モデル選択とモデル評価

- なぜモデル評価が必要か?
- AIC（赤池情報量基準）の問題点
- R_D 基準の活用

```
# ここからは簡単のため, 次の2トラップのデータを用いる（基礎編と同じ）
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)
```

検定の目的(Fisherの考え)

- かつてFisherは, 統計処理の目的は「データの縮約」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
  (2)同定されたモデルのパラメーター推定の問題
  (3)推定されたパラメーターの推測分布の問題。

- つまり, まず有意性検定を行って, 有意になったパラメーターだけを暫定的に残し, その残されたパラメーターについてだけ推定を行う。

- しかし, こうしたFisherの意図に反して, 検定を統計処理の最終目標だと解釈する人々が現在では非常に多い。

検定には意味がない

- 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。
- 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）。

検定に代わる手法とは?

- Fisherの手法「有意性検定によるモデル選択法」は実用的な方法だったが, 先述のように論理的に不備があった。モデル選択の手段として, 現在では別の手法を用いるべきであろう。
- では, どのような基準で選択するべきか?
- 一般に, モデルは手持ちのデータだけを記述すること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のが唯一の妥当な評価法だと言える。そうした評価法の一つがAIC(赤池情報量基準)である。

AIC の考え方

- 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。
- 尤度をLとし，モデルに含まれるパラメーター数（切片を含む）をkとするとき，AICは次式で定義される。

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

- 「真のモデル」を選ぶことを目指しているわけではないので，検定のような「論理的矛盾」が生じない。

AIC の問題点

- 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」の現実的意義が明確ではない。そのため，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相対的に意味がある。つまり，AICは量的変数ではなく順序変数でしかない。
- 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を用いた場合には，これは「相対的な尺度」であるから，「モデル選択」はできても絶対的な「モデル評価」を行うことができない。
- 検定と同様に，データ量が多ければ「もっとも複雑なモデル」が採用されて議論が終了するだけであり，そのモデルの有用性を評価することはできない。

R_D 指数の提案 (Yamamura 2016)

- Kullback–Leibler 情報量の尺度ではなく,「実際に当たる確率」で正しくモデルを評価するべきではないか？ そうすれば, 選択されたモデルが「有用なモデル」かどうかを判断できる。また, 単に「予測力最大」ではなく, コストが少なく「適度の予測力」を持つモデルも選択可能になる。
- 尤度ではなく発生確率で比較するために, ラプラス哲学にしたがって「真のモデル」を飽和モデルなどもっとも複雑なモデルに固定する。(固定しなければ発生確率の比較にはならない。)
- その上で, 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」が予測した場合にR_{pred}が0%となるように改善割合R_{pred}を定義する。このR_{pred}の推定値としてR_D指数が導出されている。(なお, Laplaceは無神論者だったので,「神」とは言わずに「Intelligence」と言った。)

R_D 指数の定義式

- **R_Dの定義式**

$$R_D = 1 - \frac{l_{\max} - l + k}{l_{\max} - l_{\text{null}} + 1}$$

- ✓ l は候補モデルでの対数尤度
- ✓ l_{null}は切片だけを含むモデル (null model) における対数尤度
- ✓ $l_{\max}$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 (maximum model) における対数尤度。
- ✓ ただし, まず「真のモデル」から分散パラメーターを推定し, その分散パラメーター（つまり確率の定義）を固定して, これらの値を計算。
- ✓ 分母の1は切片の数であるため, 多変量の場合は, 分母の1を変量の数に置き換える。

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

- R用の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。stepAIC関数と同じく，交互作用に関しては，パラメーターの優劣関係を自動的に判定してくれる。要因数が多すぎて誤差の自由度が少ない場合には，あらかじめ分散を推定してから，その分散を固定して指定することもできる。

- 一方，SAS用の関数を使えば一般化線型混合モデルにおいて R_D の計算を自動的に行うことができる。stepAIC関数と同じく，交互作用に関しては，パラメーターの優劣関係を自動的に判定してくれる。

R_Dの計算例(LM)

- $\log_e(x+0.5)$近似を用いた場合(ハスモンヨトウのデータで2トラップの場合)

