ と置くときはモデルはつぎのようになる

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ T_2 \\ T_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix} \tag{20}$$

三つの処理の真の値はモデルの定義から，μ, $\mu + T_2$, $\mu + T_3$ であるから，この制約条件においては「T_1 の処理の真の値を μ と定義する」という定義を採用しているわけである。ゼロ和制約は誰でも最も納得しやすい定義であると思われるが，ゼロ和制約をかけてデザイン行列を書き直すのは，若干面倒である。これに対し，端点制約を入れてデザイン行列を書き直すのは非常に簡単である。この場合には，単に，デザイン行列の，ゼロとおいたパラメーターに対応する列を消せば良いだけだからである。

R や SAS など多くの統計ソフトウエアは端点制約を用いている。しかし同じ SAS 社の製品でも JMP はゼロ和制約である。そのため，完全に同じデータで同じ分散分析を行ったとしても，SAS で出力される推定値と JMP で出力される推定値は異なったものになる。その意味では，制約条件に関する理解は非常に重要だとも言える。パラメーターの意味を理解しやすいという点ではゼロ和制約の方が優れていると言えるかもしれない。例えばゼロ和制約では切片は全体平均であるのに対して，端点制約では切片は全体平均ではなく，どれかの処理の値になっており，どこに端点を置くかによって切片の推定値は異なってくる。端点制約で「どこに端点を置くか」については，R では `relevel` 関数，SAS では `ref` オプションを使うことによって変更することができる。

次に２元配置分散分析の場面を考えてみよう。先ほどの例の場合，次のように書き表すことができた。

$$\begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ A_1 \\ A_2 \\ B_1 \\ B_2 \\ AB_{11} \\ AB_{12} \\ AB_{21} \\ AB_{22} \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ e_{121} \\ e_{122} \\ e_{211} \\ e_{212} \\ e_{221} \\ e_{222} \end{bmatrix} \tag{21}$$

このデザイン行列の場合も $\mathbf{X}'\mathbf{X}$ はフルランクでないので，通常次のようなゼロ和制約を加える。まず，A の効果に関しては

$$A_1 + A_2 = 0 \tag{22}$$

A の真の値は $\mu + A_1$，$\mu + A_2$ であるから，その平均は$[(\mu + A_1) + (\mu + A_2)]/2 = \mu$ となる。すなわち，いまの場合，A の真の値の平均値をμと定義していることと同じである。また，B の効果についても同様に定義する。

$$B_1 + B_2 = 0 \tag{23}$$

交互作用は，A の効果と B の効果を除いた余りとして定義すると，次の式を満たしていなければならないことがわかっている。

$$\begin{aligned} AB_{11} + AB_{12} = 0, \qquad AB_{12} + AB_{22} = 0 \\ AB_{21} + AB_{22} = 0, \qquad AB_{11} + AB_{21} = 0 \end{aligned} \tag{24}$$

これは，$[(\mu + A_1 + B_1 + AB_{11}) + (\mu + A_1 + B_2 + AB_{12})]/2 = \mu + A_1$ という形で AB_{11}，AB_{12} などを定義するのと同じ形になっている。三つの制約条件から，残りの一つは自動的に出てくるような性質があるので，事実上は制約条件は三つである。これらの制約を満たすように，一元配置の際に行なったような方式で慎重に先のデザイン行列を変更すると次のようになることがわかるであろう。

$$\begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ A_1 \\ B_1 \\ AB_{11} \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ e_{121} \\ e_{122} \\ e_{211} \\ e_{212} \\ e_{221} \\ e_{222} \end{bmatrix} \tag{25}$$

この行列の列はすべて線形独立であり，$\mathbf{X}'\mathbf{X}$ はフルランクとなるので，これを用いてパラメーターの推定，検定が行なえる。以後の操作は単純な行列計算だけである。上のデザイン行列では，交互作用 AB_{11} の係数は右端の行である。この係数は，A_1 の係数の列と B_1 の係数の列とを

かけたものとなっている。交互作用の係数は，その定義から，一般に，このようにふたつの要因の係数をかけたものとして現われることが示される。

分散分析表で，ある要因の平方和の自由度は通常は（水準数−１）であった。これは，上述のように１個の制約条件をかけているからである，例えばゼロ和制約の場合，平均値からのズレを問題にしているからである。市販の分散分析用のソフトウエアを用いて計算を行う場合には，自動的にゼロ和制約か端点制約がかけられるので，上のような行列の変形を意識しなくても良い。ところが，逆に言えば，これらのソフトウエアを使った場合には，制約条件をかけたくない場合の計算に工夫がいる。例えば，単純な一元配置実験で薬剤 A1, A2 の効果のデータ（処理後−処理前）をとったとき，「薬剤 A1, A2 に差があるか」を調べるのではなくて，「薬剤 A1, A2 のどちらかに効果があるか」どうかを調べるときには，基準値はゼロに決まっているので，別の基準値μを考える意味がなく，したがって制約条件はいらない。この場合，自由度は水準数と等しい。SAS では「noint」，S や R では「-1」と記して「切片なし」を指定する。

(補足) デザイン行列を先に決める場合

多くの場合，実験の際には，まず各処理組合せごとに繰り返しを設けた実験計画を考える。そして，その実験計画の意味するモデルを記述する。そしてその後で，そのモデルに従ってパラメーター行列 **b** とデザイン行列 **X** を書く。しかし，これとは逆にデザイン行列を先に決めてから，それに合うように **b** の配置を考えてモデルを記述するという手順をとることもできるであろう。このような逆の手順をとれば，実験パターンやそのアイデアは限られるかもしれないが，その代わりたいへん省力的な実験ができるはずである。

