一般化線形モデルへの道

ほんのさわりとして

三中 信宏

minaka@affrc.go.jp
http://cse.niaes.affrc.go.jp/minaka/

租界Rの門前にて：統計言語「R」との極私的格闘記録
http://cse.niaes.affrc.go.jp/minaka/R/R-top.html

〒 305-8604 茨城県つくば市観音台 3-1-3
独立行政法人 農業環境技術研究所 生態系計測研究領域 上席研究員
東京大学大学院 農学生命科学研究科 教授［生態系計測学］
京都大学大学院理学研究科　連携併任教授［進化生物学］
東京農業大学大学院 農学専攻 客員教授［応用昆虫学］

これまで説明してきた「分散分析」（analysis of variance）は，ある因子変数による測定データへの効果（処理効果）を，誤差が独立かつ同一の正規分布にしたがうという仮定のもとで検定する方法である．伝統的な手法としての分散分析は, データに対する「線形モデル」（linear model）をあてはめるという点で回帰分析や共分散分析と共通する部分が多い．さらに，誤差が必ずしも正規分布に従わないようなケースや, あてはめる関数のタイプをより広げた「一般化線形モデル」（generalized linear model）がいま広範囲に用いられるようになってきた．

ここでは，Rを用いた分散分析（「**aov**」コマンド）・線形モデル（「**lm**」コマンド）・一般化線形モデル（「**glm**」コマンド）を比較することにより，一般化線形モデルへの道程を示そう．もとより「一般化」という言葉はユーザーにとって必ずしもラクな人生を確約するものではない．むしろ，さまざまなオプション設定がユーザー側に委ねられているため，要求される知識や責任は大きくなるだろう．しかし，従来的な統計手法を縛ってきた誤差に関する制約（正規性や等分散性など）を取り除き，より現実に即した統計解析を進める上で，一般化線形モデルは有力な選択肢のひとつといえるだろう．

一般化線形モデルの「山」にいきなり登攀するのはリスクが大きすぎるかもしれない．すでに知っている分散分析を足がかりにして登り始めよう．

【1】復習：分散分析コマンド（**aov**）を利用する

```
  R : Copyright 2007, The R Foundation for Statistical Computing
  Version 2.5.1  (2007-6-27), ISBN 3-900051-07-0
```

```
> mm <- read.table("Box1_R.txt", header=T)
                          # データ Box1_R.txt を呼び出し，mm に格納.
> attach(mm)              # mm をデータフレームとして指定する.
> TRT <- factor(TRT)      # 因子指定．attach してあるので「mm$」は不要.
> fm <- aov(DATA ~ TRT)   # 1 要因完全無作為化法の分散分析をする.
                          # attach してあるので「data=mm」は不要.
```

```
> summary(fm)             # 分散分析表の表示
             Df  Sum Sq Mean Sq F value    Pr(>F)
  TRT         6 5587175  931196  9.8255 3.329e-05 ***
  Residuals  21 1990237   94773
```

完全無作為化法による実験計画では，観察されたデータが次の線形モデル：

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij} (\epsilon_{ij} \sim N(0, \sigma^2))$$

にしたがって互いに独立に生じていると仮定している．この正規性ならびに等分散性の仮定の下で，上の分散分析表は，帰無仮説 H_0「$x_{ij} = \mu + \epsilon_{ij} (\forall i, \alpha_i = 0)$」とその対立仮説である H_1「$x_{ij} = \mu + \alpha_i + \epsilon_{ij} (\exists i, \alpha_i \neq 0)$」とを対置して，分散比に関する F 検定を行なっている．同様の内容は次のコマンドによっても表示させることができる.

```
> anova(fm)               # 分散分析表の表示（ここでは summary と同じ内容）
  Analysis of Variance Table
  Response: DATA
            Df  Sum Sq Mean Sq F value    Pr(>F)
  TRT        6 5587175  931196  9.8255 3.329e-05 ***
  Residuals 21 1990237   94773
```

```
> detach()                # mm のデータフレーム指定を解除する.
> rm(list=ls())           # ワークスペースの掃除（オブジェクト削除）.
```

【２】線形モデル（**lm**）を使ってみる

```
> attach(mm)                # mm をデータフレームとして再び指定する.
> fn <- lm(DATA ~ TRT)      # 完全無作為化法による分散分析
> anova(fn)                 # 分散分析表の表示
  Analysis of Variance Table
  Response: DATA
            Df  Sum Sq Mean Sq F value    Pr(>F)
  TRT        6 5587175  931196  9.8255 3.329e-05 ***
  Residuals 21 1990237   94773
```

