

# A brief introduction to the linear model with R

凡例 はんれい legends

ver 091218 : minor updated 20161003

"=", 同義語 どうぎご synonyms: "→", 説明 せつめい explanation: "C", 含まれる ふくまれる being included in

引用文献 いんようぶんけん Abbreviation for references in this note

"Dobson" : *cited from* "Dobson, 「一般化線形モデル入門」(An introduction to generalized linear models 2nd ed.)

"Faraway" : *cited from* "Faraway, 'Extending the linear model with R', Chapman & Hall/CRC.

"Grafen" : *cited from* "Grafen & Hails, 「一般線形モデルによる生物科学のための現代統計学」(Modern statistics for the life sciences)

"Vit" : *cited from* "Vittinghoff et al., "Regression methods in biostatistics (Springer)"

"モデル選択" : *cited from* "下平 et al, 「モデル選択」(岩波書店)"

"Zar" : *cited from* "Zar, Biostatistical analyses. 5th ed. (Pearson Educational)"

"Venables" : *cited from* "Venables & Ripley 「S-PLUS による統計解析」(Modern Applied Statistics with S-PLUS, 3rd ed., Springer)

"McCarthy" : *cited from* "Bayesian Methods for Ecology (Cambridge Univ. Press)"

"Crawley" : *cited from* "Statistics: An Introduction using R"

"Johnson" Johnson & Omland (2004) Trends in Ecology & Evolution 19 (2), 101-108.

"中妻" : *cited from* "中妻, 「入門 ベイズ統計学」(朝倉書店)

用語集 ようごしゅう / 同義語 どうぎご glossary / synonyms (Dobson P2)

terms for 'X' variables :

説明変数 せつめい へんすう explanatory variable

= 予測変数 よそくへんすう

predictor/regressor variable (Zar p424)

= 独立変数 どれつ へんすう independent variable

a) 質的説明変数 しつてき せつめい へんすう

qualitative explanatory variable

= 因子 いんし factor

(例 「樹種」 e.g., "species name")

・ 因子のカテゴリ category of a factor

= 水準 すいじゆん level (例 具体的な種名

e.g., "Betula ermanii", "Quercus crispula")

b) 量的説明変数 りょうてき せつめい へんすう

quantitative explanatory variable

= 共変量 きょうへんりょう covariate

terms for 'Y' variables :

反応変数 はんのう へんすう response variable or

criterion variable (Zar p424)

= 結果変数 けつか へんすう outcome variable

= 従属変数 じゅうぞく へんすう dependent variable

X, Y の尺度 しゃくど Scales of X and Y:

1) 質的変数 for qualitative variable

カテゴリ変数 かがてりー へんすう

categorical variable

= 離散変数 りさん へんすう discrete variable

☞ 計数データ けいすう でーた count data,

☞ 度数データ とうすう でーた frequency data

1a) 名義尺度 めいぎ しゃくど nominal scale

= 名義分類 めいぎ ぶんるい nominal classification

1b) 順序尺度 じゆんじょ しゃくど ordinal scale

= 順序分類 -- ぶんるい ordinal classification

2) 量的変数 りょうてき へんすう for quantitative variable

= 連続変数 れんぞく へんすう continuous variable

## Overview of this note

ここで取り扱う GLM は基本的に正規分布型の GLM である。したがって、このための [R] の関数は `lm`、`package (nlme)` の `lme`、`package (lme4)` の `lmer` である。

Basically the GLMs in this note deal with normal type GLM. Hence the functions of [R] used in this note are `lm`, `lme` in package `(nlme)`, and `lmer` in package `(lme4)`.

間違いもたくさんあるはずなので信用しないこと。また、基本的には隅田のメモ用につくったものなので、詳しい説明はしない。

Do not trust this note as there would be a lot of mistakes. Basically descriptions are prepared for Sumida's use only, so do not complain about unkind explanations. (But please tell me if you find any serious mistakes).

0) 統計の約束などの簡単な復習 A short revision of stats before learning GLMs

1) 線形モデルの定義 Definition of linear model

- Rによる線形モデルの表記一重回帰による例 Let's try linear models using [R] with multiple regressions
- 重回帰におけるパラメタ選択 Parameter selection in ordinary multiple regression
- p値の多重性の問題: 説明変数の数 Multiplicity of p-values and the number of explanatory variables.
- 大きいモデルと小さいモデル Bigger models, smaller models

2) GLM の特徴 Characteristics of GLMs

- 連結関数 link function : なぜリンク関数について知ることが大事か? Why is link function important?
- 最大尤度と GLM との関係: 尤度的にモデル選択する理由 Why do we use likelihood method in GLM?
- なぜ AIC をモデル選択に使うのか Why do we use AIC for model selection?

`nlm()` in package `(nlme)` / `lmer()` in package `(lme4)`

3) 混合モデル Mixed models

- 固定効果、変量効果(ランダム効果) fixed effects, random effects
- 練習 Practice

付録 Appendix

- 尤度 likelihood、最尤法 maximum likelihood method、尤度比検定 likelihood ratio test
- `offset` 関数

式の表示についての約束ごと Basics for expression of equations (Dobson p41)

•列ベクトル  $x$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{bmatrix}$$

に対し、行と列を入れ替えたベクトル(=転置行列)を  $x^T$  で表す。

すなわち  $x^T$  はベクトル  $x$  の転置行列は

$$x^T = (x_1, x_2, \dots, x_m)$$

Let  $x$  a column vector :

A vector which is transposed from a column vector (=transpose matrix, i.e., row and column are exchanged),  $x$ , is expressed by  $x^T$ .

$$x^T = (x_1, x_2, \dots, x_m)$$

•線形モデル linear model の定義

せんけい もでる の ていき (Faraway p6)

1つの従属変数(=反応変数)  $Y$  と一組の説明変数  $x_1, x_2, \dots, x_{m-1}$  とを結ぶ線形モデル linear model は、

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{m-1} x_{m-1} + \epsilon$$

の形で表す  $\beta_0$  は切片 intercept、 $\epsilon$  は誤差

↑ 隅田註:  $x_1, x_2, \dots, x_{m-1}$  は、重回帰のときの個々の説明変数を想定すればよい。すなわち、 $m-1$  は説明変数の数(パラメータ数は切片をいれて  $m$  個)。

• Defining a linear model (Faraway p6)

A linear model takes the form,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{m-1} x_{m-1} + \epsilon$$

where  $\beta_i$  ( $i = 0, 1, 2, \dots, m-1$ ) are unknown parameters, and  $\beta_0$  is called the *intercept* term.

The *response* is  $Y$  and the *predictors* are  $x_1, x_2, \dots, x_{m-1}$ .  $\epsilon$  shows error term.

↑ Note: Assume  $x_1, x_2, \dots, x_{m-1}$  to be each of the explanatory variables of multiple regression. So  $m-1$  shows the number of explanatory variables, and the number of parameters including the intercept is  $m$ .

•Example) 回帰式の表現 representation of regression equation

$$y = X\beta + \epsilon$$

ただし、where

$$y = (y_1, y_2, \dots, y_n)^T,$$

$$\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T, \text{ (error)}$$

$$\beta = (\beta_0, \beta_1, \dots, \beta_{m-1})^T,$$

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,m-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,m-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,m-1} \end{pmatrix}$$

$n$ , データ数 number of data:

$m-1$ , 説明変数の数 number of predictor variables

## Rで generalized linear model による解析を行う前に思い出しておくべきこと

Something you need to review before doing analyses with generalized linear models by [R]

### 基本 Basics (DobsonP4)

• 分散分析 ぶんさん ぶんせき analysis of variance, ANOVA:

→ 説明変数がすべてカテゴリカルの時、連続的な反応変数とカテゴリカルまたは質的説明変数(因子)の関係を調べる

For the relationship between continuous response variable (s) and categorical or qualitative explanatory variables (factors) when explanatory variables are all categorical.(Crawley, Chapter 9, p167)

• 共分散分析 きょうぶんさんぶんせき analysis of covariance, ANCOVA

→ 反応変数は連続型、説明変数は連続型変数とカテゴリカル変数の両方を少なくとも1個ずつ以上含む時

Applied when response variables are continuous and explanatory variables contain both one or more continuous variables and categorical ones (Crawley, Chapter 10, p202).

• 線形重回帰 せんけい じゅうかいき multiple linear regression

→ 1つの連続的な反応変数といくつかの連続的な説明変数(あるいは予測変数)の関係を調べる(方法は App. 1) 参照)

For the relationship between one continuous response variable and several continuous explanatory (predictor) variables.

### 重回帰における標準な変数選択のひとつ、「増減法」による変数選択方法

**Ordinary method for selecting independent variables in multiple regression model (Zar p433-)**

まず、説明変数を1つだけにしたモデルのなかで、 $b_i$  ( $=x_i$  のパラメーター) のうち  $t$  値の絶対値  $|t|$  が最も大きいものを選ぶ。たとえば、 $x_{(1)}$  が選ばれたとする。次に、 $x_{(1)}$  の存在下で  $|t|$  が最も大きいものをもう一つ選ぶ(例えば  $x_{(1)}$  and  $x_{(2)}$  が選ばれたとする)。このようにして、変数を増やしていくが、次のステップで  $x_{(1)}$ ,  $x_{(2)}$ ,  $x_{(3)}$  が選ばれたとしても、 $x$  を1つ増やすごとに、それまで選ばれた変数のうち  $t$  が有意でないものがないかを確認し、1つでもあれば、もっとも  $|t|$  が小さいものを除く、という方法で変数を増やしていく。

One of the most common method is "Stepwise Regression (i.e., both step-up and stem-down procedures are included)"; Starting with the regression with only one explanatory variable, the one for which  $b_i$  (= parameter of  $x_i$ ) has the largest value of  $|t|$  is first determined, e.g.  $x_{(1)}$  is chosen. Then another dependent variable with largest  $|t|$  associated with  $x_{(1)}$  is chosen (e.g., now  $x_{(1)}$  and  $x_{(2)}$  chosen). If in the next step  $x_{(1)}$ ,  $x_{(2)}$  and  $x_{(3)}$  is chosen, but whenever an  $x$  is added, the  $b$  associated with each of  $x$ 's already chosen in the model is examined to see whether it has a non-significant  $t$ , and if any of them do, the one with the smallest  $|t|$  is eliminated at that step. (\*,  $t$ -value for  $H_0: \beta_j=0$  against  $H_0: \beta_j \neq 0$  is,  $t_{0.05(2), \nu}$  ( $\nu = n - m - 1 = (\text{number of data sets}) - (\text{number of variables}) - 1$  at that step)) A stepwise testing approach is an inferior method to variable selection compared to the criterion-based methods. Nevertheless, testing-based methods are still useful, particularly when under manual control. They allow the user to respect restrictions of hierarchy and situations where certain variables must be included for explanatory purposes. Faraway p22

——重要 p 値の多重性 Multiplicity of p-values (Grafen chap10.2 p184, chap11.2 p209) ——

各パラメータの p 値は個々に計算されるので、y 有意な説明変数がたくさんあると第 1 種の過誤を犯す確率が高くなる。(モデル選択の説明の「節減の法則」も参照。)

Probability of type I error becomes high as more parameters are included in a model.  
(See also "the principle of PARSIMONY" in the chapter of model selection)

第 1 種の過誤を犯さないためには、  
 $1 - (1 - \alpha)^k < 0.05$  になるように  $\alpha$  を設定する必要がある (k は説明変数の数)。  
ただし、p が互いに独立な場合だけ。

One easiest way to avoid type I error is to set  $\alpha$  (significance level e.g., 0.05) such that  $1 - (1 - \alpha)^k < 0.05$ , where k is the number of explanatory variables (but under an assumption that all p-values are independent, which is not the case.)

