データサイエンス講座
第4回 機械学習その3

- 重回帰分析　（AIC)
- 階層ベイズとMCMC

重回帰分析

- 重回帰分析　＝　基本的な考え方は単回帰分析と同じ
  - 単回帰分析:y = ax + b：説明変数は一つ
  - 重回帰分析:y = ax + bz + c：説明変数は複数

- 説明変数が複数の場合の問題点
  - すべての説明変数が“説明”できるとは限らない
  - 説明変数を増やした結果、説明できるモデルにならないケースが多い（オッカムの剃刀）

- どこまで“説明”できるかをチェックする必要がある　→　AIC（赤池情報基準）

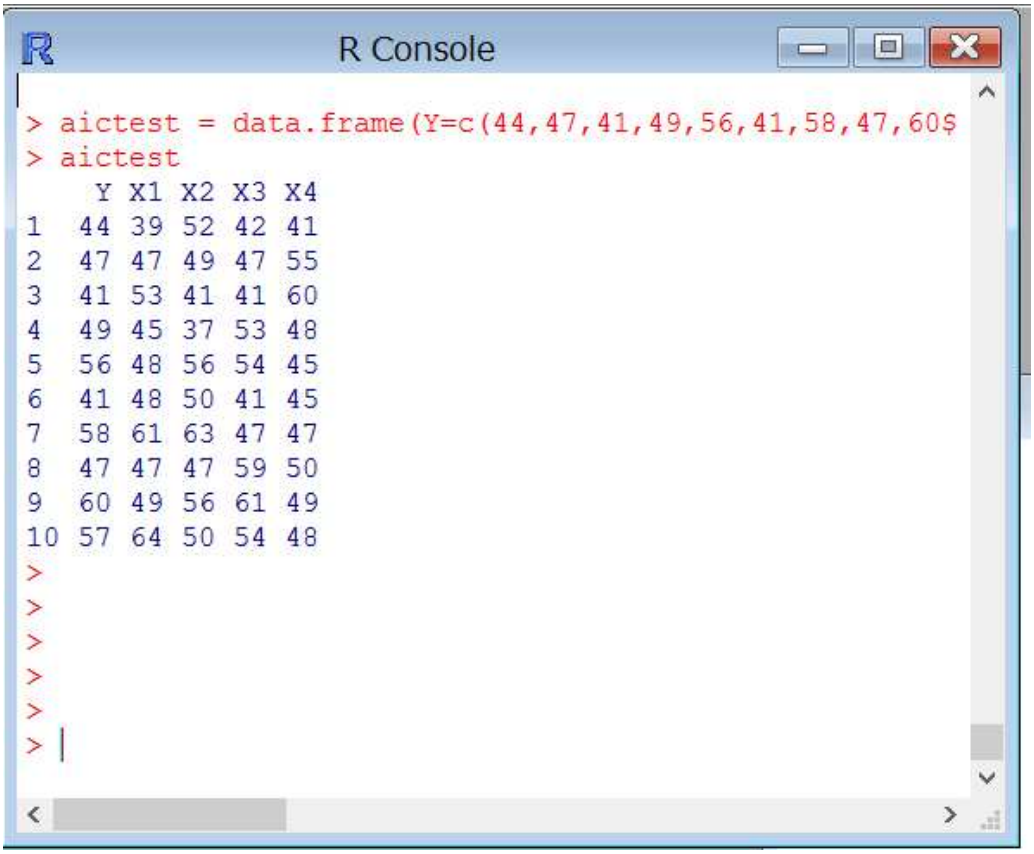
- AICの考え方

$$\mathrm{AIC} = -2\ln L + 2k$$

  - L: 最大尤度　モデルの確からしさ
  - k：パラメータの数
  - 各説明変数について、それぞれAICを計算、もっとも小さいAICが最適解

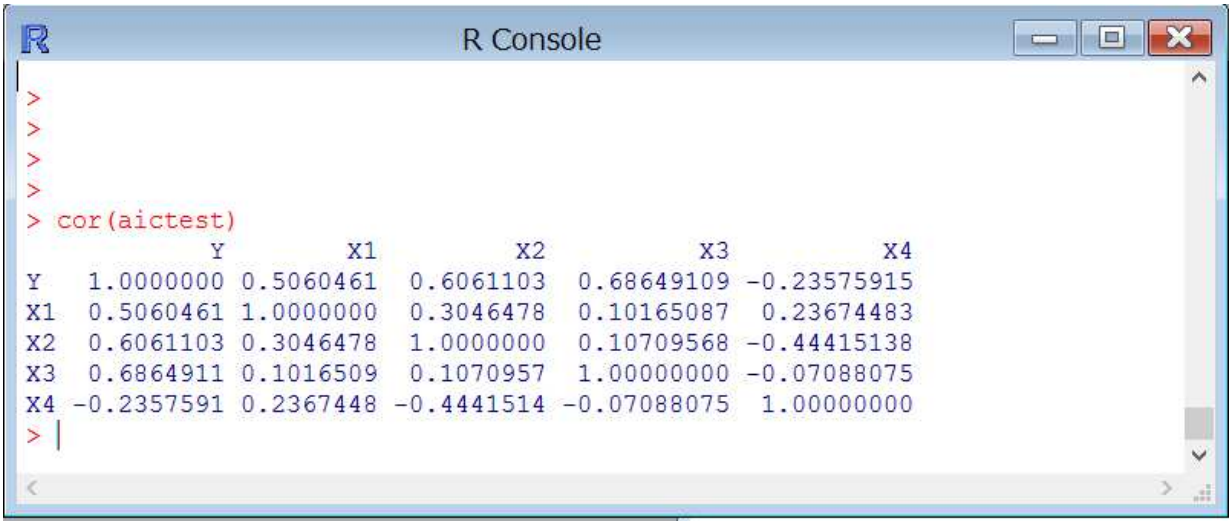
重回帰分析

- ４つの説明変数を作成(AIC.R参照)
  - aictest = data.frame(Y=c(44,47,41,49,56,41,58,47,60,57),X1=c(39,47,53,45,48,48,61,47,49,64),X2=c(52,49,41,37,56,50,63,47,56,50),X3=c(42,47,41,53,54,41,47,59,61,54),X4=c(41,55,60,48,45,45,47,50,49,48))
  - > aictest



重回帰分析

- 各説明変数の相関係数を表示
  - cor (aictest)