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

- 出力の一部

```
# RD ranking for the hierarchical family of models #
          RD                                                Model
1  0.90679867                    log(y+0.5) ~ 1 + trap + month
2  0.88850757   log(y+0.5) ~ 1 + trap + month + trap:month
3  0.84623461                       log(y+0.5) ~ 1 + month
4  0.06056406                        log(y+0.5) ~ 1 + trap
5  0.00000000                              log(y+0.5) ~ 1
```

- もっとも予測力が高いモデルは交互作用を無視したモデルである。その予測力の改善割合の推定値は$R_D = 0.90679867$であり, 予測力は十分に高いことが分かる。

R_Dの計算例（GLMM）

- R関数は一般化線形モデルまでしか対応していないが，SASマクロは一般化線形混合モデル（GLMM）まで対応している。結果は以下の通り。最新版 RDcompare 関数では method=laplace をデフォルトとしているのでnlmixed のデフォルト（qpoints=1）による計算と同じ。

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
5	1	0.00000	0.00000	

一般化線形混合モデルによる結果は$\log_e(x+0.5)$に関する線形モデル分析の結果と近い。個体数が大きい場合には，このようにGLMMをLMで近似することができる。

R_Dの計算SASプログラム(GLMM)

```
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)
```

R_Dでのモデル選択はAICよりも柔軟性が高い

- 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。
- 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，利用しやすいモデルを採用するべきである。たとえば，予測力が1%程度低下するだけなら，それは大したロスではないことも多い。むしろ，効率の悪い説明変数を測定するコストを避ける方がよい。
- 上位モデル群の中から「良いモデル」を選ぶ。(1)説明変数の数が少ないモデルや(2)容易に観測できる説明変数のみからなるモデル，(3)容易に解釈できるモデル，などが良いであろう。
- これに対して，AICは相対的な指数なのでAICが第2位のモデルや第3位のモデルを採択する理屈が存在しない。それと対照的に，このようにR_Dでは「良いモデル」を柔軟に採択することができる。

計算例2：管理者の勤務評定に関する研究

チャタジー・プライス（1980）「回帰分析の実際」新曜社より

- 大金融機関の事務員について，かれらの「管理者に対する満足度」などを調べたサーヴェイ。6個の説明変数。30の部局で得られた30個のデータがある。

Y ：管理者の行う仕事の全体的評価
X_1 ：被雇用者の不平不満の処理（管理者との人間関係）
X_2 ：えこひいきをしない（管理者との人間関係）
X_3 ：進取の気性に富む（管理者の仕事に関して）
X_4 ：管理者への昇進は仕事ができたからである（管理者の仕事）
X_5 ：まずい仕事に対する批判がきびしすぎる（人間関係）
X_6 ：仕事が段々おもしろくなる（自分の昇進についての考え）

- データは講義サイトに置いてある「`rating.txt`」

- **データ読み込み**

```
rating <- read.table("rating.txt", header=TRUE)
```

- **まずはAICで選択してみる。**

```
rating.lm <- lm(y~X1+X2+X3+X4+X5+X6, data= rating)
library(MASS)
stepAIC(rating.lm)
```

- **結果**

```
Step:  AIC=118
y ~ X1 + X3

       Df Sum of Sq    RSS    AIC
<none>              1254.7 118.00
- X3    1    114.73 1369.4 118.63
- X1    1   1370.91 2625.6 138.16
```

- 次にR_Dで分析

確かにX1+X3が最良だが, X1だけでも0.3%しか予測力は低下しない。

```
source("RDcompare.txt")
RDcompare(y~X1+X2+X3+X4+X5+X6, data= rating)
```

- 結果

X1だけのモデルを採用すべきだ。

```
# RD ranking for the hierarchical family of models #
           RD        RSD                                  Model
1   0.6464793 0.8751838                      y ~ 1 + X1 + X3
2   0.6431086 0.8706207                          y ~ 1 + X1
3   0.6409360 0.8676795                 y ~ 1 + X1 + X3 + X6
4   0.6305861 0.8536681                 y ~ 1 + X1 + X2 + X3
5   0.6240258 0.8447870                 y ~ 1 + X1 + X3 + X4
6   0.6237561 0.8444219                 y ~ 1 + X1 + X3 + X5
7   0.6229100 0.8432765                      y ~ 1 + X1 + X4
8   0.6220949 0.8421731                      y ~ 1 + X1 + X2
9   0.6218734 0.8418732            y ~ 1 + X1 + X2 + X3 + X6
10  0.6213290 0.8411361                      y ~ 1 + X1 + X6
11  0.6210653 0.8407791            y ~ 1 + X1 + X3 + X4 + X6
12  0.6203872 0.8398613                      y ~ 1 + X1 + X5
```