とにかく簡単な線形モデルを R でやってみよう Let's try a simple linear model using R

# working directoryを変えておいて

```
mydata <- read.csv("Peru.csv")
#Peru.csvというcsvファイルのデータを読み込んで"mydata"というオブジェクト名をつける
head(mydata) # mydataというデータの頭6行を表示させる
  AGE YEARS WEIGHT HEIGHT CHIN FOREARM CALF PULSE SYSTOL DIASTOL
1  21     1   71.0  1629  8.0    7.0 12.7   88    170    76
2  22     6   56.5  1569  3.3    5.0  8.0   64    120    60
3  24     5   56.0  1561  3.3    1.3  4.3   68    125    75
4  24     1   61.0  1619  3.7    3.0  4.3   52    148   120
5  25     1   65.0  1566  9.0   12.7 20.7   72    140    78
6  27    19   62.0  1639  3.0    3.3  5.7   72    106    72
```

#### 線形モデルをRでやってみる

```
# ↓ lm() はliner modelをやらせるRの関数名 lm() is an [R] function for linear model
# ↓ lm()の中に y = b0 + b1 x1 + b2 x2 + ...に相当する式を指定する。
# ↓ 上の式の場合、「y ~ x1 + x2」と書くだけ
# ↓ perulm <- は、その計算結果をperulm という名前にしたオブジェクトにいれよ、という意味
# ↓ tの値は、その説明変数を落としたときにどれだけ当てはまりが悪くなるかを示す (モデル選択p10)
# t-values represents how the fitness become worse if the predictor is removed.
```

```
2変数、交互作用なし no interaction term
perulm <- lm(mydata$SYSTOL ~ mydata$WEIGHT + mydata$YEARS)
summary(perulm) # perulmと名前をつけた計算結果を表示させる命令
Call:
lm(formula = mydata$SYSTOL ~ mydata$WEIGHT + mydata$YEARS)
```

Residuals:  
Min 1Q Median 3Q Max  
-17.469 -7.878 1.076 6.292 24.113

Coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 50.3191 15.8184 3.181 0.00302 \*\*  
mydata\$WEIGHT 1.3541 0.2672 5.067 1.22e-05 \*\*\*  
mydata\$YEARS -0.5718 0.1879 -3.043 0.00436 \*\*

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.25 on 36 degrees of freedom  
Multiple R-squared: 0.4208, Adjusted R-squared: 0.3886  
F-statistic: 13.08 on 2 and 36 DF, p-value: 5.385e-05

```
# 出力結果の意味は、
# mydata$SYSTOL = 1.3541×mydata$WEIGHT -0.5718×mydata$YEARS + 50.3191
# R2 = 0.3886, p=5.385e-05, ...
```

#### 説明変数の順番の入れ替え → 単に表示の順番が変わるだけ

```
perulm2 <- lm(mydata$SYSTOL ~ mydata$YEARS + mydata$WEIGHT)
summary(perulm2)
Call:
lm(formula = mydata$SYSTOL ~ mydata$YEARS + mydata$WEIGHT)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-17.469  -7.878   1.076   6.292  24.113
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.3191    15.8184   3.181  0.00302 **
mydata$YEARS  -0.5718     0.1879  -3.043  0.00436 **
mydata$WEIGHT  1.3541     0.2672   5.067  1.22e-05 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 10.25 on 36 degrees of freedom
Multiple R-squared:  0.4208,    Adjusted R-squared:  0.3886
F-statistic: 13.08 on 2 and 36 DF,  p-value: 5.385e-05
```

# 変数を3つに増やしてみる (交互作用なし no interaction)

```
peru3 <- lm(SYSTOL ~ YEARS + WEIGHT + AGE, mydata)
summary(peru3)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.8212    16.7196   3.159  0.00325 **
YEARS        -0.5187     0.2166  -2.394  0.02214 *
WEIGHT        1.3836     0.2762   5.010  1.56e-05 ***
AGE          -0.1410     0.2764  -0.510  0.61328
```

```
---中略
Residual standard error: 10.36 on 35 degrees of freedom
Multiple R-squared:  0.425,    Adjusted R-squared:  0.3758
F-statistic: 8.625 on 3 and 35 DF,  p-value: 0.0002023
```

次の二つは同じ

```
perulm3 <- lm(mydata$SYSTOL ~ mydata$YEARS + mydata$WEIGHT + mydata$HEIGHT)
perulm3 <- lm(SYSTOL ~ YEARS + WEIGHT + WEIGHT, mydata)
```

交互作用を含むモデルの指定 "\*" for all terms

```
perukougo <- lm(mydata$SYSTOL ~ mydata$YEARS * mydata$WEIGHT * mydata$PULSE); summary(perukougo)
Call:
lm(formula = mydata$SYSTOL ~ mydata$YEARS * mydata$WEIGHT * mydata$PULSE)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-17.224  -6.952   0.689   6.695  16.788
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  752.776561    284.212690   2.649  0.0126 *
mydata$YEARS -17.214350    12.716132  -1.354  0.1856
mydata$WEIGHT -9.765249     4.389417  -2.225  0.0335 *
mydata$PULSE  -9.953690     4.083273  -2.438  0.0207 *
mydata$YEARS:mydata$WEIGHT  0.263592     0.184640   1.428  0.1634
mydata$YEARS:mydata$PULSE  0.240460     0.170458   1.411  0.1683
mydata$WEIGHT:mydata$PULSE  0.156370     0.062203   2.514  0.0173 *
mydata$YEARS:mydata$WEIGHT:mydata$PULSE -0.003752     0.002424  -1.548  0.1317
```

```
---中略
Residual standard error: 9.729 on 31 degrees of freedom
Multiple R-squared:  0.5507,    Adjusted R-squared:  0.4493
F-statistic: 5.428 on 7 and 31 DF,  p-value: 0.0003891
```

交互作用項だけの指定 ":" を使う ":" for interaction term only

```
perukougo2 <- lm(mydata$SYSTOL ~ mydata$YEARS : mydata$WEIGHT + mydata$PULSE); summary(perukougo2)
Call:
lm(formula = mydata$SYSTOL ~ mydata$YEARS:mydata$WEIGHT + mydata$PULSE)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-21.723  -8.732  -1.992   8.017  39.255
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```

(Intercept)          1.143e+02  1.624e+01  7.037 2.94e-08 ***
mydata$PULSE         1.874e-01  2.362e-01  0.794  0.433
mydata$YEARS:mydata$WEIGHT -2.103e-05  2.955e-03  -0.007  0.994
---中略
Residual standard error: 13.35 on 36 degrees of freedom
Multiple R-squared: 0.01836, Adjusted R-squared: -0.03618
F-statistic: 0.3366 on 2 and 36 DF, p-value: 0.7164

```

**切片のないモデル "−1" をつけてやる "−1" for a no-intercept model**

```

peru00 <- lm(mydata$SYSTOL ~ mydata$WEIGHT * mydata$YEARS - 1) ; summary(peru00)
Call:
lm(formula = mydata$SYSTOL ~ mydata$WEIGHT * mydata$YEARS - 1)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-19.176  -8.993   1.142   8.122  17.456

```

```

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
mydata$WEIGHT      2.15909    0.05459  39.548 <2e-16 ***
mydata$YEARS       0.90709    0.80201   1.131  0.2655
mydata$WEIGHT:mydata$YEARS -0.02336    0.01083  -2.156  0.0378 *
---中略
Residual standard error: 10.92 on 36 degrees of freedom
Multiple R-squared: 0.9933, Adjusted R-squared: 0.9927
F-statistic: 1776 on 3 and 36 DF, p-value: < 2.2e-16

```

**full model (but with no interactions)**

```

perufull <- lm(SYSTOL ~ AGE + YEARS + WEIGHT + HEIGHT + CHIN + FOREARM + CALF + PULSE + DIASTOL,
mydata) ; summary(perufull)
Call:
lm(formula = SYSTOL ~ AGE + YEARS + WEIGHT + HEIGHT + CHIN +
    FOREARM + CALF + PULSE + DIASTOL, data = mydata)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-18.3956  -6.6134  -0.0567   6.6590  23.3068

```

```

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.65501    57.60797   2.042 0.050304 .
AGE          -0.21770    0.28583  -0.762 0.452425 .
YEARS       -0.56277    0.22116  -2.545 0.016523 *
WEIGHT       1.84852    0.48825   3.786 0.000713 ***
HEIGHT      -0.06586    0.04218  -1.561 0.129300
CHIN        -1.00756    0.88763  -1.135 0.265623
FOREARM     -0.86086    1.40854  -0.611 0.545846
CALF        -0.01008    0.63980  -0.016 0.987543
PULSE       0.05049    0.19916   0.254 0.801645
DIASTOL     0.26295    0.16539   1.590 0.122691
---中略
Residual standard error: 10.18 on 29 degrees of freedom
Multiple R-squared: 0.5399, Adjusted R-squared: 0.3972
F-statistic: 3.782 on 9 and 29 DF, p-value: 0.002991

```

---

## [R] の出力に関する補足 Note

### 線形回帰関数 lm の出力の簡単な説明:

Outputs of "lm" in R (Faraway p8-9)

**deviance**: A more general term of RSS (residual sum of squares); for linear model, deviance is the RSS.

:RSS (誤差の二乗和)みたいなものだが、より一般的な用語。ただし、線形モデルでは deviance =

RSS (Faraway p8)

**degrees of freedom**: for a linear model, it is the number of cases minus the number of coefficients.

**Adjusted R<sup>2</sup>**:  $= 1 - \text{RSS} / (n-p) / ((\text{TSS} / (n-1)))$

TSS: total sum of squares

R<sup>2</sup> can never decrease when a new predictor is added to a model. This means that it will favor the largest models. Adjusted R<sup>2</sup> makes allowance for the fact a larger model also uses more parameters.

---

### 関数 anova () による、パラメータ数の違う2つのモデル $\Omega$ 、 $\omega$ ( $\omega$ は $\Omega$ の部分モデル) の比較

(この関数分散分析をするという意味ではない; 分散分析は aov () を使う)

Comparison of two models with different dimensions (or the number of parameters), where **the parameters of  $\omega$  are a subset of the predictors of  $\Omega$** , using "anova ()" (Faraway p12)

anova ("model", "model")

仮定 assumption

p ←  $\Omega$  の大きさ (パラメータ数)

the # of parameters of  $\Omega$

q ←  $\omega$  の大きさ (パラメータ数)

the # of parameters of  $\omega$

∴ p > q

仮説 null hypothesis: 小さいモデル  $\omega$  のほうが正しい

smaller model  $\omega$  is correct

$F = (\text{RSS } \omega - \text{RSS } \Omega) / (p-q) / ((\text{RSS } \Omega / (n-p)))$

と  $F(p-q, n-p)$  とを比較し、

if  $F > F(p-q, n-p)$  または  $\text{Pr}(>F) < 0.05$

ならば null を棄却 (小さいモデル  $\omega$  のほうが正しいとは言えない)

then we would reject the null hypothesis

Example :

lmod <- lm (小さいモデルの式 small model);

lmodi <- lm (大きいモデルの式 large model)

を実行した後に after doing above,

anova (lmod, lmodi)

により F, Pr(>F) を調べる。p > q なので、大きいモデルが後に来る?

drop1 (lmodi, test = "F") Faraway p9

最も大きいモデルから1つパラメータを取ったモデルを比較する。?

confint (lm のモデル名)

モデルパラメータの95%信頼範囲を出力。ただし、同時信頼範囲ではない Faraway p14

plot (lm のモデル名)

結果を4分割で表示



## Generalized linear model の特徴    Characteristics of GLM

### 連結関数 れんけつ かんすう Link function

リンク関数とは何か？ What's link function? (after Dobson p36, p52-53, Faraway p116)

N 個の応答変数  $Y_i$  ( $i=1,2,\dots,N$ ) に対し、 $i$  番目の変数  $Y_i$  の期待値(平均値)を  $E(Y_i)=\mu_i$  と表記する。また、行列  $X$  の  $i$  列目のベクトルの転置行列を  $x_i^T$  で表す。

For  $N$  number of responses  $Y_i$  ( $i=1,2,\dots,N$ ) expressing the mean of  $i$ -th response of  $Y_i$  is expressed as  $E(Y_i)=\mu_i$ .