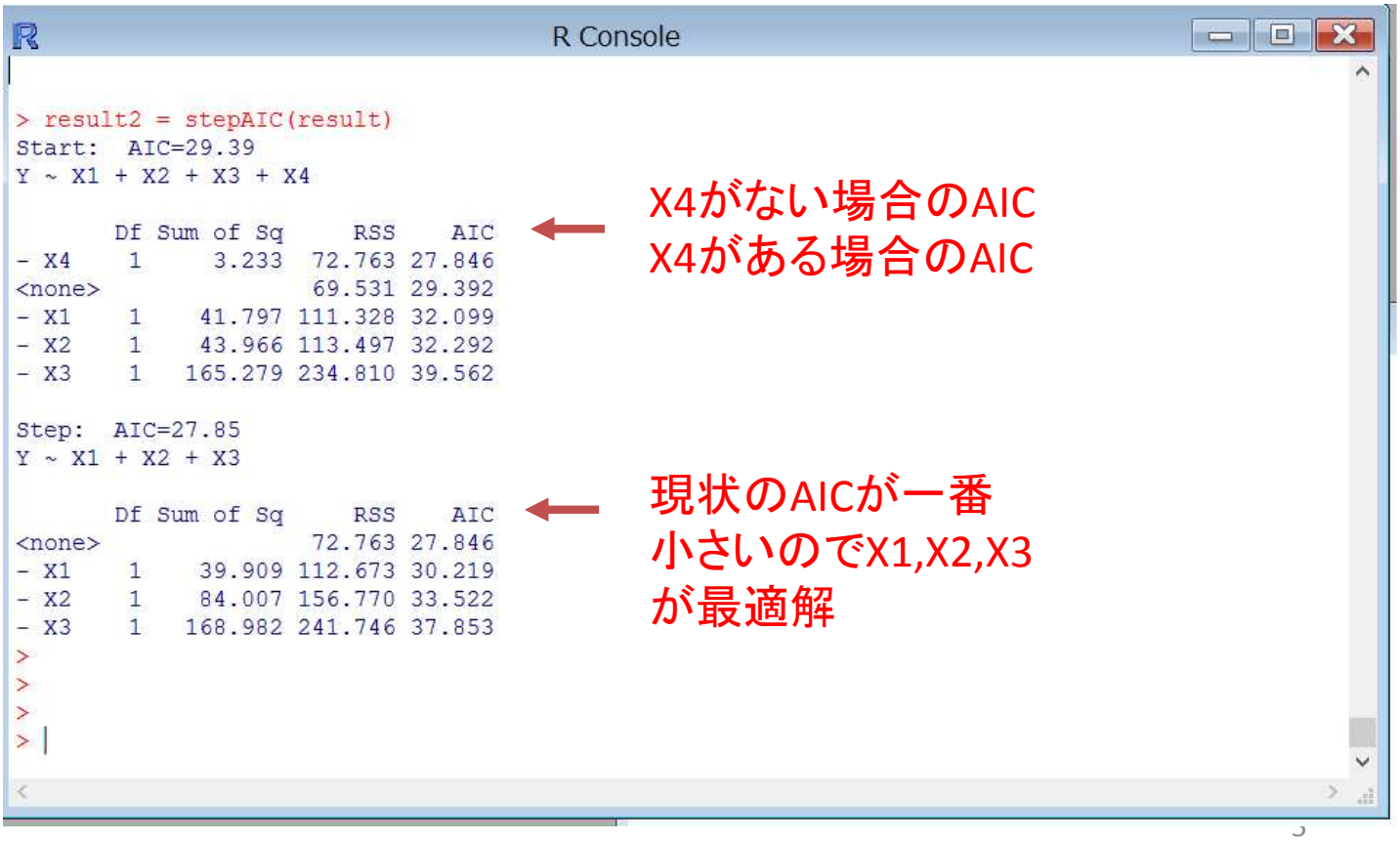
説明変数X4は、ほとんど相関に関係のない変数

重回帰分析

- MASSパッケージをインストール
  - install.packages(“MASS”)
  - library(MASS)

- X1～X4を説明変数として、重回帰分析
  - result = lm(Y ~ X1+X2+X3+X4, data=aictest)

- AICに基づく変数選択
  - result2 = stepAIC(result)



階層ベイズとMCMC

◘これまでの復習

統計手法	モデル	母集団の推定	推定計算方法
回帰分析	線形モデル	必要（ﾊﾟﾗﾒﾄﾘｯｸ）	最小二乗法
ロジスティクス回帰	一般化線形モデル	不要(ﾉﾝﾊﾟﾗﾒﾄﾘｯｸ)	最尤推定法
階層ベイズ推定	階層ベイズモデル	不要(ﾉﾝﾊﾟﾗﾒﾄﾘｯｸ)	MCMC

◘回帰分析

- 目的変数と説明変数が線形(y=ax+b)
- 基本、変数同士の相関が前提

◘ロジスティクス回帰

- 目的変数は、0～１の確率分布
- 最尤推定法で最も高い目的変数の確率を計算

階層ベイズとMCMC

- ロジスティクス回帰でできることとできないこと
- できること
  - いくつかの説明変数を投入して、その説明変数ごとのオッズを出す
  - チケットの購入確率（目的変数）
  - 説明変数：年収、性別

ランク	説明変数	P値 有意確率	オッズ比
1	40代_年収	0.1%	1.23
2	性別_男性	0.4%	1.12

- できないこと
  - 各エントリー（個人、クラスタ）に対してのオッズ、要するに各個人・クラスタがどの要素を重視しているか　→ one to one マーケティング

階層ベイズとMCMC

- 階層ベイズのざっくりとした考え方
  - 階層的にパラメータを設定することで、ターゲット（個人、クラスタ）にあわせてモデルを設定
  - 共通部分を全ユーザの情報を用いて推定
  - 基本的な考え方として、パラメータの確率分布を考える
  - パラメータの確率分布
    - MCMC　(Markov Chain Monte Carlo methods) によってパラメータの確率分布をもとめる
  - 一般的な統計の世界では、パラメータの確率分布を求めることはしない　（例： y=ax + b でaの確率分布を求めるなど）
  - ベイズ推定では、パラメータの確率分布を求めるのは自然な考え方
  - ベイズ推定とは？