デザイン行列から始める場合，当然のことながら，良い性質を持つデザイン行列を選択して始めるべきであろう。デザイン行列はいくつかの性質を備えている方が具合いがよい。まずはじめに，（当然のことながら）列が線形独立でなければならない。また，各列が直交している方がよい。デザイン行列 **X** の各列が直交していると， $\mathbf{X}'\mathbf{X}$ は対角要素以外はゼロとなる。このため，あるパラメーターだけを取り上げて推定，検定を行なうことが可能であり，計算が単純となる。この場合，パラメーターを取り込む順番に依存しないのでパラメーターの意味の解釈も簡単である。また，もし二つの列のかけ算が必ずどこかの列になるように行列を設定できたなら，その列は先に示したように交互作用要素のパラメーターになるから，この列のみから交互作用が検定できるので，これもまた非常に具合がよい。

これらの良い性質をもつ行列を表にしたのが「２水準系直交表」である。直交表のデザイン行列のそれぞれの列は (当然のことながら) 線形独立にしてある。また，この「２水準系直交表」では，それぞれの列のかけ算が必ずどこかの列になるようにつくってある。（2^n個の列を持つ形なので，そういう組合せを作ることが常に可能である。）したがって，この直交表をつかえば，交互作用の検定の際にも便利である。直交表には他にもいくつかの種類がある。「３水準系直交表」というものは，これ自体はデザイン行列ではないが，３水準系の実験のために使いやすいように工夫したものである。また「混合型直交表」とよばれるものは，任意の２列のかけ算が他の列にはならないように作ってある。交互作用を全く考えない極めて省力的な実験を行なう場合にはこの混合型が向いているであろう。

（例）Ｌ８型直交表

3 要因 (A,B,C) をそれぞれ 2 水準で試験したい。それぞれの組み合わせで繰り返しを 2 ずつ設けると $2^3\times2=16$ 回の実験を行なわなければならないが，そのような労力はないので，なんとかこれを 10 回以内の実験で済ませたい。また，AB と AC の交互作用のみを考慮したい。このと

き，L8 型直交表を見れば，次の実験計画を立てれば良いことが即座にわかる。

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ A_1 \\ B_1 \\ C_1 \\ AB_{11} \\ AC_{11} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_{222} \end{bmatrix} \tag{26}$$

パラメーターの推定値は先ほどの(3)式で得られ，検定は後述の行列演算で得られる。

1-3. 一般線形モデルの検定問題

省略

1-4. 計算例：ハスモンヨトウのトラップ捕獲数データ

一般線型モデルの概念を知っていれば，定性的要因と定量的要因を同時に扱うことができる。しかし，一般化線型モデルの講義の後で，受講者から次のような質問が寄せられることがある：「両者を同時に扱う必要がないという場面では，昔ながらの解析方法にしたがって，定性的要因については分散分析で分析を行い，定量的要因については回帰分析で分析を行ってよいのではないか？」。しかし，たとえ定性的要因と定量的要因のうちの片方にしか関心がない場合であっても，実際には一般線型モデルによって定性的要因と定量的要因の両方を同時に扱わなければならない場合が多い。ここではこのことを例によって示したい（山村 2009）。

表 1-1 はハスモンヨトウの誘殺数データの一部である。このデータには二つの要因（トラップと月）が含まれている。それらの要因がハスモンヨトウの対数誘殺数 $\log_e(x + 0.5)$ に影響を与えているかどうかを調べたいとする。二つの要因のうち，トラップは定性的要因であり，月は定量的要因であるとする。

表 1-1. ハスモンヨトウのフェロモントラップ誘殺実験結果

トラップ番号	各月の誘殺数			
	５月	６月	７月	８月
1	8	16	55	341
2	16	48	112	874

R で計算を行うためには，まずはデータオブジェクトを作成するとよい。いくつかの方法があるが，データが少量であるならば，データを別ファイルとして保存しておくのではなく，解析プログラムの一部として保存しておくとよいであろう。たとえば以下のようなプログラムが考えられる。

```
cat(file="MothData.txt",
"trap month y
```

```
1 5 8
1 6 16
1 7 55
1 8 341
2 5 16
2 6 48
2 7 112
2 8 874
")

MothData <- read.table("MothData.txt", header=TRUE)
MothData$trap <- factor(MothData$trap)
MothData$trap <- relevel(MothData$trap, ref="1")
```

ここでは，まず cat 関数で MothData.txt という名前のテキストファイルを生成する。（ファイルは作業ディレクトリに保存され，そのディレクトリの場所は getwd() で調べることができる。）次に read.table でそのテキストファイルからデータを読み込んで MothData という名前のデータオブジェクトを作成している。データの一行目は項目名なので header=TRUE として項目名を読み込む。trap は数字として読み込まれ，これを読み込んだ段階では，これは量的変数と認識されている。しかし，今の場合は trap は名義変数である。つまり，1 と 2 には量的な意味はなく，これは単に二つのトラップを識別するための値である。そこで，factor 関数を用いて trap を名義変数に変えている。デフォルト設定では R では端点制約を用いている。その端点については，数字とアルファベットの順番でもっとも先にくる水準がデフォルト設定では端点とされる。今の場合は 1 と 2 では 1 の方が数字の順番で先であるために，デフォルト設定でも自動的に trap の 1 が端点となるが，ここでは relevel 関数を用いて明示的に 1 を端点としている。端点を明示的に指定しておくほうが，後で解釈がしやすくなると思われる。なお，データ内で trap を A,B など文字として記述した場合には，trap 変数は自動的に名義変数として認識される。