The linear model (linear linkage) of  $X$  is expressed as follows:

• 行列  $X$  全体の線形モデルは次のように表される。  
 $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = X\beta$   
 $X\beta$  を  $g$  の線形成分 せんけいせいぶん (linear component) または線形予測子 せんけいよそくし (linear predictor) という。

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = X\beta$   
 which is called "linear predictor" or "linear component" (of  $g$  defined below).

Here let the function linking the mean  $\mu_i$  with linear component is expressed  $g$ . That is,

• 平均値  $\mu_i$  と線形予測子とを結びつける関数を  $g$  とすると、

$$g[\mu_i] = g[E(Y_i)] = x_i^T \beta = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_m x_{i,m}$$

$$g[\mu_i] = g[E(Y_i)] = x_i^T \beta = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_m x_{i,m}$$

Here "g" is called the "link function", i.e., link function is an equation that shows the relationship between the mean value of  $Y$  and the linear component of explanatory variable  $X$ .

この  $g$  のことを 連結関数 という。

連結関数 れんけつかんすう link function

(=リンク関数 れんくかんすう)

→  $Y$  の期待値(平均値)  $\mu$  と説明変数  $X$  の線形結合とを連結する関係式(関数)。(Dobson p36)

→ **すなわち、リンク関数  $g$  は、covariates (共変量 =  $x$  のこと) が線形予測子によってどのように応答の平均  $E(Y)=\mu$  に結びつけられるかを表す。**

→ The link function  $g$  describes how the mean response,  $E(Y)=\mu$ , is linked to the covariates (=  $x$ , or explanatory variables) through the linear predictor  $\eta$ .

• 通常データだけではどのリンク関数を使うかを決められない (Faraway p36) (but 表参照)

• It is usually not possible to choose a link function to be used based on the data alone. (Faraway p36); (but see Table)

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{1,1}, x_{1,2}, \dots, x_{1,m} \\ x_{2,1}, x_{2,2}, \dots, x_{2,m} \\ \vdots \\ x_{N,1}, x_{N,2}, \dots, x_{N,m} \end{pmatrix} \quad y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} \quad g[E(y)] = \begin{pmatrix} g[E(Y_1)] \\ g[E(Y_2)] \\ \vdots \\ g[E(Y_N)] \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}$$

簡単な例として、変数が 1 個だけ ( $i=1$ ) の場合を考えよう。たとえば体重  $Y$  を推定する説明変数を身長だけを変数  $x$  として使用する例である。このとき、線形予測子は

$$\beta_0 + \beta_1 x$$

である。 $Y$  の平均値と  $x$  の線形モデルとの関係は、連結関数  $g$  を使うと

$$g[E(Y)] = \beta_0 + \beta_1 x$$

と書ける。

As an simplest example, assume a case in which there is only a single variable (i.e.  $i=1$ ).

This is a case when the explanatory variable of body weight  $Y$  is only the body height  $X$ .

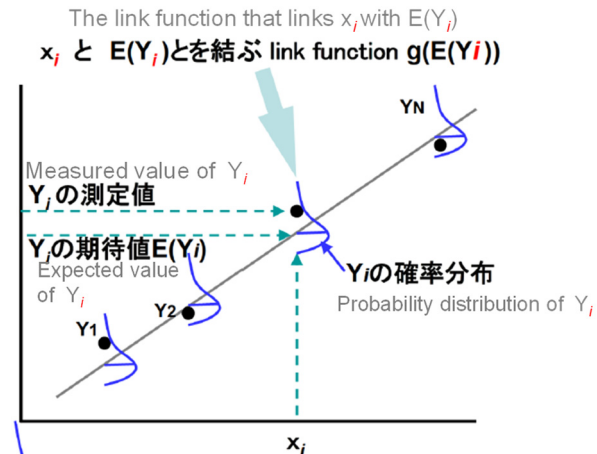
In this case the predictor is

$$\beta_0 + \beta_1 x .$$

The relationship between the mean of  $Y$  and the linear component of explanatory variable  $X$  is written, using the link function, as follows.

$$g[E(Y)] = \beta_0 + \beta_1 x$$

**連結関数 link function のイメージはこんなふう( ? ) Image of 'link function' is as follows ( ? )**



Y<sub>i</sub>の確率分布は、左図のようなひずんだ形かもしれない。左の場合、期待値の下側よりも期待値の上側のほうがデータのばらつきの範囲が大きい。

Distribution of the probability of Y<sub>i</sub> may be skewed like this; in this case the range of data > the mean is wider than the data range < mean.

**重要でない補足:リンク関数を行列表記すると...**

• mをパラメータ数、N を観測数とする。N 個の従属変数( = 反応変数 ) Y<sub>1</sub>, Y<sub>2</sub>, ..., Y<sub>N</sub> と、N 組の説明変数 x<sub>i</sub><sup>T</sup> とを結ぶ連結関数は行列表記で下のように表す。

**Trivial supplement**

• Let m the number of parameters, and N the number of observations. The link function linking N number of response variables Y<sub>1</sub>, Y<sub>2</sub>, ... Y<sub>N</sub> and N sets of explanatory variables x<sub>i</sub><sup>T</sup>, is expressed by

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{1,1}, x_{1,2}, \dots, x_{1,m} \\ x_{2,1}, x_{2,2}, \dots, x_{2,m} \\ \vdots \\ x_{N,1}, x_{N,2}, \dots, x_{N,m} \end{pmatrix}$$

$$y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}$$

$$g[E(y)] = \begin{pmatrix} g[E(Y_1)] \\ g[E(Y_2)] \\ \vdots \\ g[E(Y_N)] \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}$$

$x\beta = g[E(y)],$

where

y, the response vector

g[E(y)], the vector function of the mean value of Y<sub>i</sub>, or E(Y<sub>i</sub>) (g is common to all components)

x<sub>i</sub><sup>T</sup>, the transposed vector of vector x<sub>i</sub>.

The components of x are the measurements themselves if they are quantitative explanatory variable, and if they are qualitative variables they are constants allotted to show their levels (e.g., 0, 1, 2, etc = dummy variable, and is indicator variable if "0 or 1"). Matrix X is also called "design matrix".

β shows the vector of parameters of m number of explanatory variables.

ただし、

y : 従属変数( = 反応変数 )のベクトル、

g[E(y)] : Y<sub>i</sub> の平均値 E(Y<sub>i</sub>) のベクトル関数

(gはどの要素でも同じ)、

x<sub>i</sub><sup>T</sup> : 列ベクトルx<sub>i</sub>の転置ベクトル、

xの要素は、要素が量的な説明変数の場合は測定値そのもの、xの要素が質的な説明変数の場合は水準(level)を表す定数(0, 1, 2 など = ダミー変数 *だみーへんすう* **dummy variable** Dobson p42 ; 0,1 のみ)のときは **indicator variable** という)

また、行列 X はデザイン行列(design matrix)とも呼ばれる。

β → m個の説明変数のパラメーターのベクトル

## なぜ GLM でリンク関数が重要か？ Why is "Link Function" important in GLM?

例 ) 一般線形モデル general linear model (generalized じゃない) は次の事を仮定:

- 反応変数  $Y$  は正規分布に従う
- 反応変数  $Y$  は説明変数  $X$  の線形式で表される
- すべての反応は共通の分散  $\sigma$  を持つ

この場合(一般線形モデル general linear model の場合 = gaussian) のリンク関数は、

$$g[\mu] = \mu$$

すなわち、平均値(を求める計算関数)そのものがリンク関数である。これに対し、反応変数  $Y$  によっては、そうでない場合がある。(Faraway p117, Venables p256)

たとえば、 $\eta = \beta_0 + \beta_1 x_1 \dots + \beta_m x_m$  は負になる場合もありうるが、計数データ(カウントデータ)のモデルの場合平均値  $\mu$  は正の値でなければならないので、 $\eta = \mu$  とおくことができない。そこで、リンク関数として、 $\mu = \exp(\eta)$  とおいてやれば、計数データでも対応可能になる。すなわち、計数データモデルの場合は平均値が常に正のデータをとるリンク関数としてポアソン分布を用いる。(Faraway p117)

General linear model では  $\eta = \mu$  の場合しか取り扱いえないが、上の例のように、Generalized linear model ではそれ以外の場合にも取り扱いが可能である。

**したがって、[R] の GLM 関数(いくつかある)を使う時、どの family を使うのか(あるいはリンク関数は何なのか)を指定する必要がある。**

For example, in general  $\eta = \beta_0 + \beta_1 x_1 \dots + \beta_m$  (Faraway p116)

モデルパラメータ  $\beta$  をどうやって決めるか

通常の linear model で誤差が正規分布の場合(後述の表の分散関数 = 1 の場合)は最小二乗法による推定法は問題ないが、generalized linear model で正規分布以外の場合(分散関数  $\neq 1$  の場合)を取り扱う場合は最小二乗法による推定法は適用できず、尤度的推定方法が必要になる。Faraway p7

General (Not "generalized") linear model assumes that: (Dobson p133)

- Response variable follows normal distribution
- Response variable is expressed by a linear equation of predictors.
- All responses should have the same variance.

In this (i.e., general linear model or Gaussian family) case, the link function is:

i.e. (the function to obtain) mean value  $\mu$  is identical with the link function. However, this is not always the case. (Faraway p117, Venables p256)

$x_m$  can be negative, but in the case of count data model the mean  $\mu$  should be positive. So we set  $\mu = \exp(\eta)$  as the link function so that  $\eta = \log(\mu)$  which ensures  $\mu > 0$  in the case of count data models. (Faraway p117)

Though general linear model can deal with the case only when  $\eta = \mu$ , but generalized linear model can deal with other cases.

**Hence you need to determine which "family" is applied (or link function to be used) to your GLM function of [R].**

How is model parameter  $\beta$  determined?

In ordinary linear model with normal error distribution (variance function = 1 in the table), least square method is applicable. In other cases (when variance function  $\neq 1$ ), least square methods are not appropriate and likelihood-based method is necessary.