階層ベイズとMCMC

ベイズ統計の考え方

- 例1：サイコロを振って次に6がでる確率は？
  - 1/6 確率は固定（確率分布も固定）

- でも、世の中の多くの問題は、確率が固定であることは少ない

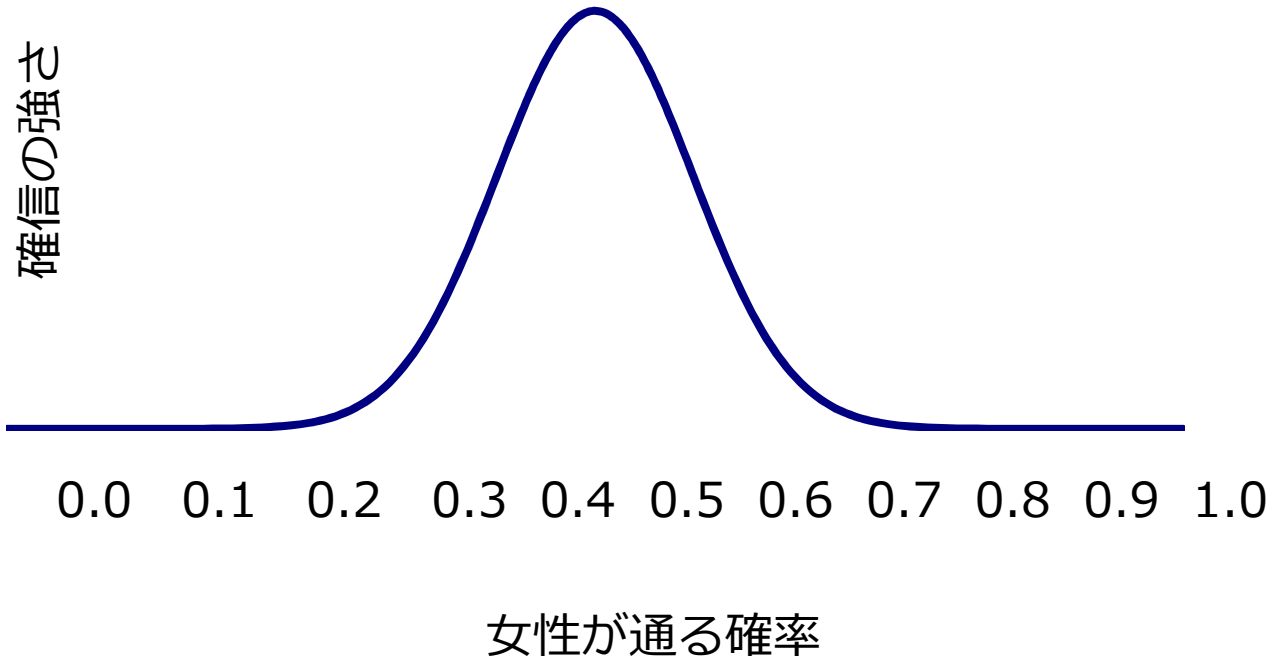
- 例2：交通量調査
  - 調査をはじめてから3人の通行人が通りました
  - 3人全員とも男性でした
  - 4人目の通行人が男性である確率と女性である確率、どちらが高いでしょうか？

  - 3人男性だったので、次も男性かもしれない
  - でも、サイコロの確率ほど、確信をもって男性であるとは断定できない
  - というわけで、男女の通る確率は不確定