「**anova()**」による上の分散分析表は「**aov**」での出力と完全に同一である．しかし, 分散分析の線形モデルとしての特徴をよりはっきりさせたのが,「**summary()**」による次の出力である：

```
> summary(fn)               # 分散分析表の表示.
                            # ここでは線形モデルとしての評価結果が表示されている。
  Call:
  lm(formula = DATA ~ TRT)

  Residuals:
      Min       1Q   Median       3Q      Max
  -572.00 -137.25  -19.25  200.25  688.00

  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept)  2128.00     153.93  13.825 5.13e-12 ***
  TRTCon       -812.00     217.69  -3.730  0.00124 **
  TRTDB        -332.00     217.69  -1.525  0.14214
  TRTDDT        423.75     217.69   1.947  0.06508 .
  TRTDK        -447.00     217.69  -2.053  0.05269 .
  TRTDM1         -1.25     217.69  -0.006  0.99547
  TRTDM2        550.00     217.69   2.527  0.01962 *

  Residual standard error: 307.9 on 21 degrees of freedom
  Multiple R-Squared: 0.7373,     Adjusted R-squared: 0.6623
  F-statistic: 9.826 on 6 and 21 DF,  p-value: 3.329e-05
```

線形モデルの一般形は「$\mathbf{y} = \mathbf{X}\beta + \epsilon$ 」となる．観察データ（y）に対して，推定されるべきパラメータ（β）とモデルの種類を決定するデザイン行列（X）ならびに誤差項がこの線形モデルを構成している．完全無作為法による実験計画の場合，パラメータは総平均（μ）ならびに処理効果（α）であり，上の出力は，「Intercept（切片）」が総平均を，そして他の六つのパラメータ（TRTCon ～ TRTDM2）は実験処理（TRT）の水準ごとの処理効果の推定値を与えている（デザイン行列がフルランクではないため推定できないパラメータが残る）．また「Residual standard error」とは誤差平均平方の平方根であり，「R^2」で表される「決定係数（coefficient of determination）」は「（処理平方和）／（全平方和）」で定義され，この線形モデルによって説明される変動の割合を表している．

```
> detach()
> rm(list=ls())
```

【３】一般化線形モデル（**glm**）を使ってみる

```
> attach(mm)
> fo <- glm(DATA ~ TRT, gaussian(identity))
                # 正規分布（gaussian）を指定した一般化線形モデルとしての分散分析
                # 平均値そのものを分析するので連結関数（link=）の指定は「identity」

> anova(fo)
  Analysis of Deviance Table    #「逸脱度（deviance）」という表現に注意
  Model: gaussian, link: identity
  Response: DATA
  Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev
  NULL                    27    7577412
  TRT    6  5587175        21    1990238
```

「逸脱度（deviance）」は「２×（帰無モデルの最大対数尤度）－２×（対立モデルの最大対数尤度）」と定義されるが，正規分布（gaussian）を仮定する上の例では分散分析での「平方和」と完全に一致する．上の出力では，処理効果を含まない「帰無仮説（null model）」に対して，処理効果（TRT）というパラメータを含む対立仮説とのモデル比較をしている．一方，下の出力では，対立仮説がモデルとしてどれほどよいかを「AIC（赤池情報量基準）」によって数値化している．

```
> summary(fo)

Call:
glm(formula = DATA ~ TRT, family = gaussian(identity))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-572.00  -137.25   -19.25   200.25   688.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2128.00     153.93  13.825 5.13e-12 ***
TRTCon       -812.00     217.69  -3.730  0.00124 **
TRTDB        -332.00     217.69  -1.525  0.14214
TRTDDT        423.75     217.69   1.947  0.06508 .
TRTDK        -447.00     217.69  -2.053  0.05269 .
TRTDM1         -1.25     217.69  -0.006  0.99547
TRTDM2        550.00     217.69   2.527  0.01962 *

(Dispersion parameter for gaussian family taken to be 94773.21)

    Null deviance: 7577412  on 27  degrees of freedom
Residual deviance: 1990238  on 21  degrees of freedom
AIC: 408.26

Number of Fisher Scoring iterations: 2
```

一般化線形モデルは，データに対する「モデル」のあてはめを行なう．モデルの適合（fit）の良さを評価する基準としてはいくつか提案されているが，上では「赤池情報量基準（AIC）」に基づくモデル評価がなされている．AIC は「－２×（モデルの最大対数尤度）＋２×（モデルのパラメータ数）」と定義される．パラメータの多い複雑なモデルはデータに対するフィットが高く，最大尤度はより大きくなる．しかし，AIC はモデルの複雑度に対するペナルティーを科しているので，結果として，過度に複雑なモデルは選ばれないことになる．AIC が最小となるモデルづくりがここでの目標となる．