Canonical Link リンク関数: → $\eta$ と $Y$ の平均値との関係 <i>Relation between <math>\eta</math> &amp; <math>\mu</math></i>	分散関数 Variance function * 平均値が変わると分散がどう変わるかを表す describes how the variance relates to the mean.	誤差族(用語 Crawley p125) モデル族(用語 Venables p257) <b>Family</b>
$\eta = \mu$ (恒等関数)	1	Normal 正規分布 = Gaussian ガウシアン
$\eta = \log(\mu)$ (対数関数)	$\mu$	Poisson ポアソン分布 count data etc カウントデータなど
$\eta = \log(\mu / (1 - \mu))$ (ロジット関数)	$\mu(1 - \mu)$	Binomial 二項分布 death/survival etc. 生死データ等
$\eta = \mu^{-1}$ (逆数関数)	$\mu^2$	Gamma ガンマ分布
$\eta = \mu^{-2}$	$\mu^3$	Inverse Gaussian 逆ガウシアン分布

リンク関数についてのその他の用語と解説 Other notes on link function (Faraway p115/p117)

GLM は反応変数  $Y$  の分布が Exponential family と呼ばれる分布をもつものに対して定義される (Dobson p58)。たとえば、 $\theta =$  平均値、 $\Phi =$  分散の正規分布の場合のように、 $\theta$  は canonical parameter (正準パラメータ (=標準とするパラメータ) =natural parameter) と呼ばれる、確率分布の「位置」を示すパラメータで、 $\Phi$  は dispersion parameter と呼ばれる、確率分布の「スケール scale, 尺度」を表すパラメータである。 $y$  の確率分布は  $f(y | \theta, \Phi)$  で示す。

GLM の

$$\eta (= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m) = g(\mu) = \theta$$

のように、リンク関数  $g$  で表される、 $\eta = \theta$  を満たす関係を自然な連結/正準連結 canonical link (=標準とするパラメータとの連結?) と呼ぶ。

(Faraway p115、Venables p257)

GLM is defined in terms of the distribution of the response variable  $Y$  that belongs to a member of the "exponential family distribution" (Dobson p58). As in the case of normal distribution with  $\theta =$  mean and  $\Phi =$  variance,  $\theta$  is called the "canonical parameter" and represents location, while  $\Phi$  is called the "dispersion parameter" and represents the 'scale'.  $y$  is represented by a form  $f(y | \theta, \Phi)$

As in the relationship of GLM:

$$\eta (= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m) = g(\mu) = \theta$$

the relationship represented by the link function  $g$  that satisfies  $\eta = \theta$  is called the "canonical link". (Faraway p115、Venables p257)

---

## GLM と最大尤度との深い関わり：モデル選択：詳しくは App.A 参照

### Close relationship between GLM and maximum likelihood : Model selection (see also App A)

Johnson p101, Box.3

モデル選択は尤度理論に基礎をおいている。

モデル選択には通常3つの方法がある。

#### 1)適合度を最大にする

$R^2$  などが最も大きいものを選ぶ。ただし、principle of parsimony「節減の法則」すなわち簡単なモデルのほうが良いという原理に反し、パラメタ数が多いモデル(大きいモデル)のほうが  $R^2$  が高くなる結果、意味のないパラメタがはいったモデルが選択される可能性がある。

#### 2)帰無仮説(null hypothesis) 検定

尤度比検定(likelihood ratio test, LRT)は最もよく使われる「帰無仮説」的方法である。LRT は入れ子関係にある大小二つのペアでモデルを比較する。大きいモデルとその部分モデルとの尤度の比を調べ、モデルが大きくなってもモデルを複雑にする意味があるかどうかを検定する。これは、重回帰において簡単なモデルからパラメターを増やしていく、「前進法」に似ている。ただし、独立でない「複数」の検定をやることになるのでタイプ I エラー(第 1 種の誤り:帰無仮説 (null hypothesis) が正しいのに、これを棄却する誤り)をおかす可能性が高くなる。

#### 3)モデル選択基準

AIC などのように、モデルの適合性と複雑性の両方を考慮し、複数のモデルを「同時」に比較できるようにしたもの。

**GLM のパラメター  $\beta$  は最大尤度で求めることができる。通常 Gaussian GLM のときだけパラメタを解析的に求められるが、それ以外では一般的に解析的に求めることができないので数値的に求める。(Fitting a GLM Faraway p117)**

Model selection is grounded in likelihood theory. Typically one of three kinds of statistical approach is used to compare models: (see Johnson p101, Box.3)

#### 1) maximizing fit

Maximizing fit (e.g.,  $R^2$ ), with no consideration of model complexity, always favors fuller (i.e. more parameter rich) models. However, it neglects the principle of PARSIMONY and, consequently, making it a poor technique for model selection.

#### 2) null hypothesis tests

The likelihood ratio test (LRT) is the most commonly used null hypothesis approach. LRT compare pairs of nested models. When the likelihood the larger (i.e., the more complex) model is significantly greater than that of the smaller (i.e., simpler) model, the complex model is chosen, and vice versa. Selection of the more complex model indicates that the benefit of improved model fit outweighs the cost of added model complexity. LRTs are often analogous to forward selection in multiple regression, where the analyst starts with the simplest model and adds terms. A drawback is that it requires several non-independent tests, thus inflating type I error.

#### 3) model selection criteria.

Model selection criteria (e.g., AIC) consider both fit and complexity, and enable multiple models to be compared simultaneously. An important advantage is that they can be used to make inferences from more than one model, something that cannot be done using the fit maximization or null hypothesis approaches.

**The parameters  $\beta$  of a GLM can be estimated using maximum likelihood. (Faraway p117) The parameters can be analytically estimated in the Gaussian GLM, but in general it is not possible and so parameter estimation is made numerically.)**

## なぜ AIC を GLM のモデル選択に使うのか？ 最尤法/尤度比検定の弱点 (モデル選択 p24)

### *Why AIC in model selection in GLM?: Drawbacks of maximum likelihood and likelihood ratio tests*

どちらも対数尤度差に着目してモデル選択を行っている。しかし、

Both methods carry out model selection by focusing on the difference of log-likelihoods. However,

最尤法・・・モデルのパラメータ数 ( $\dim\theta$ ) が増加するにつれて対数尤度が大きくなる、という影響を考慮できない

Maximum likelihood ... We cannot take into account the effect that log-likelihood becomes greater as the number of model parameters ( $\dim\theta$ ) increases

対数尤度差・・・パラメータ数の差はカイ二乗分布の自由度として考慮されるが、包含関係にあるモデルの比較しかできない

Log-likelihood-ratio ... Difference in the number of parameters is taken into account as the degree of freedom of  $\chi^2$ -distribution, but comparison of models is possible only between models that one includes the other.

そこで、対数尤度の補正をしてパラメータ数の影響を調整したものが赤池情報量基準 AIC である。

$$AIC_k = -2 (\ell_k (\theta_k | X) - \dim\theta_k)$$

を  $k$  ごとに計算し、最も AIC が小さいモデルを選ぶ。これは、パラメータ数が増えることに対するペナルティーを、 $\dim\theta_k$  を引くことで与えている。

AIC (Akaike's information criterion) corrects log-likelihood estimates by taking into account the number of parameters. That is, by calculating  $AIC_k = -2 (\ell_k (\theta_k | X) - \dim\theta_k)$  for each  $k$ , and we choose the model with the smallest AIC. Put another way, it gives "penalty" against the increase of the number of parameters by subtracting  $\dim\theta_k$ .

---

## [R] の linear model で AIC でモデル選択する現実的理由

[R] の library lme4 の lmer から p 値が消えた :

Why has p-values disappeared from a linear mixed model function lmer (package lme4) of [R]

そのいきさつを知りたいければ以下を参照。 See below if you want to know the reason

<https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html>

一部の混合モデル (後述) 対応のパッケージ nlme の線形モデル関数 lme は p 値を出力してくれるが、Gaussian しか扱えないため、Gaussian 以外の glm の場合は lmer は使えず、パッケージ lme4 の関数 lmer を使うことになる。しかし、lmer では p 値が出ない。その他の両者の主な違いは赤字のところ参照。

したがって、lmer で glm のモデルを決める場合は、AIC によるモデル選択を行うことになる。

ただし、モデルによっては mcmcscamp 関数などでパラメタの信用区間 (信頼区間みたいなもの) を計算することは可能 ; パラメタの信頼区間に相当するものをほしければ lme4 の関数 mcmcscamp でパラメタの信用区間 credible interval (highest posterior density (HPD) interval ともいう;次ページ参照) を Bayes 推定する。

このほか、package (languageR) の pvals.fnc () を使う方法もある (例 :

```
mymodel <- lmer (y ~ aaa + (1| bbb), mydata)
```

```
mymcmc <- pvals.fnc ( mymodel, nsim =1000)
```

ただし、crossed random factors には対応してないので、

```
mymodel <- lmer (y ~ x + (1| aaa) + (1| bbb),  
mydata) のような場合には pvals.fnc は使えない
```

Function lme in package (nlme) is for general linear mixed model to fit and compare Gaussian linear and nonlinear mixed-effects models, which outputs p-values. However, p-values disappeared from an advanced version lmer in package (lme4). Hence, for generalized linear models other than Gaussian, we need to use lmer in package (lme4), but p-values are not available. See below for other differences

If you want to determine a model for glm in lmer, model selection method by AIC is applicable.

Incidentally, for some mixed models, we can estimate credible interval (=something like confidence interval; also called highest posterior density (HPD) interval ; see next page) of parameters with a Bayesian way using a function such as "mcmcscamp" in lme4. Function pvals.fnc () in package (languageR) is also available, e.g., by

```
mymodel <- lmer (y ~ aaa + (1| bbb), mydata)
```

```
mymcmc <- pvals.fnc ( mymodel, nsim =1000)
```

However, pvals.fnc () cannot be applied to crossed random factors, so it is not applied in a case like

```
mymodel <- lmer (y ~ x + (1| aaa) + (1| bbb),  
mydata)
```

====



## nlme と lme4 の比較 Comparison between nlme and lme4

### lme4

- does mcmc for the posterior distribution of parameters in Gaussian models
- handles glm's, crossed random factors, very large data sets

### nlme:

- implements mixed effects models for continuous data with Gaussian errors with nested random effects
- has a good predict method
- does NOT have mcmc: but has an approximate version of confidence intervals for parameters.
- does NOT handle glm's, crossed random factors, and very large data sets.

### HPD Highest Posterior Density Regions

The Bayesian "confidence interval" is called a highest posterior density (HPD) region or credible set. For one parameter the HPD region is sometimes called a credible interval (CI).

<http://math.bu.edu/people/dlgold/courses/HPD.pdf>

For the time being, I would recommend using a Markov Chain Monte Carlo sample (function `mcmc`) to evaluate the properties of individual coefficients (use `HPDinterval` or just `summary` from the "coda" package). Evaluating entire terms is more difficult but you can always calculate the F ratio and put a lower bound on the denominator degrees of freedom.