  - 不確定・主観的という前提で確率を求めるアプローチ　→　ベイズ統計

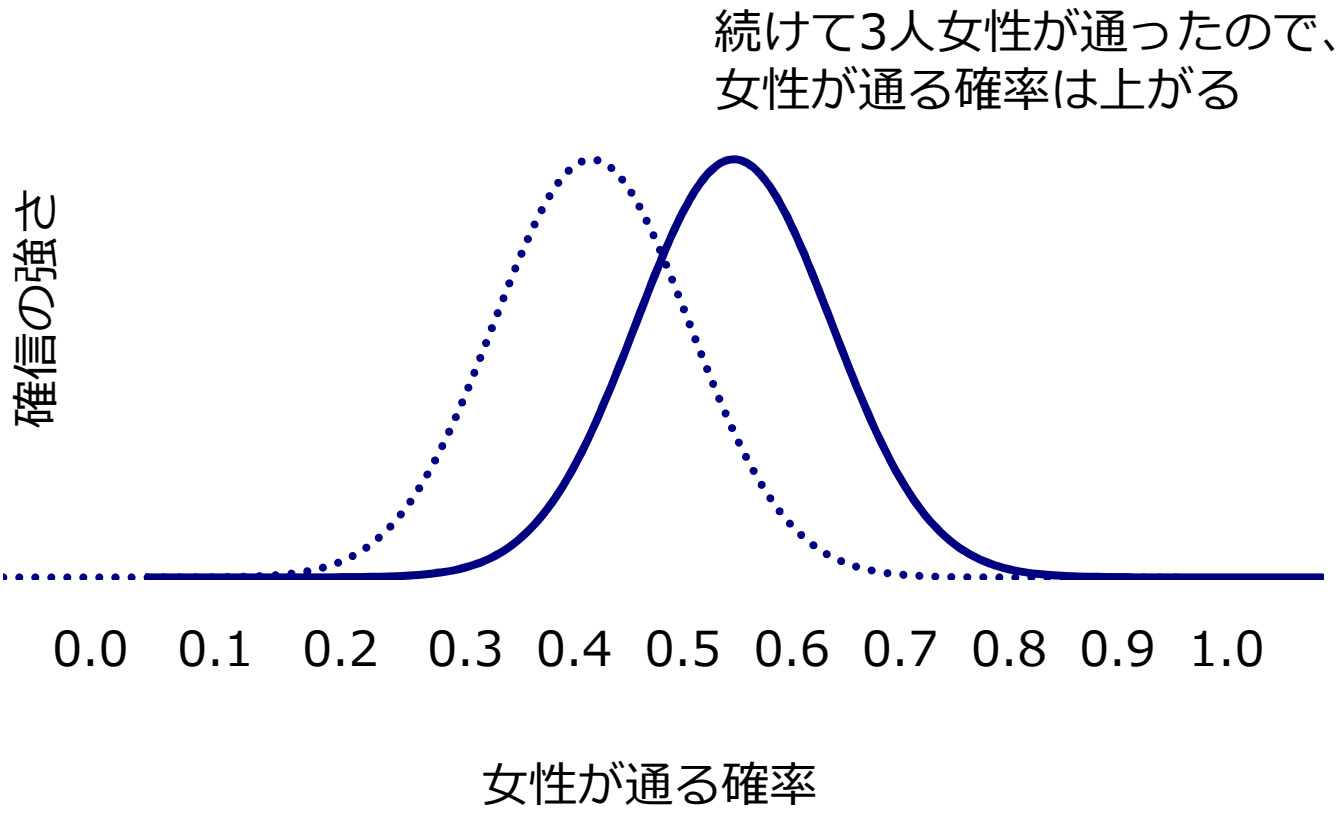
階層ベイズとMCMC

- ベイズ推定
  - 結局、確率は個人の主観でしかない
  - なので、個人個人の主観を表現する
    - Aさん：男女比　1：1　確率 0.5
    - Bさん：男女比　2：1　確率 0.33
    - Cさん：男女比　1：3　確率 0.75
    - …
  - この結果をグラフにすると、個人の主観を確率分布として表現できる



階層ベイズとMCMC

- ベイズ推定
  - 個人の主観による確率分布は、変更することができる



  - 観測するたびに確率を変更できる
    - 女性が通るたびに女性が通る確率が上がる
    - 確信の度合いも強くなる
  - ベイズ更新

階層ベイズとMCMC

- ベイズ推定
  - 「主観的にしか確率はわからない」という概念は、コンピュータ、とくに、ネットと相性が良い
  - 主観的なものを集めれば、集めるほど、真理に近づく、アンサンブル学習的な考え方
  - そして、ネット経由でたくさんの情報を集められるようになった
  - というわけで、Google,インテル、マイクロソフトなど、こぞってベイズ推定を自社プロダクト・サービスに活用
  - 「コンピュータ化や帯域幅の方向性を見極める時にベイズ理論研究を使う。不確実な状況においては全てを知ることは不可能だ。そんな時、確率論は全ての知能の基礎にあると個人的には考えている」マイクロソフト　シニアリサーチャー Eric Horovitz（*）

http://japan.cnet.com/news/special/20052855/

階層ベイズとMCMC

- ベイズ推定の応用例
  - 迷惑メール判定
  - ベイズ推定を分類問題に応用したナイーブベイズ分類の一つ
  - 迷惑メール/正常メールの区別は結局のところ主観的なもの
  - 人によっては正常メールであっても、ある人によっては迷惑メールの可能性もある
  - 各単語（無料、未公開情報など）に正常・迷惑の確率分布を計算
  - ベイズ更新で、確率分布を更新