R では一般線形モデルの分析は lm 関数で実行する。この lm というのは「linear model（線形モデル）」の略である。まず，すべての説明変数と交互作用を入れたモデルを考えよう。こうしたモデルは「飽和モデル」と呼ばれる。今の場合は，トラップ，月，トラップと月の交互作用をいれたモデルが飽和モデルである。この場合には，たとえば次のようなプログラムとなる。

```
MothData.lmlog <- lm(log(y+0.5)~trap + month + trap:month, data= MothData)
summary(MothData.lmlog)
anova(MothData.lmlog)
```

ここでは実行結果を MothData.lmlog という名前の lm オブジェクトに格納している。オブジェクトの名前は何でもよいが，使用したデータ名と解析の方法が分かるような名前がよいと思う。次に，その lm オブジェクトに summary 関数を適用して，主要な結果を表示させている。さらに anova 関数を適用して Type I の分散分析表を表示させている。lm 関数の中の trap:month はトラップと月の交互作用であるが，このモデルは次のように簡略化して書くこともできる。

```
MothData.lmlog <- lm(log(y+0.5)~trap*month, data= MothData)
```

ここに「*」は最高次までモデルを展開することを意味する簡略記号である。また，model.matrix(~ trap*month, data=MothData) と入力すれば，分析に用いたデザイン行列が表示される。このデザイン行列を見れば，端点制約の置き方を確認することができる。

```
  (Intercept) trap2 month trap2:month
1           1     0     5           0
2           1     0     6           0
3           1     0     7           0
4           1     0     8           0
5           1     1     5           5
6           1     1     6           6
7           1     1     7           7
8           1     1     8           8
attr(,"assign")
[1] 0 1 2 3
attr(,"contrasts")
attr(,"contrasts")$trap
[1] "contr.treatment"
```

今の場合は trap 1 が現れないことから，trap 1 が端点になっていることがわかる。

表 1-1 のデータをいくつかのモデルで分析した結果を表 1-2 に示した。モデル A では，定量的要因（月）を無視して定性的要因（トラップ）に関する分散分析を行っている。モデル B では定性的要因（トラップ）を無視して定量的要因（月）に関する回帰分析を行っている。モデル C では一般線型モデルの考えに従って定性的要因（トラップ）と定量的要因（月）の分析を同時に行っており，モデル D では定性的要因（トラップ）と定量的要因（月）の交互作用までを含めた分析を行っている。トラップ要因に関する有意確率 P 値はモデルによって大きく異なっている。モデル A では有意差は見られないのに対して，モデル C および D では 5%水準で有意となっている。平方和（SS）の部分をみると，該当する要因に対応する平方和はモデル A～D のいずれでも同じであるのに対して，誤差分散の推定値 $\hat{\sigma}^2$ はモデル A で特異的に大きくなっている。

今の場合はモデル A の分散分析の結果は誤りである。一般に，このように一部の要因だけを取り出して分析を行うと結果が誤ったものになる。観測値は次のような成分からなっている。

(観測値) = (要因効果による期待値) + (誤差成分)　　(27)

右辺の第 1 項は観測した要因効果の真の値をすべて用いたときの期待値を表している。手持ちの要因の真の値をすべて組み込んだとしても制御できない変動成分があり，これが右辺の第 2 項である。モデル A の分散分析では要因効果のうちのトラップの成分だけを取り込んで，残りは誤差成分とみなしてしまっている。これでは誤差成分を過大に推定してしまう。モデル A の分散分析で有意差が見られなかったのはそのためである。誤差成分の推定に関しては，誤差成分の自由度の問題もあるが，できるかぎりたくさんの要因を考慮して推定を行うべきであろう。今のデータの場合は，すべての要因を組み込むと，例えば表 1-2 のモデル D のような分析となり，分散分析と回帰分析を組み合わせた「一般線形モデル」となる。

表 1-2. ハスモンヨトウのデータの対数変換値 $\log_e(x + 0.5)$に 5 種類のモデルを当てはめた分析結果。k はモデルに含まれるパラメーター数，df は自由度，SS は Type I 平方和，$\hat{\sigma}^2$ は誤差分散の推定値を示す。

変動因	df	SS	F値	P値	$\hat{\sigma}^2$
A. トラップに関する分散分析 ($k = 2$)					
トラップ	1	1.44	0.5	0.50	2.73
誤差	6	16.36			
B. 月に関する回帰分析 ($k = 2$)					

月	1	15.68	44.5	<0.01	0.35
誤差	6	2.11			
C. トラップと月に関する一般線形モデル分析 ($k = 3$)					
トラップ	1	1.44	10.6	0.02	0.14
月	1	15.68	115.4	<0.01	
誤差	5	0.68			
D. 交互作用までを入れた一般線形モデル分析 ($k = 4$)					
トラップ	1	1.44	8.5	0.04	0.17
月	1	15.68	93.1	<0.01	
月×トラップ	1	0.01	0.0	0.87	
誤差	4	0.67			
E. 切片のみを当てはめた分析 ($k = 1$)					
誤差	7	17.80			2.54