## 混合モデル Mixed models

### fixed effect / random effect / mixed model について (Grafen P224)

- **固定効果** こてい こうか **fixed effect**  
= 母数効果 ぼすう こうか  
→ その要因の取る水準それ自身に関心があるとき  
の変数 例: 水準の平均値
  - **Fixed effect** is an unknown constant that we try to estimate from the data. e.g. mean value of a level
- **変量効果** へんりょう こうか **random effect**  
→ 興味の対象でないが興味の対象に影響する変量(分散など)。ランダム効果を推定しても意味はないが、ランダム効果の「分布」は推定する。その水準が、ある「大きな母集団」からの標本と見なせるような **カテゴリカル変数** (Grafen P223)
  - **Random effects** are **categorical variables** whose levels are viewed as a sample from some alrege population (Grafen p 223). **It does not make sense to estimate a random effect; instead, we try to estimate the parameters that describe the distribution of this random effect.** (Faraway p153) e.g., variance, which is a parameter that affects the distribution of the parameter we want to know.
- **混合(効果)モデル** こんごう(こうか) もでる  
mixed (effect) model
  - fixed effect と random effect の両方入ったモデル
  - GLM の混合モデル版  
= generalized linear mixed model GLMM
    - **Mixed (effect) model**  
→ A model containing both fixed- and random-effect models  
→ mixed-effect version of GLM  
= generalized linear mixed model GLMM

---

### 反応変数が正規データである場合の混合モデルの一般形

General expression of mixed model when response is normally distributed (Faraway p155)

$$y = X \beta + Z \gamma + \epsilon$$

$y$ : 反応変数 Response variable

$\beta$  固定効果: 長さ  $p$  のベクトル Fixed effect: vector of length  $p$

$X$   $n \times p$  の行列  $n \times p$  model matrix

$\gamma$   $q$  個のランダム効果を持つベクトル Vector with  $q$  random effects

$Z$   $n \times q$  の行列  $n \times q$  matrix

$\epsilon$  正規分布に従う誤差 Normal errors

## Practice #Chapter 8 (of Faraway's Book) : Random effects

library (faraway)

注) MCMCによるp値の信頼範囲の計算は、Confidence intervals of p-values by MCMC was obtained by package"languageR" の" pvals.fnc (lmerのオブジェクト名, nsim=1000) で求めた。

### p156 8.1 data (pulp) の内容: random effects

紙の明るさとオペレータのデータ

```
> head(pulp)
  bright operator
1   59.8         a
2   60.0         a
3   60.8         a
4   60.8         a
5   59.8         a
6   59.8         b
```

#### モデル化 (O) : ANOVA

最もふつうに思いつくのは4人のオペレータを固定効果とするANOVA

```
lmod <- aov(bright ~ operator, pulp)
```

注) aovはlmをANOVA風に出力するためのwrapper関数 (p157)

これによる出力: 事前に

```
op <- options(contrasts = c("contr.sum", "contr.poly")) # uses 'sum to zero contrasts'
summary(lmod)
```

```
          Df Sum Sq Mean Sq F value Pr(>F)
operator   3  1.34000  0.44667   4.2039 0.02261 *
Residuals 16  1.70000  0.10625
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coef(lmod)
```

```
(Intercept) operator1 operator2 operator3
      60.40      -0.16      -0.34       0.22
```

オペレーターは4人だが、固定効果の和は0となるはずなので、4人目の値は表示されていない。

しかし、 $0 - (-0.16 + 0.34 + 0.22) = -0.40$  で計算できる。

ちなみに lm で出力させると・・・

```
lmod <- lm(bright ~ operator, pulp)
```

```
summary(lmod)
```

```
Call:
```

```
lm(formula = bright ~ operator, data = pulp)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.440 -0.195 -0.070  0.175  0.560
```

```
Coefficients:
```

```
          Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.40000    0.07289  828.681 <2e-16 ***
operator1    -0.16000    0.12624   -1.267   0.223
operator2    -0.34000    0.12624   -2.693   0.016 *
operator3     0.22000    0.12624    1.743   0.101
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.326 on 16 degrees of freedom
```

```
Multiple R-squared:  0.4408,    Adjusted R-squared:  0.3359
```

```
F-statistic: 4.204 on 3 and 16 DF,  p-value: 0.02261
```

## モデル化(1) :

```
mmod <- lmer(bright ~ 1 + (1| operator), pulp)
```

random effect の表現法

(1|operator) 意味 meaning

→ データはoperatorによってグループ化されている Data is grouped or nested by "operator".

(1| )の"1"の 意味 meaning → ランダム効果は各グループ内では一定

(1| )、 random effect is constant within each group

REML=FALSE が指定されていないので、Restricted maximum likelihood (p156) でやっている

これによる出力 :

```
summary(mmod)
Linear mixed model fit by REML
Formula: bright ~ 1 + (1 | operator)
Data: pulp
   AIC   BIC logLik deviance REMLdev
24.63 27.61 -9.313   16.64   18.63
Random effects:
 Groups Name          Variance Std.Dev.
operator (Intercept) 0.068084 0.26093
Residual              0.106250 0.32596
Number of obs: 20, groups: operator, 4

Fixed effects:
              Estimate Std. Error t value
(Intercept)  60.4000     0.1494   404.2
```

意味 :

総平均=60.4 =fixef(mmod) で出力可能(下記)

ランダム効果(operator)の分散=0.068084

## 固定効果(平均値)の出力法 How to output fixed effects

fixef(lmerを格納した変数名)

```
例) p162 fixef(mmod)
      (Intercept)
      60.4
```

## ランダム効果の出力法 How to output random effects

ranef(lmerを格納した変数名)\$ランダム効果に指定した変数名

```
例) p161 ranef(mmod)$operator
      (Intercept)
a -0.1219427
b -0.2591282
c  0.1676712
d  0.2133997
```

## 固定効果とランダム効果を合わせた出力 (BLUPs, the best linear unbiased predictor)

fixef(lmerを格納した変数名) + ranef(lmerを格納した変数名)\$ランダム効果の変数名

```
例) p161 fixef(mmod) + ranef(mmod)$operator
      (Intercept)
a  60.27806
b  60.14087
c  60.56767
d  60.61340
```

pvals.fnc による係数の信頼範囲 confidence interval of parameters using function "pvals.fnc"

```
mcmc <- pvals.fnc (mmod, nsim=1000)
mcmc
$fixed
  Estimate MCMCmean HPD95lower HPD95upper pMCMC Pr(>|t|)
1      60.4      60.4      60.09      60.71 0.001      0

$random
  Groups      Name Std. Dev. MCMCmedian MCMCmean HPD95lower HPD95upper
1 operator (Intercept) 0.2609      0.1874 0.2083      0.000      0.4958
2 Residual              0.3260      0.3523 0.3632      0.247      0.5011
```

モデル化 (2) :

```
smold <- lmer (bright ~ 1+ (1| operator), pulp, REML=FALSE)
```

REML=FALSE が指定され、通常の最大尤度でやっている

## p163 8.4 data (penicillin) の内容: blocks as random effects

ペニシリン製造の4つの方法 (treat) A,B,C,Dによる生産量の比較。ただし、corn steep liquor (\*) の種類にも関係するので、5種類のBlendそれぞれに対して4つの方法が試されている。

```
  treat blend yield
1     A Blend1    89
2     B Blend1    88
3     C Blend1    97
4     D Blend1    94
5     A Blend2    84
6     B Blend2    77
7     C Blend2    92
8     D Blend2    79
. . . . .
19    C Blend5    80
20    D Blend5    88
```

(\*) トウモロコシから溶出した可溶性成分と乳酸発酵で生成した成分を含む浸漬液を濃縮した液状のもの。コーンスターチの副産物の一つ。抗生物質、酵母等の培地。

モデル化 (1) :

両方を固定効果として扱う場合

```
lmod <- aov (yield ~ blend + treat, penicillin)
```

注) aovはlmをANOVA風に出力するためのwrapper関数 (p157)

モデル化 (2) :

blendをランダム効果として扱う場合

```
op <- options (contrasts = c ("contr.sum", "contr.poly"))
```

もとに戻すときは `op <- options (contrasts = c ("contr.treatment", "contr.sum" ))`

`contr.helmert` returns Helmert contrasts, which contrast the second level with the first, the third with the average of the first two, and so on.

`contr.poly` returns contrasts based on orthogonal polynomials.

`contr.sum` uses 'sum to zero contrasts'

```
mmod <- lmer(yield ~ treat + (1 | blend), penicillin)
```

```
#これによる出力
```

```
summary(lmod)
```

```
Linear mixed model fit by REML
Formula: yield ~ treat + (1 | blend)
Data: penicillin
AIC      BIC    logLik deviance REMLdev
118.6 124.6 -53.3  117.3  106.6
Random effects: # ランダム効果はblend
Groups      Name      Variance Std.Dev.
blend      (Intercept) 11.792  3.4339
Residual                    18.833  4.3397
Number of obs: 20, groups: blend, 5

Fixed effects: # 固定効果はtreat
              Estimate Std. Error t value
(Intercept)  86.000     1.817   47.34
treat1       -2.000     1.681   -1.19
treat2       -1.000     1.681   -0.59
treat3        3.000     1.681    1.78
```

意味

緑字の部分：固定効果の結果（平均） ↓

```
fixef(mmod) #とやると
(Intercept)      treat1      treat2      treat3
           86          -2          -1           3
```

↑ただし、固定効果treatは順序変数なので、順序が出る。

ランダム効果blendの結果は出ていないが、ランダム効果ごとのBLUPsは ↓

```
ranef(mmod) #とやると出る（和が0になっている ↓）
```

```
$blend
(Intercept)
Blend1  4.2878788
Blend2 -2.1439394
Blend3 -0.7146465
Blend4  1.4292929
Blend5 -2.8585859
```

coef(mmod) #とやると、「ランダム効果ごとの固定効果の平均」がでる

```
$blend
(Intercept)      treat1      treat2      treat3
Blend1  90.28788 (=86+4.28)  -2  -1  3
Blend2  83.85606 (=86-2.14)  -2  -1  3
Blend3  85.28535             -2  -1  3
Blend4  87.42929             -2  -1  3
Blend5  83.14141             -2  -1  3
```

pvals.fnc による係数の信頼範囲

```
mcmc <- pvals.fnc(mmod, nsim=1000)
```

```
mcmc
```

```
$fixed
```

	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	Pr(> t )
(Intercept)	84	84.0866	79.309	89.456	0.001	0.0000
treatB	1	0.8506	-5.367	6.657	0.764	0.7204
treatC	5	4.9987	-1.121	11.428	0.104	0.0872
treatD	2	1.8845	-4.725	8.209	0.534	0.4767

```
$random
```

Groups	Name	Std.Dev.	MCMCmedian	MCMCmean	HPD95lower	HPD95upper
1 blend	(Intercept)	3.4339	2.0085	2.1783	0.0000	5.6180
2 Residual		4.3397	4.9426	5.1224	3.1499	7.0999

## p 167 8.5 data (irrigation) の内容: Split plots

8つのfieldの各々のうち、2つずつに同じ灌漑 (irrigation) パターン (すなわち4種のirrigation) を与え、さらにその各々に異なる2品種の植物を栽培し、このときの収量yieldを記録した。したがって16パターンがある。

```
irrigation
field      irrigation variety yield
1         f1         i1      v1    35.4
2         f1         i1      v2    37.9
3         f2         i2      v1    36.7
4         f2         i2      v2    38.2
5         f3         i3      v1    34.8
6         f3         i3      v2    36.4
7         f4         i4      v1    39.5
8         f4         i4      v2    40.0
9         f5         i1      v1    41.6
10        f5         i1      v2    40.3
11        f6         i2      v1    42.7
12        f6         i2      v2    41.6
13        f7         i3      v1    43.6
14        f7         i3      v2    42.8
15        f8         i4      v1    44.5
16        f8         i4      v2    47.6
```

f 1 i 1 v 1	f1 i 1 v 2	f 5 i 1 v 1	f 5 i 1 v 2
f 2 i 2 v 1	f 2 i 2 v 2	f 6 i 2 v 1	f 6 i 2 v 2
f 3 i 3 v 1	f 3 i 3 v 2	f 7 i 3 v 1	f 7 i 3 v 2
f 4 i 4 v 1	f 4 i 4 v 2	f 8 i 4 v 1	f 8 i 4 v 2

summary (irrigation)

モデル化 (1) :

**失敗例 WRONG** : irrigationと品種とを、交互作用も入れて固定効果にする。また、fieldをランダム効果、さらに、品種もfieldにnestしてランダム効果にいれると・・・

```
lmod <- lmer (yield ~ irrigation * variety + (1|field) + (1|field:variety), data=irrigation)
```

Number of levels of a grouping factor for the random effects must be less than the number of observations (ランダム効果のグルーピングの水準の数は (Yの) 観測数よりも少なくなければならない) と出る

fieldのレベルの数=8、varietyのレベルの数=2で、

$8 * 2 = 16$ なので、 $length(yield) = 16$ と同じになるから???