- **パラメーター推定値**

```
# Parameters of the best 5 model #
# Model 1 #
             Estimate Std. Error
(Intercept) 9.8708805  7.3214464
X1          0.6435176  0.1228436
X3          0.2111918  0.1393568

# Model 2 #
              Estimate Std. Error
(Intercept) 14.3763194 6.69067404
X1           0.7546098 0.09857435

# Model 3 #
              Estimate Std. Error
(Intercept) 13.5777411  7.9177491
X1           0.6227297  0.1240013
X3           0.3123870  0.1618413
X6          -0.1869508  0.1520322
```

「被雇用者の不平不満の処理」の係数は 0.75**である。**

計算例3：大気汚染研究における変数選択

McDonald and Schwing (1973) は「気候, 社会経済, および公害を表す変数に対して総死亡率がどのような関わりをもっているか」についての研究を発表した。その研究では次スライドに掲載されているような15個の独立変数が選ばれた。従属変数は, すべての原因による総死亡率を年齢に応じて調整したものである（1960年の60地域のデータ）。下記サイトにそのデータがある。これもチャタジー・プライスに出てくるデータの一つ。説明変数の組み合わせは2^{15}=32768とおりある。

https://www4.stat.ncsu.edu/~boos/var.select/pollution.html

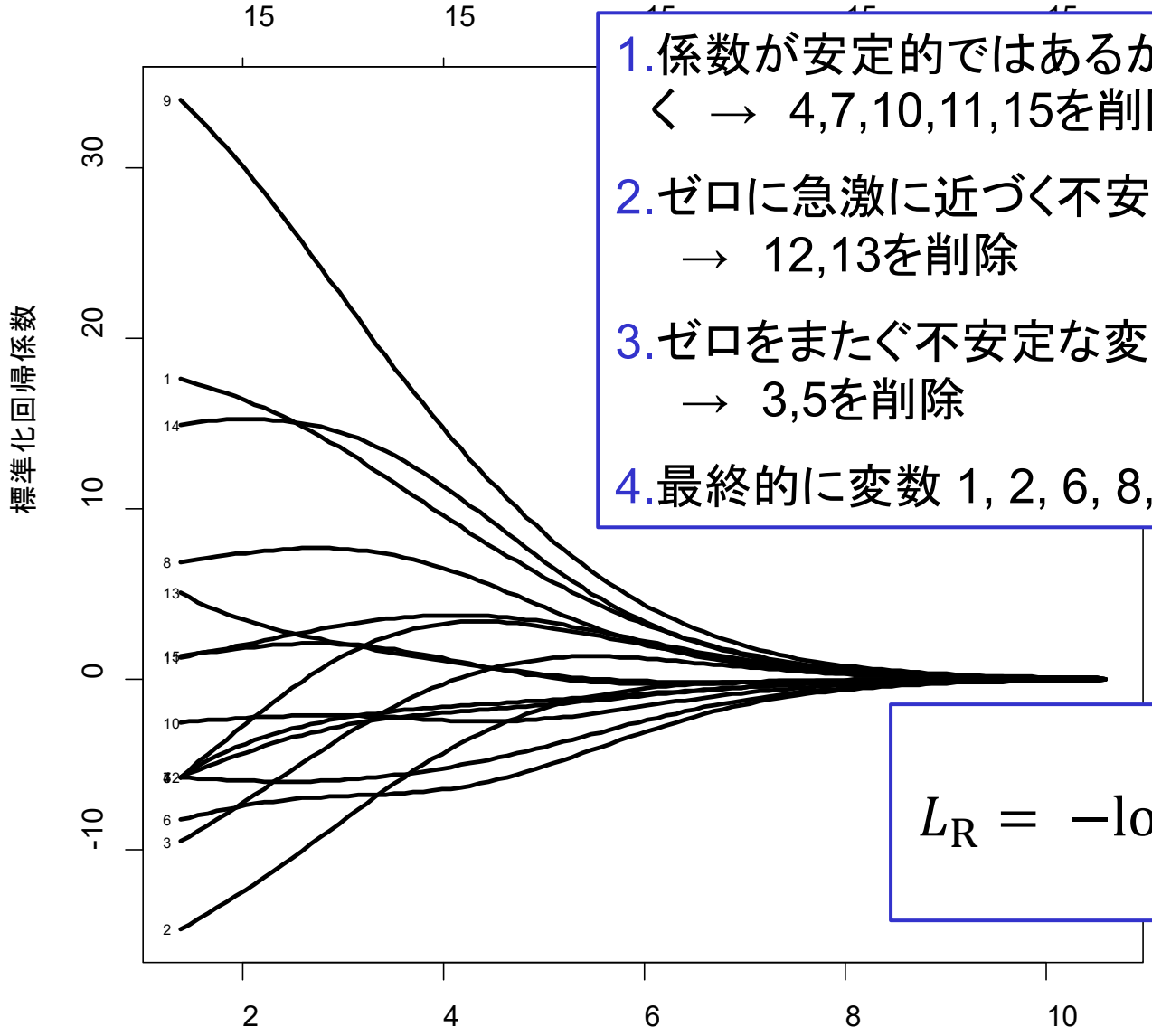
文献： McDonald GC, Schwing RC (1973) Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15:463-481