2. 等分散性の問題

省略

3. 一般化線型モデル

省略

4 . モデル選択とモデル評価

4-1. モデル選択手順としての有意性検定

かつて Fisher（1922, p313 ; 1973a, p8）は，統計処理の目的は「データの縮約（reduction of data）」にあるとして，次の三つのプロセスを示唆していた：（1）*P* 値を用いた有意性検定によるモデル同定の問題（Problems of specification）（2）同定されたモデルのパラメーター推定の問題（Problems of estimation）（3）推定されたパラメーターの推測分布（fiducial distribution）の問題（Problems of distribution）。Fisher（1973b, p52）は「仮説検定の段階では推定理論は必要ない」と述べており，有意になったパラメーターについてのみ，次のステップとしてパラメーターの推定に進むとしていた。つまり，Fisher（1973b）にとって有意性検定は統計処理の最終目標ではなく，有意性検定は推定の前に行う一時的な作業に過ぎなかった。しかし，こうした Fisher の意図に反して，有意性検定を統計処理の最終目標だと解釈する人々が実際には非常に多い。このような「検定偏重」の風潮が生じたのは，Fisher（1973a）自身がその著書の中で有意性検定の使用を強調しすぎたからであろうと Yates（1951, p32）は指摘している。

検定が「統計処理の最終目標」になりえないことについては，Fisher 以降も一部の統計学者によって繰り返し指摘されてきた。検定で有意差が出なかったときには「差を検出するのにサンプル数が足りなかった」ことを示しているにすぎず，一方，有意差が出たときには「差を検出するのにサンプル数が十分に多かった」ことを示しているにすぎない。つまり，統計的な有意差の有無は，単に私らが用いたサンプル数の大きさによって決まる問題であり，それは私らが探求している真実とは無関係であるとも言える．AIC の提案者である赤池（1976）は，「あ

るサイコロの正しさを検定するという問題も全く同様で，現実のサイコロで完全に対称なものが存在しえないことは明らかである。（中略）データによる検定結果を待つまでもなく結論は見えている」とし，検定の「論理的矛盾」を指摘していた。「日本の品質管理の父」と言われる Deming（1975）も，「わたしらは二つのムギ品種や二つの薬が等しいことを見出すために実験を行っているのではない。実験にお金を費やすまでもなく，それらが等しくないことは最初からわかっている」と指摘していた。最近になって，ようやくアメリカ統計学会が公式にこの問題を認知しはじめたようである（Baker 2016；Wasserstein and Lazar 2016）。

上の Fisher の第 1 段階部分は「モデル選択」を意味しており，現代ではこの部分は有意性検定ではなく AIC や R_Dによるモデル選択によって置き換えることができる（山村 2018）。現在は，まだまだ「検定偏重」の時代であるが，こうしたアメリカ統計学会の動きを受けて，今後は「モデル選択と推定」の時代へとゆっくりとシフトしてゆくのではないかと予想される。これは，ある意味では Fisher への原点回帰だとも言える。

4-2. AIC によるモデル選択

モデル選択の手段として，現在ではいくつかの指標を用いることができる。いま手元のデータによく当てはまるモデルは，次にデータをとったときに，その新しいデータにもうまく当てはまるとは限らない。そこで，次にデータをとったときの確率分布が真の確率分布に Kullback-Leibler 情報量の尺度でもっとも近くなるようにモデルを選択することを考えて赤池氏は情報量基準 AIC を導いた。最尤推定法でモデルを当てはめときの尤度をLとし，そのモデルに含まれるパラメーター数（切片を含む）をkと書くとき，AIC は次のように定義される。

$$\mathrm{AIC} = -2\log(L) + 2k \tag{28}$$

等分散正規分布の線型モデルで誤差項が一つだけの場合に AIC の小標本の偏りを修正するものとして AICc と呼ばれる修正版が Sugiura（1978）によって提案された。AICc は海外では線型モデル以外の場合にも無批判に用いられている。また，すべてのパラメーターが事前分布にしたがって変動するという仮定に基づき，AIC とは完全に別の考えから BIC（Bayesian Information Criterion）という指標が導かれており，こちらも AICc と並んでよく用いられている

R では AIC は `AIC` 関数を用いて出力できる。第 1-4 節で計算したハスモンヨトウのデータ分析において，すべての説明変数を考慮したモデル（飽和モデル）の場合には，`lm` オブジェクト `MothData.lmlog` に `AIC` 関数を適用することにより `AIC(MothData.lmlog)` で求められ，この出力は `12.90994` である。なお，対数尤度の値は `logLik(MothData.lmlog)` で計算され，その出力は`-1.454968 (df=5)` である。ここに `df=5` はモデルのパラメーター数を意味しており，切片，トラップ，月，トラップと月の交互作用，分散パラメーターの 5 個のパラメーターのことを指している。これを(68)式に代入すれば，`-2*(-1.454968)+2*5= 12.90994` であり，AIC が `12.90994` であることを確認できる。しかし，この AIC の値自体には実用的には意味がなく，他のモデルをあてはめたときの AIC と比較したときにだけ意味がある。そこで，月の量的変数だけを入れた単回帰分析の場合を計算しよう。

```
MothData.reglog <- lm(log(y+0.5)~ month, data= MothData)
AIC(MothData.reglog)
```