→field内の品種による変動 (分散) と誤差による変動が区別できない。これらを区別するには、各field内で1つの品種につき2個以上の測定が必要。(p168) と注釈あり

モデル化 (2) : **正解 CORRECT**

```
lmodr <- lmer (yield ~ irrigation * variety + (1|field), data=irrigation)
```

灌漑レベルと品種とを交互作用も入れて固定効果にし、fieldのみをランダム効果にする

#これによる出力

```
summary(lmodr)
Linear mixed model fit by REML
Formula: yield ~ irrigation * variety + (1 | field)
Data: irrigation
   AIC   BIC logLik deviance REMLdev
65.4 73.12 -22.70   68.61   45.39
Random effects:
Groups   Name      Variance Std.Dev.
field    (Intercept) 16.2000  4.0249
Residual                    2.1075  1.4517
Number of obs: 16, groups: field, 8
```

```

Fixed effects:
              Estimate Std. Error t value
(Intercept)   38.500    3.025  12.726
irrigationi2    1.200    4.279   0.280
irrigationi3    0.700    4.279   0.164
irrigationi4    3.500    4.279   0.818
varietyv2       0.600    1.452   0.413
irrigationi2:varietyv2 -0.400    2.053  -0.195
irrigationi3:varietyv2 -0.200    2.053  -0.097
irrigationi4:varietyv2  1.200    2.053   0.584

Correlation of Fixed Effects:
(Intr) irrgrt2 irrgrt3 irrgrt4 vrtyv2 irr2:2 irr3:2
irrigation2 -0.707
irrigation3 -0.707  0.500
irrigation4 -0.707  0.500  0.500
varietyv2   -0.240  0.170  0.170  0.170
irrgrtn2:vr2  0.170 -0.240 -0.120 -0.120 -0.707
irrgrtn3:vr2  0.170 -0.120 -0.240 -0.120 -0.707  0.500
irrgrtn4:vr2  0.170 -0.120 -0.120 -0.240 -0.707  0.500  0.500

```

pvals.fnc による係数の信頼範囲 `mcmc <- pvals.fnc(lmodr, nsim = 1000)`

```

$fixed
              Estimate MCMCmean HPD95lower HPD95upper pMCMC Pr(>|t|)
(Intercept)   38.5    38.5742    31.855    45.707  0.001  0.0000
irrigationi2    1.2    1.2128    -7.979    10.582  0.760  0.7862
irrigationi3    0.7    0.7811    -7.416     9.640  0.844  0.8741
irrigationi4    3.5    3.5910    -6.049    13.149  0.424  0.4370
varietyv2       0.6    0.5479    -7.968     9.758  0.896  0.6902
irrigationi2:varietyv2 -0.4 -0.6091 -13.092    11.945  0.908  0.8504
irrigationi3:varietyv2 -0.2 -0.3427 -11.322    12.282  0.962  0.9248
irrigationi4:varietyv2  1.2  1.2085 -12.293    13.860  0.802  0.5750

$random
  Groups      Name Std. Dev. MCMCmedian MCMCmean HPD95lower HPD95upper
1  field (Intercept)  4.0249    1.0615    1.2249    0.0000    3.3114
2 Residual          1.4517    3.9660    4.2242    2.1251    7.3168

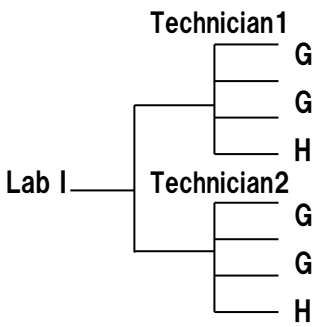
```

## p 170 8.6 data (eggs) の内容: Nested effects

6つのラボに、8つずつサンプルを送る。この8つを、各ラボは2人の技師に4個ずつ配る。この4個は、GとHという名前の2種類のサンプルが2個ずつである。これらの脂肪の量を量る。研究の目的はラボ間で一貫した結果が得られるかということ。また、実はGとHは全く同じもの。

eggs

	Fat	Lab	Technician	Sample
1	0.62	I	one	G
2	0.55	I	one	G
3	0.34	I	one	H
4	0.24	I	one	H
5	0.80	I	two	G
...	...	...	...	...
8	0.65	I	two	H
9	0.30	II	one	G
...	...	...	...	...
47	0.26	VI	two	H
48	0.06	VI	two	H



Technician には one, two という名前がついているが、同じ one でもラボ間では異なる人なので nest する

モデル化 (1) :

固定効果はFatの量で、これは同じであるはず。技師とサンプルは無作為に選ばれたものだと考えるので、ランダム効果。また、研究の目的はラボ間で一貫した結果が得られるかということなので、ラボもランダム効果。(もしラボごとに注目する場合は固定効果にする) サンプルは技師内でネストする。ただし、これらのランダム効果は、ラボ別、各ラボ内の技師別、各ラボの各技師のサンプル別、にnestする。



```
cmodr <- lmer (Fat ~ 1 + (1|Lab) + (1|Lab:Technician) + (1|Lab:Technician:Sample), data=eggs)
```

**#これによる出力：**

```
Linear mixed model fit by REML
Formula: Fat ~ 1 + (1 | Lab) + (1 | Lab:Technician) + (1 | Lab:Technician:Sample)
Data: eggs
   AIC   BIC logLik deviance REMLdev
-54.24 -44.88  32.12   -68.7   -64.24
Random effects:
Groups                Name          Variance Std.Dev.
Lab:Technician:Sample (Intercept) 0.0030646 0.055359
Lab:Technician        (Intercept) 0.0069802 0.083548
Lab                   (Intercept) 0.0059199 0.076941
Residual              0.0071958 0.084828
Number of obs: 48, groups: Lab:Technician:Sample, 24; Lab:Technician, 12; Lab, 6

Fixed effects:
              Estimate Std. Error t value
(Intercept)  0.38750    0.04296   9.019
```

**pvals.fnc** による係数の信頼範囲

```
mcmc
$fixed
  Estimate MCMCmean HPD95lower HPD95upper pMCMC Pr(>|t|)
1  0.3875   0.3869   0.3105   0.4604 0.001      0

$random
      Groups          Name Std. Dev. MCMCmedian MCMCmean HPD95lower
1 Lab:Technician:Sample (Intercept)  0.0554   0.0106   0.0152   0.0000
2      Lab:Technician (Intercept)  0.0835   0.0475   0.0480   0.0000
3              Lab (Intercept)  0.0769   0.0633   0.0631   0.0000
4      Residual              0.0848   0.1108   0.1123   0.0889
HPD95upper
1  0.0477
2  0.0937
3  0.1172
4  0.1441
```

**モデル化(2)：**

Sample は無しにしてもよいかもしれないので次も試す

```
cmodr <- lmer (Fat ~ 1 + (1|Lab) + (1|Lab:Technician), data=eggs)
```

**#これによる出力：**

```
Linear mixed model fit by REML
Formula: Fat ~ 1 + (1 | Lab) + (1 | Lab:Technician)
Data: eggs
   AIC   BIC logLik deviance REMLdev
-54.63 -47.15  31.32   -67.1   -62.63
Random effects:
Groups                Name          Variance Std.Dev.
Lab:Technician        (Intercept) 0.0080017 0.089452
Lab                   (Intercept) 0.0059199 0.076941
Residual              0.0092389 0.096119
Number of obs: 48, groups: Lab:Technician, 12; Lab, 6

Fixed effects:
              Estimate Std. Error t value
(Intercept)  0.38750    0.04296   9.019
```

*AICはこっちのほうが低いのでこのモデルのほうがよい???*

*p173-4では尤度比検定LRTでやっている。*

**pvals.fnc** による係数の信頼範囲

```
mcmc <- pvals.fnc(cmodr, nsim=1000)
mcmc
$fixed
  Estimate MCMCmean HPD95lower HPD95upper pMCMC Pr(>|t|)
1  0.3875   0.3894   0.3223   0.4668 0.001      0
```

```

$random
      Groups      Name Std. Dev. MCMCmedian MCMCmean HPD95lower HPD95upper
1 Lab:Technician (Intercept) 0.0895    0.0532    0.0516    0.0000    0.0927
2 Lab (Intercept)          0.0769    0.0629    0.0625    0.0000    0.1129
3 Residual                  0.0961    0.1122    0.1138    0.0892    0.1452

```

```

> 2*(logLik(cmod) - logLik(cmodr))
REML
1.603423

```

これと、ブートストラップでやって求めたLRT (likelihood ratio test statistic, p158) から、sampleの変動は無視できる、としている (p172)

## P173 8.7 data (abrasion) の内容: Crossed effects

ラテン方格構造のデータ (縦方向も横方向も要素が全部違う構造)。4サンプルが同時に試験できる摩耗検査機に4つの材料A,B,C,Dを入れ、摩耗度を検査する。この4つの位置ごとにも、試験runごとにも結果が違ふようであり、4回試験した。(右図では行がposition,列がrun)

```

abrasion
run position material wear
1 1 1 C 235
2 1 2 D 236
3 1 3 B 218
4 1 4 A 268
5 2 1 A 251
6 2 2 B 241
...
16 4 4 D 225

```

C	A	D	B
D	B	C	A
B	D	A	C
A	C	B	D

### モデル化 (1) :

摩耗度wearが材料material、position、runで変わるか

```
lmod <- aov(wear ~ material + run + position, abrasion)
```

#これによる出力:

```

どれも有意だった (結果省略)
summary(lmod)
      Df Sum Sq Mean Sq F value    Pr(>F)
material  3 4621.5  1540.5  25.1510 0.0008498 ***
run       3  986.5   328.8   5.3687 0.0390130 *
position  3 1468.5   489.5   7.9918 0.0161685 *
Residuals 6  367.5    61.2

```

### モデル化 (2) :

摩耗度wearに対し、材料materialだけを固定効果、position、runはランダム効果と見なす。nestされていないので・・・

```
mmod <- lmer(wear ~ material + (1|run) + (1|position), abrasion)
```

#これによる出力:

```

Linear mixed model fit by REML
Formula: wear ~ material + (1 | run) + (1 | position)
Data: abrasion
      AIC      BIC logLik deviance REMLdev
114.3 119.7 -50.13  120.4  100.3
Random effects:
Groups      Name      Variance Std. Dev.