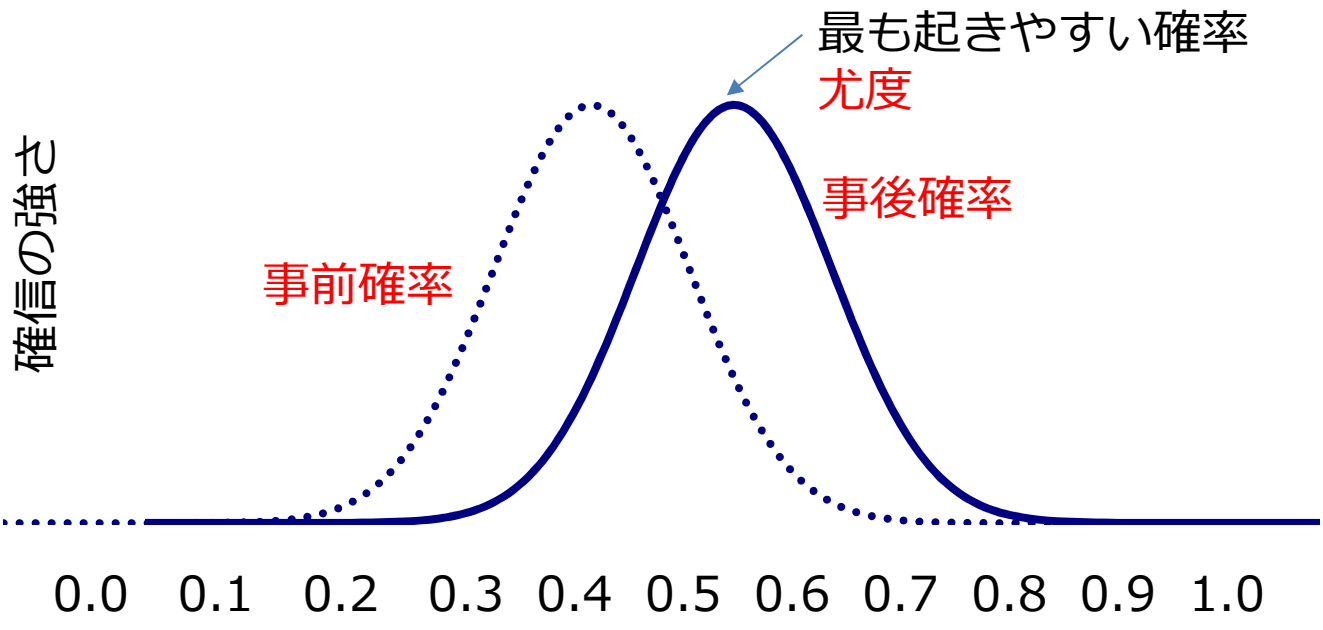
階層ベイズとMCMC

ベイズ推定

- どうやって確率分布を変化させるか
- ベイズの基本公式

$$p(q|Y) = \frac{p(Y|q) \times p(q)}{p(Y)}$$

- $p(q|Y)$：データ（Y)入力でパラメータqが得られる確率　（事後確率）
- $p(Y|q)$ ：パラメータqを決めることで最も起きやすい確率（尤度）
- $p(q)$：あるパラメータqが得られる確率(事前確率)
- $p(Y)$ ：　規格化定数



階層ベイズとMCMC

ベイズの公式

$$p(q|Y) = \frac{p(Y|q) \times p(q)}{p(Y)}$$

– 規定化定数を考慮しなければ、以下のようになる

$$p(q|Y) \propto p(Y|q) \times p(q)$$

事後確率　　尤度　　事前確率

– 決めたいもの

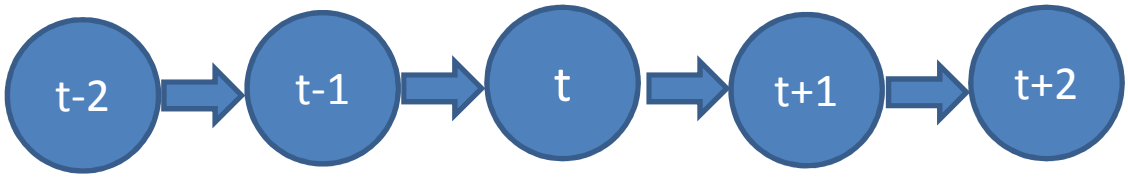
1. 事前確率
2. 尤度

階層ベイズとMCMC

- まず、事前確率をもとめたい
- どう求めるか？
  - ２つの方法
    1. 主観的な確率分布　－　一般的なベイズ推定（次に女性がくる確率）
    2. 階層事前確率分布　－　データによって確率分布が変わる
  - 階層ベイズでやりたいこと　－　ユーザー人一人のクラス分類
  - なので、事前確率を１．共通部分、２．個別部分とした階層事前確率分布にすれば、ユーザごとのグルーピングができそう

階層ベイズとMCMC

- 次に尤度$p(Y|q)$をもとめたい
- どう求めるか？
  - MCMC法の登場
  - MCMC　マルコフ連鎖モンテカルロ法
  - マルコフ連鎖
    - 1つ前の状態によって次の状態が決まる連鎖