この値は `AIC = 18.05726` であり，飽和モデルの方が AIC が小さいことから，飽和モデルの

方が AIC 基準において優れていることがわかる。

AIC を計算する別の関数として MASS ライブラリに extractAIC がある。この関数で計算すれば，飽和モデルでは extractAIC(MothData.lmlog)によりパラメーター数 4 で AIC = -11.79308 となり，月の単回帰では extractAIC(MothData.reglog)によりパラメーター数 2 で AIC= -6.645757 となる。これらの値は先ほど AIC 関数で計算した結果と全く異なる。extractAIC 関数では（1）対数尤度の定数部分を省略しており，かつ（2）分散パラメーターをパラメーター数として数えていない，という二つの特徴があるために計算値が異なっている。extractAIC(MothData.lmlog)で実施しているのは次の計算である。

```
n <- 8; k <- MothData.lmlog$rank
RSS <- sum(MothData.lmlog$residuals^2)
n*log(RSS/n) + 2*k
```

(68)式のもともとの定義どおりの AIC の値は AIC 関数で計算される値の方である。しかし，extractAIC 関数を用いて自動的に最良モデルを選択する関数として stepAIC 関数があり，この stepAIC を使えば最適なモデルを自動的に探索してくれるので非常に便利である。先ほどの飽和モデルの場面で stepAIC を使ってみよう。

```
library(MASS)
stepAIC(MothData.lmlog)
```

この出力は以下のとおりである。ここで計算される AIC は extractAIC 関数によるものなので，もともとの(68)式による定義の AIC とは定数だけ異なっている。

```
Start:  AIC=-11.79
log(y + 0.5) ~ trap * month

             Df Sum of Sq     RSS     AIC
- trap:month  1 0.0052762 0.67916 -13.731
<none>                    0.67389 -11.793

Step:  AIC=-13.73
log(y + 0.5) ~ trap + month

        Df Sum of Sq     RSS      AIC
<none>                0.6792 -13.7307
- trap   1    1.4351  2.1143  -6.6458
- month  1   15.6815 16.3606   9.7235

Call:
lm(formula = log(y + 0.5) ~ trap + month, data = MothData)

Coefficients:
(Intercept)         trap2         month
    -4.4414        0.8471        1.2523
```

探索の結果として，交互作用項を除いてトラップと月の主効果の項だけを含むモデルが選択される。先ほど議論したように，交互作用は主効果で説明できない残りの成分を説明するためのものであった。そのため，交互作用項を残したまま主効果を除いたモデルには普通は意味がない。交互作用の係数など劣位のパラメーターを組み込む際には，普通はそれに関係する優位のパラメーターを同時にすべて組み込んでおく必要がある。そうした「意味のあるモデル群」を Yamamura（2016）では「階層型モデル群（hierarchical family）」と呼んだ。モデル選択においては，このような「階層型モデル群」に絞って比較を行わなければならない。先述のように，有意性検定を用いてモデル選択を行う場合には，検定のタイプ（Type I, II, III など）を変えることにより階層型モデル群を考慮することができた。上の試行錯誤の経過を見れば，stepAIC

関数は，こうした階層型モデル群に絞って探索を行ってくれることが分かる。いわば `stepAIC` 関数は自動的に Type の判別を行ってくれると言える。

4-3. AIC の問題点

上の例では AIC を計算してきたが，AIC が根拠としている「予測における Kullback-Leibler 情報量の尺度で測った近さ」とはそもそも何を意味するのであろうか。この尺度を考えれば計算が簡単になったという事情もあるようだが「Kullback-Leibler 情報量の尺度」は現実的には意味不明だともいえる。たとえば，AIC=15.2 という値が出たときに，この値（15.2）自体には意味はない。AIC の使用においては，同じデータのもとで二つ以上のモデルの AIC を比較した場合にのみ意味があった。それらのモデルのうちで AIC が最小となるモデルを採用するのが普通である。しかし AIC や BIC は，その選択された最良モデルが「どれだけ良いのか」については何も示してくれない。選ばれたモデルは所定の基準では確かに最良モデルではあるが，それはほとんど役に立たないモデルだという場合もある。特に，データの量や質が不十分な場合には，そのようなことが普通に生じる。現在のデータの量や質が十分であるかどうかを判断するためには，「モデル選択」だけでなく「モデル評価」を行うことが極めて重要である。しかし，AIC や BIC を用いた場合には「モデル選択」はできても「モデル評価」を行うことができない。

また，一つのモデルに頼るのは良くないとして，モデル平均化（model averaging）として，AIC や BIC で重み付けしたパラメーターを採用する人々が最近では増えている（たとえば Claeskens and Hjort 2008）。しかし，モデルのパラメーター定義を変えれば，実際には論理的に完全に同じモデルであっても推定されるパラメーター値が異なってくる（Yamamura 2016）。モデル平均化アプローチはその基本的な考え方において大きな間違いを包含しているようである。

4-4.「実際に当たる確率」によるモデル評価

予測力でモデルを評価したい場合には，むしろ「実際に当たる確率」で評価するべきではないだろうか。ところが「実際に当たる確率」で評価しようとすると「Kullback-Leibler 情報量の尺度」で測るよりも問題は複雑になる。というのも「確率とは何か」という哲学を持ち出さなければならなくなるからである。もともと，尤度の式と確率の式は同一だが「尤度は確率ではない」ことに注意しなければならない。例えば，いま，あるデータが生じる確率について，二項分布を仮定した場合と正規分布を仮定した場合とで比較したとしよう。この比較は「尤度の比較」にはなっている。しかし，これは「確率の比較」ではない。というのも，確率の定義が異なっているために，この二つの量は確率としては比較できないのである。確率として比較できないのであるから「実際に当たる確率」を比較することはもちろん不可能である。第 2 節で述べたように，パラメーターには優劣関係がある。確率そのものの定義にかかわるパラメーターは他のパラメーターより先に推定して，それを固定しなければならない。