変数の意味と平均, SD（$n = 60$）

変数番号	意味	平均	標準偏差
1	年間平均降雨量（インチ単位）	37.37	9.98
2	1月の平均気温（カ氏(F)単位）	33.98	10.17
3	7月の平均気温（カ氏(F)単位）	74.58	4.76
4	65才以上の人口の割合（パーセント）	8.80	1.46
5	家計当りの人口	3.26	0.14
6	終了学業年数のメディアン	10.97	0.85
7	健康な家庭の割合（パーセソト）	80.92	5.15
8	一平方マイル当りの人口	3876.05	1454.10
9	非白人の割合（バーセント）	11.87	8.92
10	ホワイトカラーの従業員の割合（バーセント）	46.08	4.61
11	所得が3000ドル以下の世帯の割合（パーセソト）	14. 37	4.16
12	炭化水素の相対汚染度	37.85	91.98
13	窒素酸化物の相対汚染度	22.65	46.33
14	亜硫酸ガスの相対汚染度	53.77	63.39
15	相対湿度の割合（パーセント）	57.67	5.37

「リッジトレース」を用いた「職人わざ的」なモデル選択

McDonald and Schwing (1973)



1. 係数が安定的ではあるが，値が小さい変数は除く → 4,7,10,11,15を削除
2. ゼロに急激に近づく不安定な変数を除く
   → 12,13を削除
3. ゼロをまたぐ不安定な変数を除く
   → 3,5を削除
4. 最終的に変数 1, 2, 6, 8, 9, 14 を採用

$$L_{\mathrm{R}} = -\log(L) + \lambda \sum_{i=1}^{k} {\beta_i}^2$$

RDcompare 関数で計算 計算時間は97秒

```
pollution <-
read.table("https://www4.stat.ncsu.edu/~boos/var.select/pollution.data.txt", header=TRUE)
source("RDcompare.txt")
RDcompare(y ~ ., data=pollution, max.ranking= 20, ShowModel= 5)
```

```
# RD ranking for the hierarchical family of models #
          RD                                                  Model
1  0.6636175                  y ~ 1 + x1 + x2 + x3 + x6 + x9 + x14
2  0.6623753             y ~ 1 + x1 + x2 + x3 + x5 + x6 + x9 + x14
3  0.6602734             y ~ 1 + x1 + x2 + x3 + x6 + x8 + x9 + x14
4  0.6564475                       y ~ 1 + x1 + x2 + x6 + x9 + x14
5  0.6561624      y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x9 + x14
6  0.6560376                  y ~ 1 + x1 + x2 + x5 + x6 + x9 + x14
7  0.6551305      y ~ 1 + x1 + x2 + x3 + x5 + x6 + x8 + x9 + x14
8  0.6551299             y ~ 1 + x1 + x2 + x3 + x7 + x8 + x9 + x14
```