```

```
run      (Intercept) 66.896  8.1790
position (Intercept) 107.062 10.3471
Residual                61.250  7.8262
Number of obs: 16, groups: run, 4; position, 4
```

```
Fixed effects:
              Estimate Std. Error t value
(Intercept)  265.750    7.668    34.66
materialB    -45.750    5.534    -8.27
materialC    -24.000    5.534    -4.34
materialD    -35.250    5.534    -6.37
```

```
Correlation of Fixed Effects:
              (Intr) matrIB matrIC
materialB  -0.361
materialC  -0.361  0.500
materialD  -0.361  0.500  0.500
```

固定効果の有意性はパラメトリックbootstrapで検定できるが、「t が大きいのでmaterialの効果があることは明らか」らしい (Errata p9より)

### pvals.fnc による係数の信頼範囲

```
mcmc <- pvals.fnc(mmod, nsim=10000)
```

```
mcmc
$fixed
```

	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	Pr(> t )
(Intercept)	265.75	265.84	248.32	283.074	0.0001	0.000
materialB	-45.75	-45.74	-63.95	-26.753	0.0004	0.000
materialC	-24.00	-24.04	-43.48	-6.047	0.0194	0.001
materialD	-35.25	-35.26	-54.02	-16.309	0.0024	0.000

```
$random
```

Groups	Name	Std. Dev.	MCMCmedian	MCMCmean	HPD95lower	HPD95upper
1	run (Intercept)	8.1790	4.7182	5.6719	0.0000	15.5149
2	position (Intercept)	10.3471	6.0975	6.7799	0.0000	16.3480
3	Residual	7.8262	12.0840	12.6926	6.5554	19.9656

## p174 8.8 data (jsp) の内容 : 多層モデル multilevel models

学校school、クラスclass、性別gender、親の社会的ランクsocial、入学時の知能テストの成績raven、個人番号id、1～3年時の英語のテスト、1～3年時の数学のテスト、入学後の学年year、別のデータ

```
head(jsp)
```

```
  school class gender social raven id english math year
1      1     1  girl     9     23  1      72   23    0
2      1     1  girl     9     23  1      80   24    1
3      1     1  girl     9     23  1      39   23    2
4      1     1  boy     2     15  2       7   14    0
5      1     1  boy     2     15  2      17   11    1
6      1     1  boy     2     22  3      88   36    0
```

3年次の数学の成績が何の影響を受けるか、を考える。

```
jspr <- jsp[jsp$year==2,]
```

性別gender、親の社会的ランクsocial、入学時の知能テストの成績ravenに目をつけて、学校school および クラスclassをランダム効果にいれこむ。(ただし予備解析でgenderは関係なさそうだったのでこれもはずす。また、ravenも平均値で標準化)

```
jspr$craven <- jspr$raven - mean(jspr$raven)
```

```
mmod <- lmer(math ~ craven*social + (1|school) + (1|school:class), data=jspr)
```

#これによる出力:

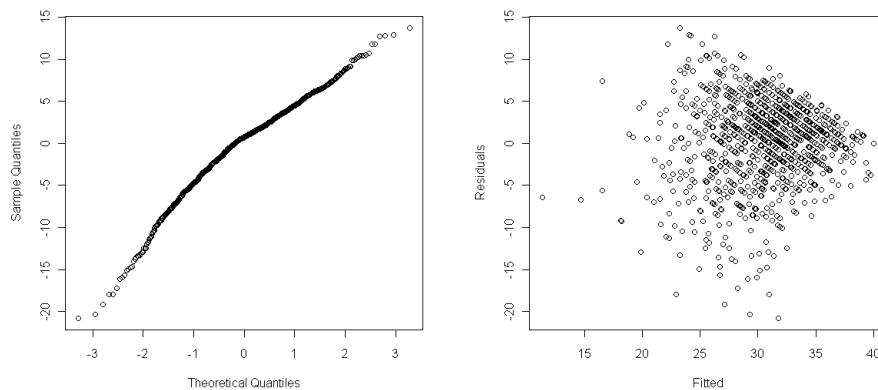
```
summary(mmod)
Linear mixed model fit by REML
Formula: math ~ craven * social + (1 | school) + (1 | school:class)
Data: jspr
   AIC   BIC logLik deviance REMLdev
5963 6065  -2961   5907    5921
Random effects:
Groups      Name      Variance Std. Dev.
school:class (Intercept)  1.1774  1.0851
school      (Intercept)  3.1477  1.7742
Residual                                27.1412  5.2097
Number of obs: 953, groups: school:class, 90; school, 48
```

```
Fixed effects:
Estimate Std. Error t value
(Intercept) 31.91127  1.19554  26.692
craven      0.60585  0.18854   3.213
social2     0.02362  1.27219   0.019
social3    -0.63073  1.30887  -0.482
social4    -1.96707  1.19707  -1.643
social5    -1.35849  1.30022  -1.045
social6    -2.26870  1.37373  -1.651
social7    -2.55182  1.40554  -1.816
social8    -3.39499  1.80135  -1.885
social9    -0.83133  1.25346  -0.663
craven:social2 -0.13208  0.20579  -0.642
craven:social3 -0.22433  0.21888  -1.025
craven:social4  0.03581  0.19488   0.184
craven:social5 -0.15035  0.20889  -0.720
craven:social6 -0.03861  0.23259  -0.166
craven:social7  0.39825  0.23176   1.718
craven:social8  0.25599  0.26154   0.979
craven:social9 -0.08103  0.20550  -0.394
```

mcmc1 <- pvals.fnc(mmod, nsim=1000)

```
mcmc1
$fixed
Estimate MCMCmean HPD95lower HPD95upper pMCMC Pr(>|t|)
(Intercept) 31.9113 31.9345 29.7232 34.3475 0.001 0.0000
craven      0.6059  0.6051  0.2568  1.0237 0.002 0.0014
social2     0.0236  0.0340 -2.4319  2.4343 0.972 0.9852
social3    -0.6307 -0.6547 -3.3087  1.7078 0.634 0.6300
social4    -1.9671 -2.0022 -4.4037  0.4998 0.098 0.1007
social5    -1.3585 -1.2986 -3.6729  1.3829 0.304 0.2964
social6    -2.2687 -2.2801 -5.2775  0.3244 0.100 0.0990
social7    -2.5518 -2.5315 -5.2890  0.1648 0.068 0.0698
social8    -3.3950 -3.3847 -6.9614  0.1097 0.052 0.0598
social9    -0.8313 -0.8399 -3.4581  1.5226 0.492 0.5073
craven:social2 -0.1321 -0.1337 -0.5406  0.2822 0.538 0.5212
craven:social3 -0.2243 -0.2182 -0.6473  0.2023 0.312 0.3057
craven:social4  0.0358  0.0354 -0.3558  0.4306 0.872 0.8543
craven:social5 -0.1504 -0.1552 -0.5768  0.2727 0.456 0.4718
craven:social6 -0.0386 -0.0263 -0.4752  0.4259 0.876 0.8682
craven:social7  0.3982  0.4020 -0.0822  0.8259 0.110 0.0861
craven:social8  0.2560  0.2477 -0.3089  0.6867 0.346 0.3279
craven:social9 -0.0810 -0.0779 -0.4971  0.3188 0.714 0.6934
```

```
$random
Groups      Name      Std. Dev. MCMCmedian MCMCmean HPD95lower HPD95upper
1 school:class (Intercept)  1.0851  0.6668  0.6809  0.0000  1.4764
2 school      (Intercept)  1.7742  1.6493  1.6232  0.9086  2.2875
3 Residual                                5.2097  5.2695  5.2703  5.0106  5.5282
```



```
qqnorm (resid (mmod) ,main="")
```

```
plot (fitted (mmod) ,resid (mmod) ,xlab="Fitted",ylab="Residuals")
```

とやると（上図）、推定値fittedが大きくなるほど分散が小さくなる傾向がわかる。

## Chapter9 Repeated measures and longitudinal data

### 9.1 data (psid) の内容

1968時点で25-39歳の1968-1990の85人の世帯データ

```
head (psid)
```

	age	educ	sex	income	year	person
1	31	12	M	6000	68	1
2	31	12	M	5300	69	1
3	31	12	M	5200	70	1
4	31	12	M	6900	71	1
5	31	12	M	7500	72	1
6	31	12	M	8000	73	1

```
psid$year <- psid$year - 78
```

```
mmod <- lmer (log (income) ~ cyear * sex + age + educ + (cyear | person) , psid)
```

```
print (summary (mmod) , correlation = FALSE)
```

```
Linear mixed model fit by REML
Formula: log (income) ~ cyear * sex + age + educ + (cyear | person)
Data: psid
   AIC   BIC logLik deviance REMLdev
3840 3894  -1910   3786    3820
Random effects:
Groups   Name      Variance Std. Dev.  Corr
person  (Intercept)  0.2816564  0.53071
        cyear      0.0024000  0.04899   0.187
Residual    0.4672724  0.68357
Number of obs: 1661, groups: person, 85
```

```
Fixed effects:
              Estimate Std. Error t value
(Intercept)  6.674178   0.543334  12.284
cyear        0.085312   0.008999   9.480
sexM         1.150315   0.121293   9.484
age          0.010932   0.013524   0.808
educ         0.104212   0.021437   4.861
cyear:sexM  -0.026307   0.012238  -2.150
```

```
mcmc1 <- pvals.fnc (mmod, nsim = 1000)
```

以下にエラー pvals.fnc (mmod, nsim = 1000) : と出て計算してくれない

MCMC sampling is not yet implemented in lme4\_0.999375  
for models with random correlation parameters

```
mcmc1 <- mcmc1 <- mcmc1(mmod, n=1000, saveb = TRUE)
以下にエラー .local(object, n, verbose, ...) : と出て計算してくれない
Code for non-trivial theta_T not yet written
```

しかし、lmeを使うとうまくいく・・・

```
mmodlme <- lme(log(income) ~ cyear * sex + age + educ, random = ~ cyear | person, psid)
summary(mmodlme)
Linear mixed-effects model fit by REML
Data: psid
      AIC      BIC    logLik
3839.776 3893.892 -1909.888
Random effects:
Formula: ~cyear | person
Structure: General positive-definite, Log-Cholesky parametrization
      StdDev      Corr
(Intercept) 0.53071321 (Intr)
cyear       0.04898952 0.187
Residual    0.68357323

Fixed effects: log(income) ~ cyear * sex + age + educ
      Value Std. Error  DF  t-value p-value
(Intercept) 6.674204 0.5433252 1574 12.283995 0.0000
cyear       0.085312 0.0089996 1574  9.479521 0.0000
sexM        1.150313 0.1212925   81  9.483790 0.0000
age         0.010932 0.0135238   81  0.808342 0.4213
educ        0.104210 0.0214366   81  4.861287 0.0000
cyear:sexM -0.026307 0.0122378 1574 -2.149607 0.0317
Correlation:
      (Intr) cyear  sexM  age  educ
cyear  0.020
sexM   -0.104 -0.098
age    -0.874  0.002 -0.026
educ   -0.597  0.000  0.008  0.167
cyear:sexM -0.003 -0.735  0.156 -0.010 -0.011

Standardized Within-Group Residuals:
      Min      Q1      Med      Q3      Max
-10.23102885 -0.21344108  0.07945029  0.41471605  2.82543559

Number of Observations: 1661
Number of Groups: 85
```

```
glmres <- glm(PULSE ~ 1, offset=log(YEARS), family = poisson, mydata)
summary(glmres)
Call:
glm(formula = PULSE ~ 1, family = poisson, data = mydata, offset = log(YEARS))

Deviance Residuals:
      Min       1Q   Median       3Q      Max
-10.742  -2.489   1.018   4.724  18.618
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.5621     0.0191   81.8 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1803.3 on 38 degrees of freedom
Residual deviance: 1803.3 on 38 degrees of freedom
AIC: 2042.6
```

Number of Fisher Scoring iterations: 5

以下は上と同じ

```
glmres2 <- glm(PULSE ~ offset(log(YEARS)), family = poisson, mydata)
summary(glmres2)
```

**尤度、最大尤度 likelihood, maximum likelihood (モデル選択 p 13、中妻 p 56)**

尤度=データ  $X$  を観察したときのパラメタ  $\theta$  の尤もらしさ。尤度関数は次のように定義される。

$$p(X|\theta) = p(x_1|\theta)p(x_2|\theta) \cdots p(x_N|\theta) = \prod [p(x_i|\theta)]$$

ここで、 $X$  は  $N$  個の観測地から成るベクトル

$$X = (x_1, x_2, \dots, x_N)$$

であり、 $p(X|\theta)$  はパラメタ  $\theta$  が仮に真の値であった時にデータ  $X$  が得られる確率である (中妻「入門ベイズ統計学」p 56)。実際に尤度を計算する際は、単調増加関数にするために、この両辺の対数をとった対数尤度関数を取り扱う。すなわち、(対数)尤度関数は次式で定義される。

$$\ell(\theta|X) = \log p(X|\theta) = \log \prod [p(x_i|\theta)] = \sum \log p(x_i|\theta)$$

(尤度は、データ  $X$  が与えられた時のパラメタ  $\theta$  の尤もらしさ、としてあら得られていることに注意。)

尤度を最大にするような  $\theta$  が最大尤度推定量 maximum likelihood estimate, MLE。たぶん確率  $p$  を何かの式で表したとき、 $\theta \sim \ell(\theta|X)$  関係が最大になるように (実際には  $\ell(\theta|X)$  を  $\theta$  で微分した  $\sum \{d \log [p(x_i|\theta)] / d\theta\}$  が 0 になるように) 数値的に  $\theta$  を決めている。

likelihood = likelihood of parameter  $\theta$  given data  $X$  is observed. Likelihood function is defined as:

$$p(X|\theta) = p(x_1|\theta)p(x_2|\theta) \cdots p(x_N|\theta) = \prod [p(x_i|\theta)],$$

where  $X$  is a vector composed of  $N$  observations

$$X = (x_1, x_2, \dots, x_N)$$

and  $p(X|\theta)$  represents the probability of observing data  $X$  if parameter  $\theta$  is the true value.