Yamamura（2016）は，確率の定義を固定した上で「実際に当たる確率」で評価を行うことを提案し，この考え方に基づいて R_D という指標の使用を提案した。ここで「確率の定義の固定」においては，Laplace（1825）の考えから「手持ちの全知識を組み込んでも予測できない残りの部分」とする。知識は技術の進歩などにより変化するため，確率の定義は時代とともに変化する。すなわち真の確率は手持ちの知識に依存して変化する。このため，確率は主観的な存在だと間違われることも多い（例えば de Finetti 1937）。しかし，確率は手持ちの情報のもとで客観的に決まるのであって「私がこう思うから」とか「私がこう確信するから」といった理由で主

観的に決めてよいものではないであろう。後にマクロ経済学を確立させることになる Keynes（1921）の言葉を借りれば「知識を決定する事実がひとたび与えられれば，この環境でどれが起こりそうでどれが起こりそうでないかは客観的に決まるのであって，それは個人のオピニオンとは無関係である。」Laplace（1825）はハレー彗星の例をあげていた。昔はハレー彗星がいつ出現するかは不明であり，その出現は確率現象（probabilistic events）であった。しかし，現代では知識の進歩により，いつハレー彗星が出現するかを計算することができる。現代ではハレー彗星の出現は確率現象ではなく決定論的現象（deterministic events）になっている。科学の進歩により確率成分が減少して，ほとんどゼロにまで小さくなったといえる。降水確率の計算についても同様であろう。100 年後の将来の技術で計算される降水確率は，仮にまったく気象条件が同一であったとしても，現在の技術で計算される降水確率とは異なるであろう。将来の技術で計算される確率が「真の確率」であって，現在の技術で計算される確率が「間違った確率」だというわけではない。どちらの確率も，その時代の情報のもとで正しい。このように「真の確率」は手持ちの知識に依存して変化する。そして「手持ちの全知識を組み込んでも予測できない残りの部分」が「真の確率」であるから「手持ちの全知識を組み込んだモデル」が Laplace 確率においては必然的に「真のモデル」である。

一般に，モデルは手持ちのデータを記述することだけを目的とするのではなく，何らかの別のデータにも適用できることを暗黙の前提としている。こうしたモデルの性質から考えれば，予測力でモデルの妥当性を評価するというのは必然的な評価法だと言えるであろう。たとえモデルが「真のモデル」であったとしても，その予測力が低ければ，その「モデルとしての価値」は低いと考えられる。

R_D 指数は以下のようにして導かれている。まず，手持ちのデータを用いてモデルを構築して，次にそのモデルを新しいデータに適用する場面を考える。このときに「構築したモデルの元でその新しいデータが発生する確率」の対数値として予測力を定義する。そして，その予測力の改善割合 R_{pred} を考える。将来のデータをすべて知っている「神」がわれわれの知る誤差構造を用いて予測した場合に R_{pred} は 100%となり，説明変数をまったく持たない「凡人」が同様にして予測した場合に R_{pred} が 0%となるように改善割合 R_{pred} は定義される。いわば「後出しジャンケン」ができる場合に予測力が 100%となるように尺度化されていることになる。この R_{pred} の推定値として R_D が導出されている。

$$R_D = 1 - \frac{l_{\max} - l + k}{l_{\max} - l_{\mathrm{null}} + \theta} \tag{29}$$

ここにkはR_Dの計算対象とする候補モデルに含まれるパラメーター数である。lは候補モデルでの対数尤度であり，l_{null}は切片だけを含むモデル（null model）における対数尤度である。$l_{\max}$は固定効果パラメーター数とデータ数が等しい最大モデル（maximal model）における対数尤度である。ただし，すべての要因効果を入れたモデルのもとで推定した分散 $\hat{\sigma}^2$ を用いて l, l_{null}, $l_{\max}$は計算される。その点で，計算手順は AIC と違ってかなり複雑になる。θは切片の数であり，1 変量の場合は$\theta = 1$である。

4-5. R_D 指数の使い方

表 4-1 には表 1-1 のハスモンヨトウのデータをさまざまなモデルで分析して指数を比較した結果が示されている。AICc の値を比較すると，これはモデル B で最小になることから，AICc を用いた場合にはモデル B が採択される。一方，BIC を用いた場合には同様の比較からモデル

C が採択される。この AICc と BIC の計算においては確率成分の推定値は固定されておらず，モデル毎に異なる誤差推定が行われている。このようにモデル毎に別々の誤差推定を行うと，検定や推論で誤りが生じやすいことが，以前から経験的に知られていた。Draper and Smith（1966）は「すべての要因を組み込んだ後の誤差」を「純誤差（pure error）」と呼び，推論は常に純誤差に基づくべきだと指摘している。R_Dの計算では，確率の比較を可能とするために，モデル選択に先立って確率分布の推定が行われる。このため，確率成分の推定については結果的に Draper and Smith（1966）と同様の手順となる。