- モデル4でも0.7%ぐらいしか低下しないので，モデル4を採用。採用された変数は 1, 2, 6, 9, 14。
- McDonald and Schwing (1973) は，リッジ回帰をもとに変数 1, 2, 6, 8, 9, 14 を採用した。ほぼ同じモデルが採用されている。

RDcompareLasso 関数で計算 計算時間は2秒

lasso を活用して計算時間を圧倒的に短縮できる。

```
source("RDcompareLasso.txt")
library(grpreg)
RDcompareLasso(y ~ ., data = pollution, max.variable = 10,
max.ranking = 20, ShowModel = 5)
```

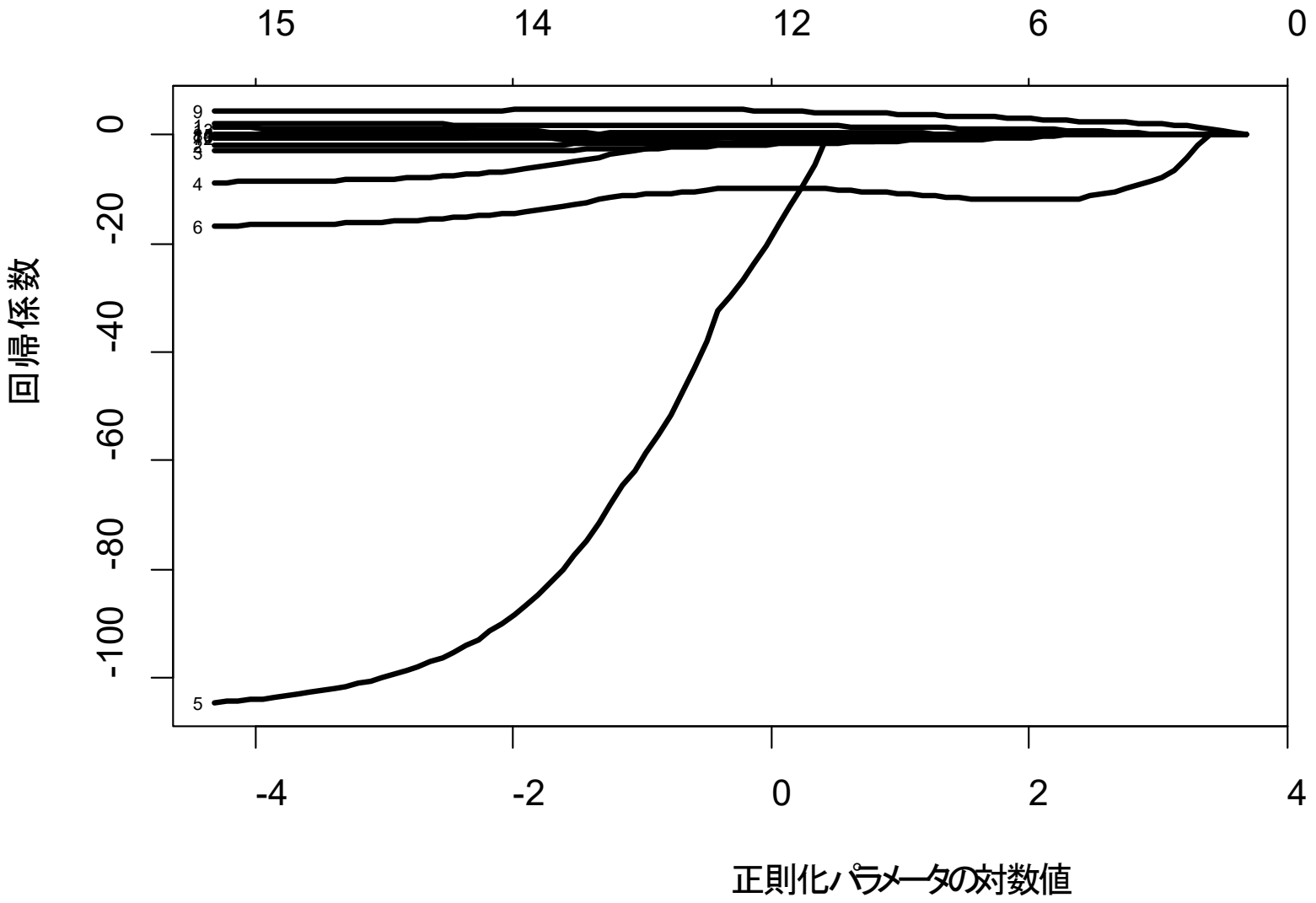
```
# RD ranking for the hierarchical family of models #
          RD                                                    Model
1  0.6636175              Lasso.Y ~ 1 + x9 + x6 + x14 + x1 + x2 + x3
2  0.6623753        Lasso.Y ~ 1 + x9 + x6 + x14 + x1 + x2 + x3 + x5
3  0.6602734        Lasso.Y ~ 1 + x9 + x6 + x14 + x1 + x2 + x8 + x3
4  0.6564475                 Lasso.Y ~ 1 + x9 + x6 + x14 + x1 + x2
5  0.6560376              Lasso.Y ~ 1 + x9 + x6 + x14 + x1 + x2 + x5
6  0.6551305  Lasso.Y ~ 1 + x9 + x6 + x14 + x1 + x2 + x8 + x3 + x5
7  0.6551299        Lasso.Y ~ 1 + x9 + x14 + x1 + x2 + x8 + x7 + x3
8  0.6535472        Lasso.Y ~ 1 + x9 + x6 + x14 + x1 + x2 + x7 + x3
```