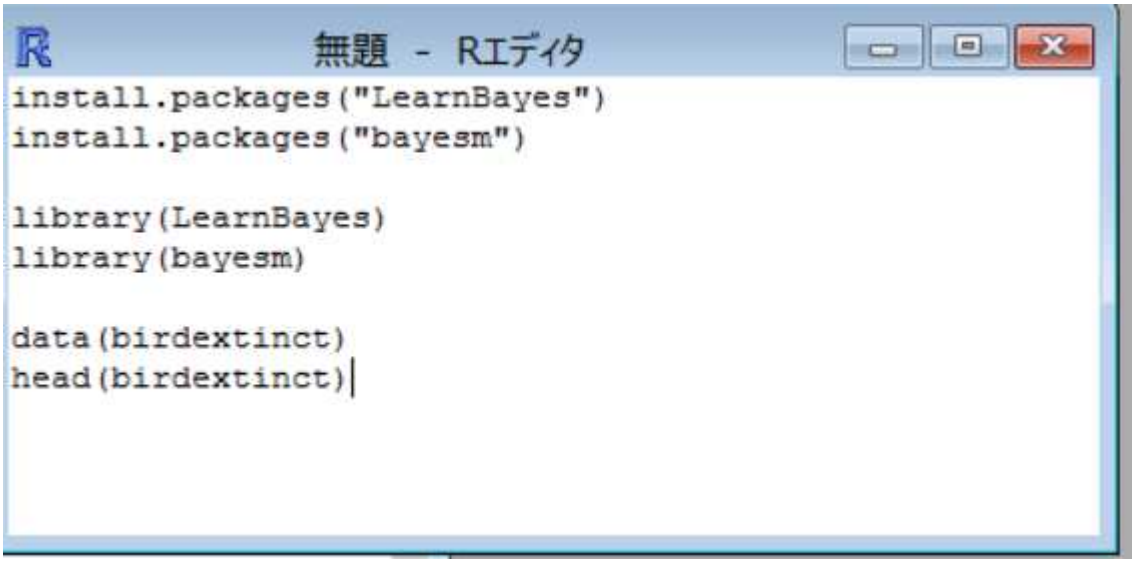
  - モンテカルロ法
    - 乱数を発生させるアプローチ
    - モンテカルロシミュレーション

- ざっくりMCMC法は、乱数を発生させて、試行錯誤で値を変えて尤度の確率分布を求めるアプローチ
- 結局のところ、y(事後確率) = q(尤度) × 事前確率なので、回帰のようなアプローチになる