表 4-1 を見ると，今の場合はモデル C において R_Dがもっとも大きく，予測力の改善割合の推定値は $R_D = 0.91$ である。予測力の改善割合は十分に大きいといえる。ただし，**予測力を問題にする場合であっても，必ずしも R_Dが最大となるモデルを採択する必要はない。順位が 2 位以下のモデルであっても，最良モデルと比較してあまり R_Dが低下しておらず，かつ，利用しやすい性質を持っているモデル（たとえば，容易に測定できる説明変数からなるモデルや，容易に解釈できるモデル）であれば，そちらのモデルを採択するべきである。**Yamamura et al.（2018）では，土壌中の放射性セシウムが玄米へ移行する際の「移行係数」を管理する式を提案する際に，「最良でないモデルを採用する根拠」として R_Dを用いた。

なお，予測が目的の場合にはなるべく大きな R_Dを示すモデルを採択するべきであるが，予測が目的ではなく「変数が生じる主な理由を把握する」のが目的の場合には，R_Dが大きいモデルよりも，あえて R_Dが 0.8 程度のモデルを採用するのが好ましい場合もあるであろう。今の場合は，月だけを用いたモデル（モデル B）で $R_D = 0.85$ であり 0.8 に近いことから，モデル B を「要約モデル」として採用し，「ハスモンヨトウの個体数は主として月によって決まっている」と解釈するのも妥当であろう。

また，もし R_Dの最大値がたいへん小さく，たとえば $R_D = 0.3$ といった具合であったならば，その場合はデータの量や質が悪すぎることを意味している。したがって，そのようなときには，その段階での最良モデルを報告するのではなく，データ量を増やしたり，別の説明変数を準備したりして，データ自体を改善して再解析を行ってから最良モデルを報告する必要があるであろう。

4-6. R_D 指数の計算プログラム

R_Dを計算するための 1 変量用の R 関数（`RDcompare`）および SAS マクロ（`RDcompare`）が以下のサイトにおいてある。論文の著者版原稿もここに置いてある。

http://cse.naro.affrc.go.jp/yamamura/RD_criterion_program.html

この R 関数を使えば，一般化線型モデルにおいて R_Dの計算を自動的に行うことができる。その計算の際には表 4-1 にあるような「階層型モデル群」だけを比較の対象とし，モデルに順位をつけ，そのモデルの予測力を計算してくれる。

表 4-1.ハスモンヨトウのデータの対数変換値 $\log_e(x + 0.5)$に 5 種類のモデルを当てはめた分析結果。kはモデルに含まれるパラメーター数，df は自由度，SS は Type I 平方和，$\hat{\sigma}^2$ は誤差分散の推定値を示す。R_Dの計算では，検定の場合と同様に，もっとも多くの要因効果を含むモデル（モデル D）から計算された不偏分散推定値（$\hat{\sigma}^2 = 0.17$）を使用している。一方，AICc の計算ではモデル毎に別々の σ^2 を最尤推定するため，パ

ラメーター数 k は一つ多くなる。

変動因	df	SS	F値	P値	$\hat{\sigma}^2$	R_D	AICc	BIC
A. トラップに関する分散分析 ($k = 2$)								
トラップ	1	1.44	0.5	0.50	2.73	0.06	40.43	34.66
誤差	6	16.36						
B. 月に関する回帰分析 ($k = 2$)								
月	1	15.68	44.5	<0.01	0.35	0.85	24.06	18.30
誤差	6	2.11						
C. トラップと月に関する一般線形モデル分析 ($k = 3$)								
トラップ	1	1.44	10.6	0.02	0.14	0.91	24.31	11.29
月	1	15.68	115.4	<0.01				
誤差	5	0.68						
D. 交互作用までを入れた一般線形モデル分析 ($k = 4$)								
トラップ	1	1.44	8.5	0.04	0.17	0.89	42.91	13.31
月	1	15.68	93.1	<0.01				
月×トラップ	1	0.01	0.0	0.87				
誤差	4	0.67						
E. 切片のみを当てはめた分析 ($k = 1$)								
誤差	7	17.80			2.54	0	35.50	33.26

`RDcompare` 関数を使用可能にするために，まずテキストファイル RDcompare.txt をウェブからダウンロードして，それを R の作業ディレクトリにコピーしておく。（作業ディレクトリの場所が不明であれば `getwd()` で調べることができる。）それを `source` 関数で最初に読み込んでおく。そうすれば `RDcompare` 関数が使用可能となる。

```
source("RDcompare.txt")
```

あるいは，`file.choose()` 関数を用いれば，RDcompare.txt の場所をブラウザから指定することができる。

```
source(file.choose())
```

R_D を計算する計算プログラムは例えば次のようになる。

```
RDcompare(log(y+0.5)~trap*month, data= MothData)
```

モデル式としては「飽和モデル」あるいは「最も複雑なモデル」を指定する。飽和モデルというのは，先述のように「すべての要因効果を含んだモデル（手持ちの情報をすべて組み込んだモデル）」のことであり，今の場合は，`trap*month` が飽和モデルである。これは `trap + month + trap:month` と同じ意味である。`ShowModel` オプションを使えば上位モデルのパラメーター推定値とその標準誤差を表示できる。デフォルト設定は `ShowModel=5` であり，上位 5 モデルのパラメーターが表示される。プログラム出力は次のようである。

```
# RD ranking for the hierarchical family of models #
          RD        RSD                                    Model
1 0.90679867 0.94179985             log(y+0.5) ~ 1 + trap + month
```

```
2 0.88850757 0.92280275 log(y+0.5) ~ 1 + trap + month + trap:month
3 0.84623461 0.87889810                     log(y+0.5) ~ 1 + month
4 0.06056406 0.06290175                      log(y+0.5) ~ 1 + trap
5 0.00000000 0.00000000                             log(y+0.5) ~ 1

# Parameters of the best 5 model #
# Model 1 #
              Estimate Std. Error
(Intercept) -4.4413787  0.8682802
trap2        0.8470931  0.2902341
month        1.2522566  0.1297966
# Model 2 #
              Estimate Std. Error
(Intercept) -4.2920744  1.2106622
trap2        0.5484844  1.7121350
month        1.2292867  0.1835601
trap2:month  0.0459398  0.2595932
# Model 3 #
             Estimate Std. Error
(Intercept) -4.017832  0.8560675
month        1.252257  0.1297966
# Model 4 #
             Estimate Std. Error
(Intercept) 3.6982893  0.2052265
trap2       0.8470931  0.2902341
# Model 5 #
  Estimate Std. Error
  4.121836   0.145117
```