- モデル4でも0.7%ぐらいしか低下しないので，モデル4を採用。採用された変数は 1, 2, 6, 9, 14。
- 選ばれたモデルは RDCompare 関数で選ばれたモデルと同じ。RDCompareLasso 関数ではモデルがいくつか省略されている。

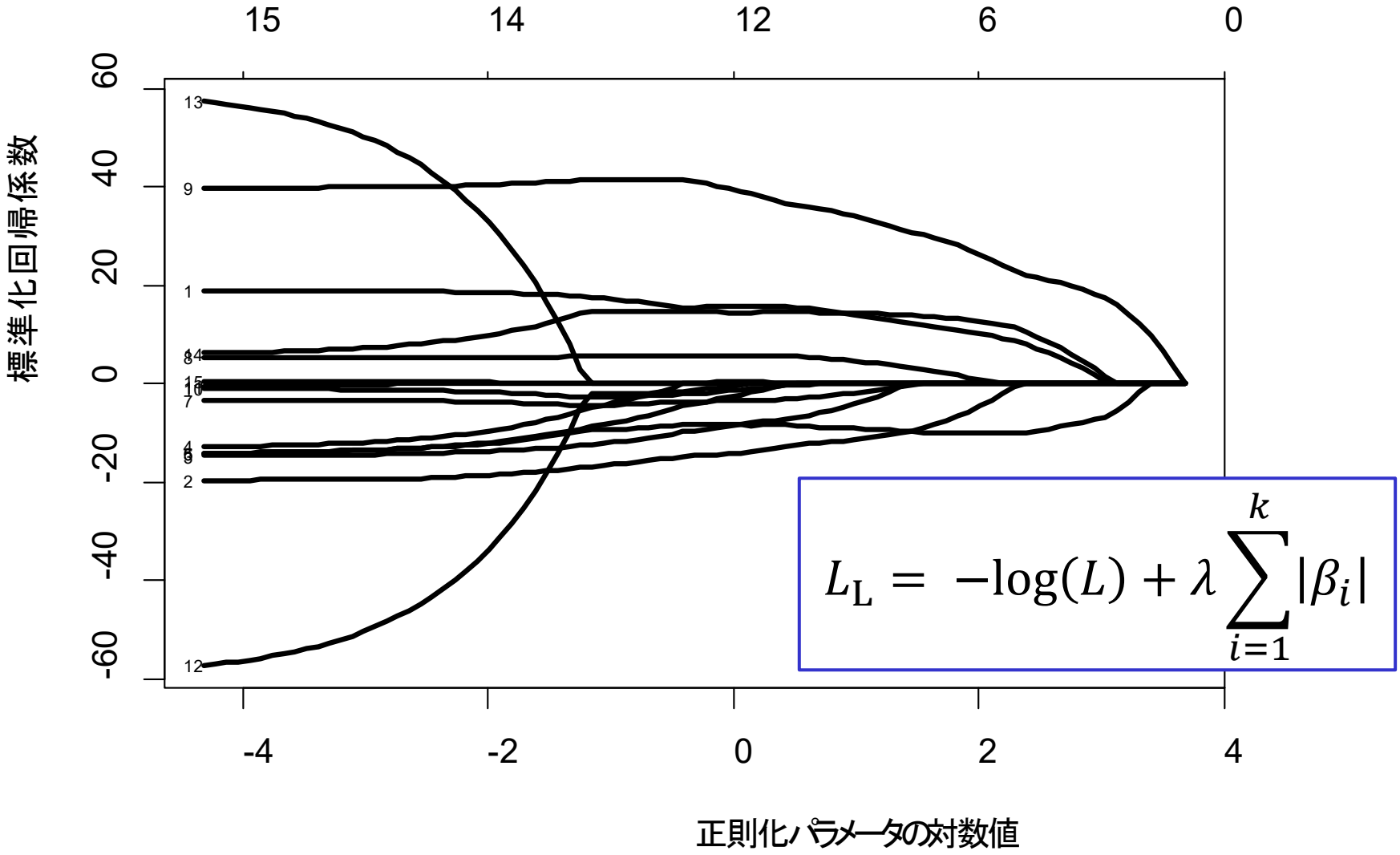
普通に lasso 選択を行った場合には（ glmnet を使用 ）

```
library("glmnet")
pollutionM <- as.matrix(pollution)
X <- pollutionM [, 1:15]  # 説明変数
y <- pollutionM [, 16]    # 目的変数
ycentered <- y - mean(y)   # 目的変数を中心化

# Lasso推定（標準化なし）
res <- glmnet(x=X, y=ycentered)
# 解パス図描画
windows(width = 8*0.8, height = 6*0.8)
plot(res, xvar="lambda", label=TRUE, xlab="正則化パラメータの対数値",
     ylab="回帰係数", col="black", lwd=2.5)

# Lasso推定（標準化あり：解パス図が見やすくなるが，パラメーター値は異なる）
Xstd <- scale(X)
resStd <- glmnet(x=Xstd, y=ycentered)
windows(width = 8*0.8, height = 6*0.8)
plot(resStd, xvar="lambda", label=TRUE, xlab="正則化パラメータの対数
値", ylab="標準化回帰係数", col="black", lwd=2.5)
```

解パス図 (もとのパラメーター)