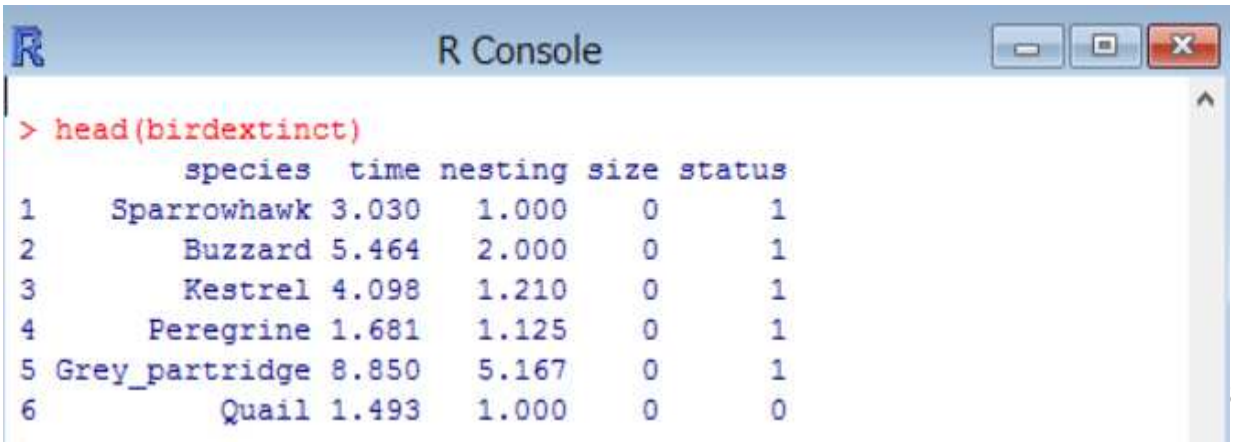
階層ベイズとMCMC

◘階層ベイズパッケージ

- LeanBayes bayesmをインストール

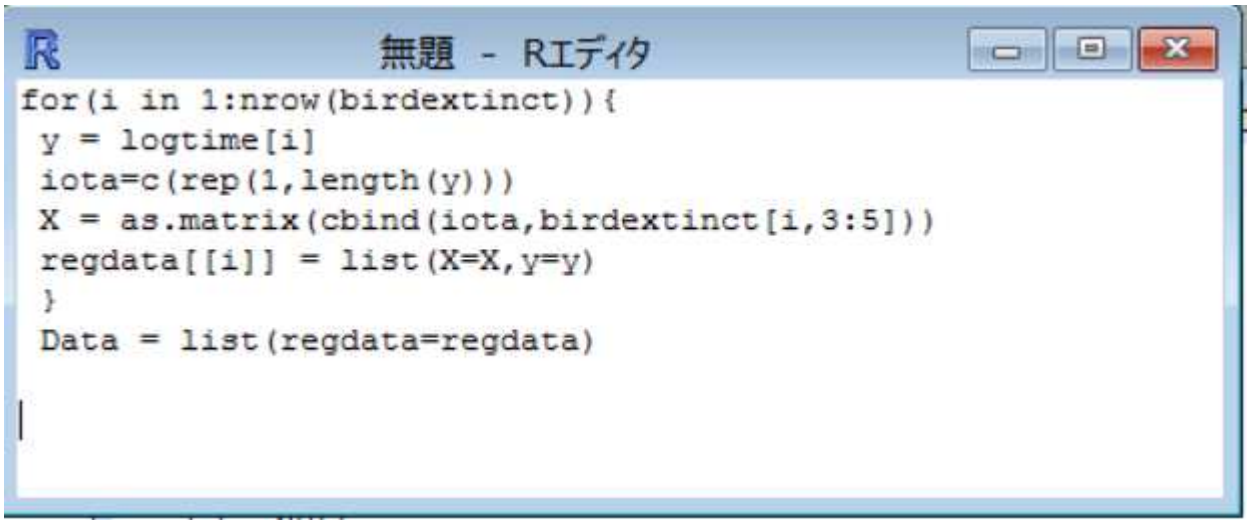


- birdextinct (鳥の絶滅種類）データ