一方，SAS マクロでは一般化線型モデルだけでなく一般化線形混合モデルでも R_D の計算を自動的に行うことができる。表 1-1 のような個体数のデータは正確には一般化線形混合モデルで扱うのが妥当であり，$\log_e(x + 0.5)$で扱うのはあくまでも近似である（山村 2009；Yamamura 2016；山村・鈴木 2006）。ここでは近似を使用せずに厳密に扱ってみよう。表 1-1 のデータを一般化線形混合モデルで計算するための SAS プログラムは次のようになる。

```
title 'Trap data of Oriental leafworm moth';
data MothData;
input Trap Month y;

datalines;
1  5  8
1  6  16
1  7  55
1  8  341
2  5  16
2  6  48
2  7  112
2  8  874
;

* Specify the location of RDcomparSAS.txt.;
%inc '/folders/myfolders/RDcompareSAS.txt' / nosource; run;
* RD ranking for Poisson GLMM, using RDcompare.;
%RDcompare(data = MothData,
       class = Trap(ref="1"),
       DepVar = y,
       TrueModel = Trap Month Trap*Month,
       dist = poisson,
       link = log,
       ShowModel = 2)
```

この出力は次のようである。log(x+0.5) 変換による計算は，この一般化線形混合モデルの良い

近似になっていることがわかる。

RANK	ModelDF	RD	RSD	Model
1	3	0.93664	0.97052	Trap Month
2	4	0.92718	0.96072	Trap Month Trap*Month
3	2	0.86550	0.89680	Month
4	2	0.05197	0.05385	Trap

Effect	Trap	Estimate	StdErr
Intercept		-5.0823	0.7347
Trap	2	0.8832	0.2152
Trap	1	0	.
Month		1.3390	0.103

Effect	Trap	Estimate	StdErr
Intercept		-5.1392	1.1124
Trap	2	0.9811	1.4515
Trap	1	0	.
Month		1.3472	0.1583
Month*Trap	2	-0.01425	0.2088
Month*Trap	1	0	

5. 引用文献

赤池弘次 (1976) 情報量基準 AIC とは何かーその意味と将来への展望ー. 数理科学, 153:5-11

Baker M (2016) Statisticians issue warning on *P* values. Nature, 531:151

Claeskens G, Hjort NL (2008) Model Selection and Model Averaging. Cambridge University Press, Cambridge, UK

de Finetti B (1937) Foresight: its logical laws, its subjective sources (translated and reprinted in Kyburg HE, Smokler HE (eds) Studies in Subjective Probability, 97–158. Wiley, New York, 1964). Annales de l'Institut Henri Poincaré, 7:1–68

Deming WE (1975) On probability as a basis for action. American Statistician, 29:146–152

Draper NR, Smith H (1966) Applied Regression Analysis. Wiley, New York

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London Series A, Mathematical and Physical Sciences, 222:309–368

Fisher RA (1973a) Statistical Methods for Research Workers. 14th edn. Hafner, New York

Fisher RA (1973b) Statistical Methods and Scientific Inference. 3rd edn. Hafner Press, New York

Keynes JM (1921) A Treatise on Probability. Macmillan and Co, London

Laplace PS (1825) A Philosophical Essay on Probabilities (Translated from the fifth French edition of 1825 by Andrew I. Dale, 1995). Springer, New York

Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. Communications in Statistics - Theory and Methods, A7:13–26

Wasserstein RL, Lazar NA (2016) The ASA's statement on *p*-values: context, process, and purpose. The American Statistician, 70:129-133

山村光司 (2009) 一般化線型モデルとモデル選択 ー統計解析の新しい流れー. 植物防疫, 63:324-329 <http://www.jppa.or.jp/shiryokan/pdf/63_05_46.pdf>

Yamamura K (2016) Estimation of the predictive ability of ecological models. Communications in Statistics - Simulation and Computation, 45:2122–2144

山村光司 (2018) ベイズ推定法の適切な活用についてーエゾシカ個体数推定の例ー. 保全生態学研究, 23:39-56

山村光司・鈴木芳人 (2006) 農薬の効果判定：密度指数と補正密度指数. 植物防疫, 60:112-116

<http://www.jppa.or.jp/shiryokan/pdf/60_03_14.pdf>
Yamamura K, Fujimura S, Ota T, Ishikawa T, Saito T, Arai Y, Shinano T (2018) A statistical model for estimating the radiocesium transfer factor from soil to brown rice using the soil exchangeable potassium content. Journal of Environmental Radioactivity, 195:114–125
Yates F (1951) The influence of "Statistical methods for research workers" on the development of the science of statistics. Journal of the American Statistical Association, 46:19–34