解パス図（説明変数を標準化した場合）

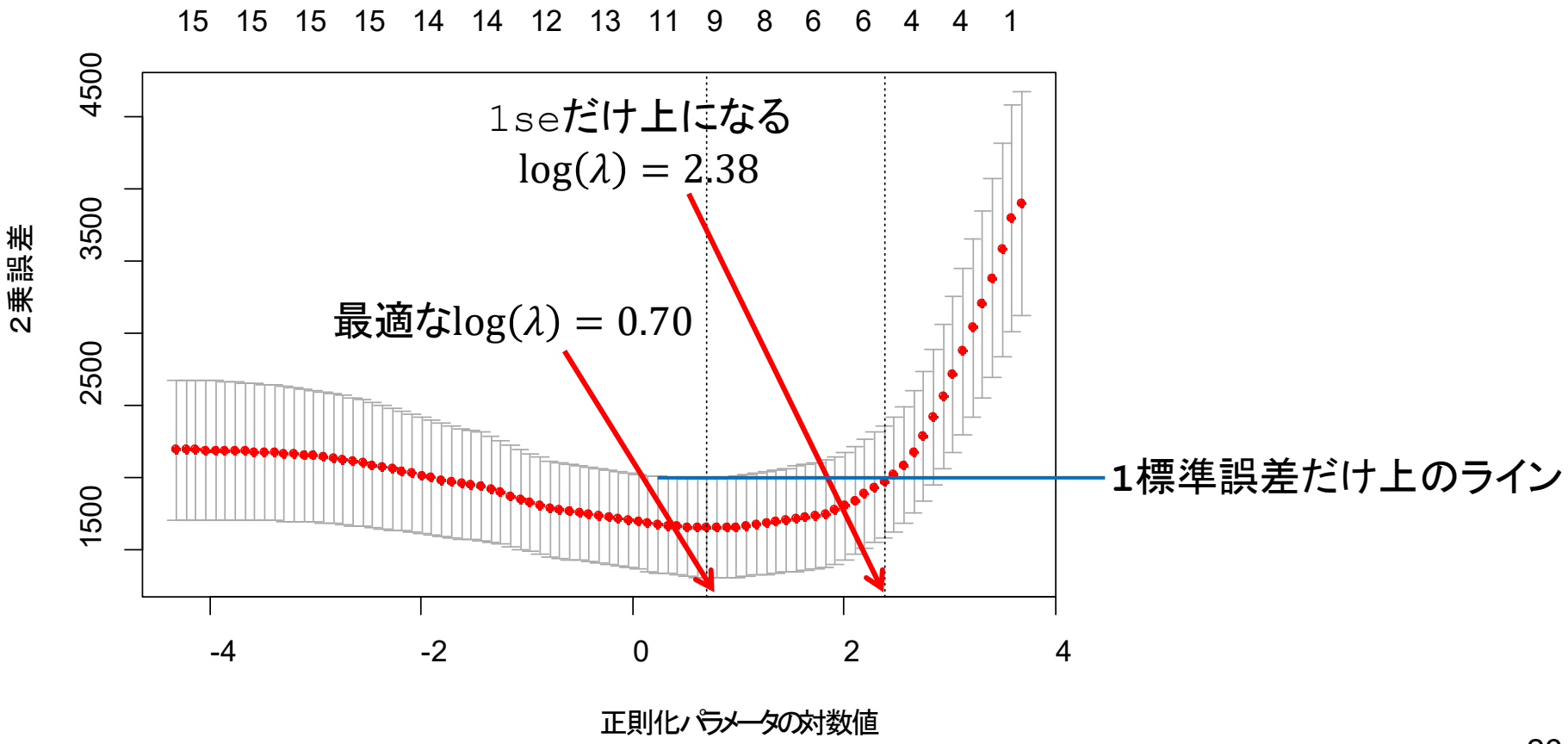


$$L_{\mathrm{L}} = -\log(L) + \lambda \sum_{i=1}^{k} |\beta_i|$$

パラメーターを入れる順番が決まるが，どこまで入れればよいかは不明

クロスバリデーションで最適なλを決定

```
# cvの計算
res.cv <- cv.glmnet(x=X, y= ycentered)
# cv値の推移をプロット
windows(width = 8*0.8, height = 6*0.8)
plot(res.cv, xlab="正則化パラメータの対数値", ylab="2乗誤差")
```



lasso 推定結果

CV値が最小となる正則化パラメータの値を用いた場合に選択されるモデル

```
resCVmin <- glmnet(x=X, y= ycentered, lambda=res.cv$lambda.min)
resCVmin$beta # 係数の推定値(1,2,3,6,7,8,9,10,14)。
```

⇒パラメーター数が多すぎる。

経験的な「1標準誤差ルール」により選択された正則化パラメータ用いた場合

```
res2 <- glmnet(x=X, y=ycentered, lambda=res.cv$lambda.1se)
res2$beta  #係数の推定値(1,2,6,9,14)
```

⇒選択されたパラメーターはR_Dと同じ。

```
     x1          x2          x6          x9         x14
0.9188579  -0.2137127  -12.0000483   2.7100360  0.1888903
```

一方，R_Dで推定された係数は下記のとおり。上の lasso 推定値では，係数の絶対値がかなり過小推定されていることがわかる。

```
     x1          x2          x6          x9         x14
1.4883049  -1.6229083  -12.7640647   4.0660834  0.2839027
```

もともと「正則化法をモデル選択に流用する」ということ自体に無理がある。正則化によって偏った推定値が出力される。「1標準誤差ルール」という経験的なルールを使わないと現実に合致しないことも問題。