When calculating likelihood, logarithm of the equation is taken to make it monotone increasing function. That is, the (log) likelihood function is defined as follows:

$$\ell(\theta|X) = \log p(X|\theta) = \log \prod [p(x_i|\theta)] = \sum \log p(x_i|\theta)$$

(Note that the likelihood is defined such that the likelihood of parameter  $\theta$  given data  $X$ .)

The  $\theta$  that maximizes the likelihood is called maximum likelihood estimate (MLE). *Perhaps* if the probability  $p$  is expressed by an equation, MLE is numerically estimated such that the derivative of  $\theta \sim \ell(\theta|X)$  relationship differentiated with  $\theta$  is maximum.

**最尤法とモデルの包括関係 (モデル選択 p 20)**

最尤法によるモデル選択を変数選択にそのまま用いてもうまくいかない。一般に2つの集合  $S_k$  と  $S_{k'}$  が包含関係  $S_k \subset S_{k'}$  にあるとき、 $\sigma_k^2 \geq \sigma_{k'}^2$  (残差分散) であるから、 $\ell_k \leq \ell_{k'}$  (尤度) となり、最尤法ではどのようなデータを入力しても  $k$  (パラメタ数) が最大のモデルが常に最大対数尤度が大きい。

Applying model selection by maximum likelihood to variable selection does not work when maximum likelihood is used directly. In general, if a set  $S_k$  is included in a set  $S_{k'}$  or  $S_k \subset S_{k'}$ , the variance of residuals are in the relationship  $\sigma_k^2 \geq \sigma_{k'}^2$ , and so their likelihoods are  $\ell_k \leq \ell_{k'}$ , so that the maximum likelihood for the model with the largest  $k$  (or those with largest number of parameters) are selected with any data.

## (対数) 尤度比検定 (log) likelihood-ratio test (モデル選択 p 21~)

包含関係にある二つのモデル  $M_k \subset M_{k'}$  のうち、小さいほうのモデル  $M_k$  が正しいと仮定する。すなわち真の確率密度関数  $q(X)$  が  $M_k$  の要素だと仮定する。このとき、 $q \in M_k \subset M_{k'}$  なので、大きいほうのモデル  $M_{k'}$  も正しいくなり、よって小さいモデルを選びたくなるが、最大対数尤度の差

$$\Delta \ell = \ell_{k'}(\theta_{k'} | X) - \ell_k(\theta_k | X)$$

は 0 または正であり、最尤法によるモデル選択では  $M_{k'}$  が選ばれてしまう。もし  $M_{k'}$  は正しいが  $M_k$  は正しくない場合は  $\Delta \ell$  はさらに大きくなる傾向があるはずだから、ある閾値 (=  $p$  のこと) を決めておいて、それより  $\Delta \ell$  が小さい場合には  $M_k$  を選ぶことにする。

2  $\Delta \ell$  は自由度  $m = \dim \theta_{k'} - \dim \theta_k$  ( $\theta_{k'}, \theta_k$  の自由度  $\equiv$  データ数) のカイ二乗分布に 近似的 (\*) に従うことが知られているので、これを利用して検定を行う。(\* Faraway p159-160 では問題有り (conservative になる) として parametric bootstrap で  $p$  を計算することを推奨) ]

Assume that smaller model  $M_k$  is correct between two models with  $M_k \subset M_{k'}$ . Put another way, assume that actual probability function  $q(X)$  is a component of  $M_k$ . In this case, we may want to select smaller model as  $q \in M_k \subset M_{k'}$  and hence larger model  $M_{k'}$  is also correct. However, difference in the two maximum log-likelihoods,

$$\Delta \ell = \ell_{k'}(\theta_{k'} | X) - \ell_k(\theta_k | X)$$

is 0 or positive, so that  $M_{k'}$  is always selected by the model selection using maximum likelihood. If  $M_{k'}$  is correct but  $M_k$  is not correct, then  $\Delta \ell$  should become greater, so we determine a threshold value beforehand (i.e,  $p$  value), and we select  $M_k$  if  $\Delta \ell$  is smaller than the threshold.

Because it is known that  $2 \Delta \ell$  approximately (\*) follows  $\chi^2$ -distribution with degrees of freedom  $m = \dim \theta_{k'} - \dim \theta_k$  (d.f. of  $\theta_{k'}, \theta_k \equiv$  the number of data), a test is possible. (\* Note: in Faraway p159-160, as the result by this method can become conservative, it is advised to calculate  $p$  value by parametric bootstrapping.)

### 検定法 how to test

自由度  $m$  のカイ二乗分布に従う確率変数が  $x$  以下になる確率を  $p = F_m(x)$  とする。

仮説:  $M_k$  (小さいモデル) が正しい とする。

$p$  (例, 0.05) は、 $M_k$  が正しいにもかかわらず  $M_{k'}$  を選ぶ確率、すなわち

$M_k$  を選ぶこと = 仮説の受容 ( $p > 0.05$ )

$M_{k'}$  を選ぶこと = 仮説の棄却 ( $p < 0.05$ )

パラメーターが ( $2 \Delta \ell$ , 自由度  $m$ ) のカイ二乗分布の  $p$  値を  $R$  で求める (Faraway p159)

`pchisq(2 *  $\Delta \ell$ , m, lower = FALSE) > 0.05` なら小さいモデルを選ぶ。

Let the probability  $p$  that a probability variable, which follows  $\chi^2$ -distribution with degrees of freedom  $m$ , be equal to or smaller than  $x$  is

$$p = F_m(x)$$

Now we set  $H_0$ :  $M_k$  (smaller model) is correct).

The  $p$  (e.g., 0.05) shows the probability of selecting  $M_{k'}$  despite that  $M_k$  is correct, i.e.:

To select  $M_k$  = acceptance of  $H_0$  ( $p > 0.05$ ).

To select  $M_{k'}$  = rejection of  $H_0$  ( $p < 0.05$ .)

We estimate the  $p$ -value of  $\chi^2$ -distribution with parameters of ( $2 \Delta \ell$ , d.f. =  $m$ ) using [R] (Faraway p 159). If

`pchisq(2 *  $\Delta \ell$ , m, lower = FALSE) > 0.05`, then choose smaller model.



Appendix B

「比率」の変数を使うな、offset 関数を使え！ Do not use "ratio" variable: use offset function!

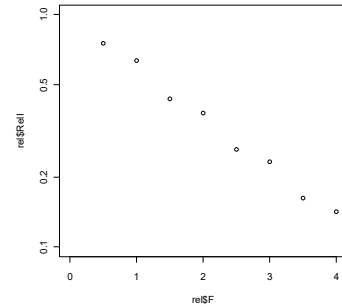
offset ( ) (Vit pp197, 293)

オフセット項：(通常係数が1になるように) あらかじめ係数を固定した説明変数

Predictors with fixed coefficients are referred to as offsets (Vit P197) , with a coefficient of 1 (Vit p 293)

```
rel <- read.csv("RelL.csv")
rel
```

	F	I0	I	RelI
1	0.5	1000	751.41485	0.7514149
2	1.0	800	507.67630	0.6345954
3	1.5	1500	651.18446	0.4341230
4	2.0	250	94.02838	0.3761135
5	2.5	743	195.52361	0.2631543
6	3.0	291	67.72321	0.2327258
7	3.5	323	52.46260	0.1624229
8	4.0	1100	156.33296	0.1421209



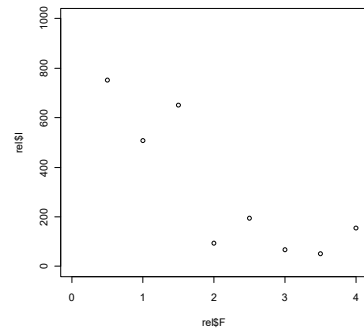
実は Beer-Lambert の法則  $I/I_0 = \exp(-0.5 F)$  のデータ(右図)

```
plot(rel$F, rel$RelI, xlim = c(0, 4), ylim=c(0.1, 1), log="y")
```

すなわち、 $\log(\text{relative light}) = \log(I/I_0) = -0.5 F$  になっている。

F と群落内の光 I との関係を、群落上の光 I0 も考慮して解析しようとした

```
plot(rel$F, rel$I, xlim = c(0, 4), ylim=c(0, 1000))
RIRI2 <- lm( (log(I)) ~ F + log(I0), data = rel)
summary(RIRI2)
```



```
Call:
lm(formula = (log(I)) ~ F + log(I0), data = rel)
```

Residuals:

	1	2	3	4	5	6	7	8
	-0.00578	0.06891	-0.04926	0.01811	-0.06790	0.03792	-0.07123	0.06924

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.11296	0.28034	0.403	0.704
F	-0.49666	0.02238	-22.190	3.45e-06 ***
log(I0)	0.97906	0.04018	24.364	2.17e-06 ***

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06849 on 5 degrees of freedom  
 Multiple R-squared: 0.9969, Adjusted R-squared: 0.9957  
 F-statistic: 808.5 on 2 and 5 DF, p-value: 5.276e-07

これで得られる式は、

$$\log(I) = -0.49666 F + 0.97906 \log(I_0) + 0.11296$$

となり、 $\log(I_0)$ に係る係数がじゃまになって相対光強度 ( $I/I_0$ ) に換算できない。 $0.97906 \log(I_0)$ の係数は1であってほしい。そこで、offset ( ) を使うと、係数を1に固定してくれる。

```
RIRI <- lm ( log (I) ) ~ F + offset ( log (I0) ), rel)
```

```
summary (RIRI)
```

```
Call:
```

```
lm(formula = (log(I)) ~ F + offset(log(I0)), data = rel)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.07251 -0.06286  0.01465  0.05075  0.06851
```

```
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03045    0.05002  -0.609    0.565
F            -0.49282    0.01981 -24.876 2.78e-07 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.06419 on 6 degrees of freedom
Multiple R-squared:  0.9968,    Adjusted R-squared:  0.9963
F-statistic: 1870 on 1 and 6 DF,  p-value: 1.024e-08
```

これで得られた式は、

$$\log(I) = -0.49282 F + \log(I_0) - 0.03045$$

すなわち、 $\log(I/I_0) = -0.49282 F - 0.03045$

$$I/I_0 = \exp(-0.49282 F) * \exp(-0.03045) = 0.97009 \exp(-0.49282 F)$$

となり、ほぼ Beer-Lambert の法則に従うことがわかる。

```
curve(exp(-0.03045)*exp(-0.49282*x), add = TRUE)
```