上の例について，帰無モデルと対立モデルとの間で AIC がどれほどちがうかを調べよう：

```
> model <- fo          #glm の結果（fo）を「model」に移す.
> AIC(model)           # 対立モデルの AIC を求める.
      [1] 408.2642
> null.model <- update(model, ~.-TRT)      # 処理効果を除去し，帰無モデルをつくる
> AIC(null.model)      # 帰無モデルの AIC を求める.
      [1] 433.6979
```

確かに対立モデルの方が，帰無モデルよりも，AIC の点で優れていることがわかる．次に，二要因乱塊法による BOX3 のデータについて，**glm** での解析を行なう．

```
> mm <- read.table("Box3_R.txt", header=T)        # データ「Box3_R.txt」の読み込み
> attach(mm)
> REP <- factor(REP)
> N <- factor(N)
> V <- factor(V)
> model <- glm(DATA ~ REP + N + V + N:V, gaussian(identity)) #「glm」での分析
> anova(model)
      Analysis of Deviance Table
      Model: gaussian, link: identity
      Response: DATA
      Terms added sequentially (first to last)

           Df Deviance Resid. Df Resid. Dev
      NULL                    59     53.531
      REP   3    2.600        56     50.931
      N     4   41.235        52      9.696
      V     2    1.053        50      8.644
      N:V   8    2.291        42      6.353

> summary(model)
      Call:
      glm(formula = DATA ~ REP + N + V + N:V, family =
gaussian(identity))

      Deviance Residuals:
```

```
      Min         1Q     Median         3Q        Max
-0.878300  -0.177425   0.007083   0.242392   0.574133

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3003     0.2130  15.493  < 2e-16 ***
REP2         -0.2203     0.1420  -1.551  0.12840
REP3          0.0140     0.1420   0.099  0.92194
REP4         -0.4989     0.1420  -3.513  0.00107 **
NN1           1.5995     0.2750   5.816 7.30e-07 ***
NN2           1.3355     0.2750   4.856 1.70e-05 ***
NN3           2.5930     0.2750   9.429 6.30e-12 ***
NN4           2.6990     0.2750   9.814 1.96e-12 ***
VV2           0.4240     0.2750   1.542  0.13063
VV3           0.6540     0.2750   2.378  0.02203 *
NN1:VV2      -0.3415     0.3889  -0.878  0.38490
NN2:VV2       0.5525     0.3889   1.421  0.16281
NN3:VV2      -0.4015     0.3889  -1.032  0.30782
NN4:VV2      -0.5665     0.3889  -1.457  0.15266
NN1:VV3      -0.6240     0.3889  -1.604  0.11611
NN2:VV3       0.2065     0.3889   0.531  0.59824
NN3:VV3      -0.8185     0.3889  -2.105  0.04135 *
NN4:VV3      -0.5905     0.3889  -1.518  0.13643
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for gaussian family taken to be 0.1512571)

Null deviance: 53.5309  on 59  degrees of freedom
Residual deviance:  6.3528  on 42  degrees of freedom
AIC: 73.546

Number of Fisher Scoring iterations: 2
```

上の解析におけるモデルが AIC の点でどれくらい優れているかは，要因をひとつずつ削除することによってその効果を確かめることができる.

```
> AIC(model)          # 元のモデルの AIC
      [1] 73.54565
> AIC(update(model, ~.-N:V))        # モデルから交互作用「N:V」の除去したときの AIC
      [1] 76.02056
> AIC(update(model, ~.-N:V -V))     # モデルからさらに「V」を除去する
      [1] 78.91666
> AIC(update(model, ~.-N:V -V -N))          # さらに「N」も除去する
      [1] 170.4403
> AIC(update(model, ~.-N:V -V -N -REP))  # 最後に「REP」を除去した帰無モデルの AIC
      [1] 167.4275
```

要因を削除するたびに，AIC の値が増加し，モデルとしての良さが失われていくことに注意されたい.

参考文献

1) Michael J. Crawley (2005), *Statistics : An Introduction Using R*. John Wiley & Sons, Chichester.［Michael J. Crawley 著（野間口謙太郎・菊池泰樹訳）『統計学：R を用いた入門書』2008 年，共立出版］

2) Peter Dalgaard (2002), *Introductory Statistics with R*. Springer Verlag, New York.［ピーター・ダルガード著（岡田昌史監訳）『R による医療統計学』2007 年，丸善出版事業部］

3) Julian J. Faraway (2004), *Linear Models with R*. Chapman & Hall /CRC, Boca Raton.

4) Julian J. Faraway (2006) , *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall /CRC, Boca Raton.