  - Species: 鳥の種類
  - time：平均絶滅機関
  - nesting:つがいの平均数
  - status:わたり鳥かどうか

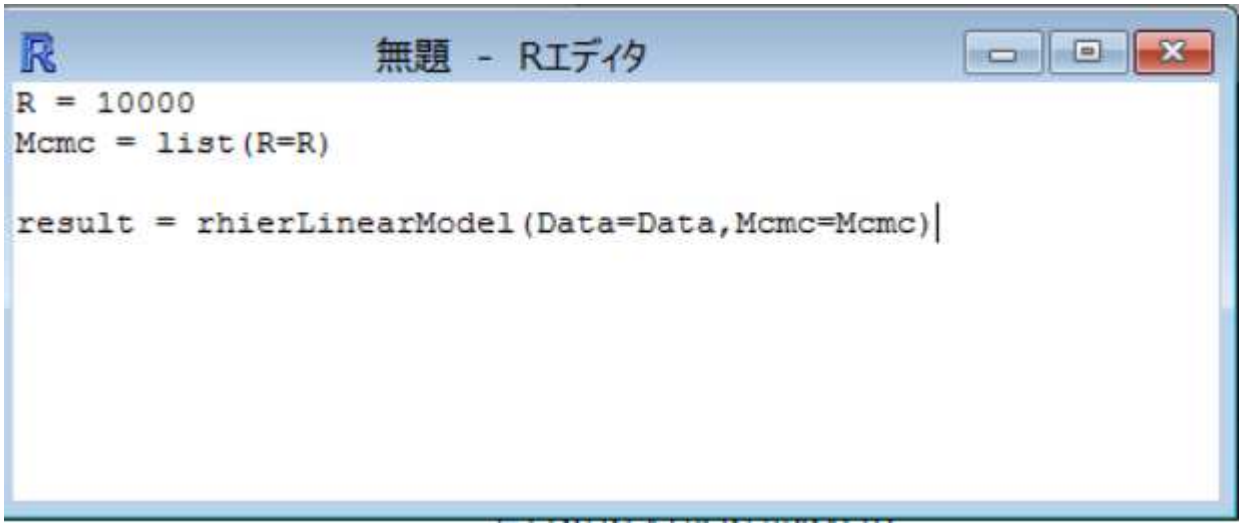


階層ベイズとMCMC

- データを用意する
  - birdextinctをDataに代入

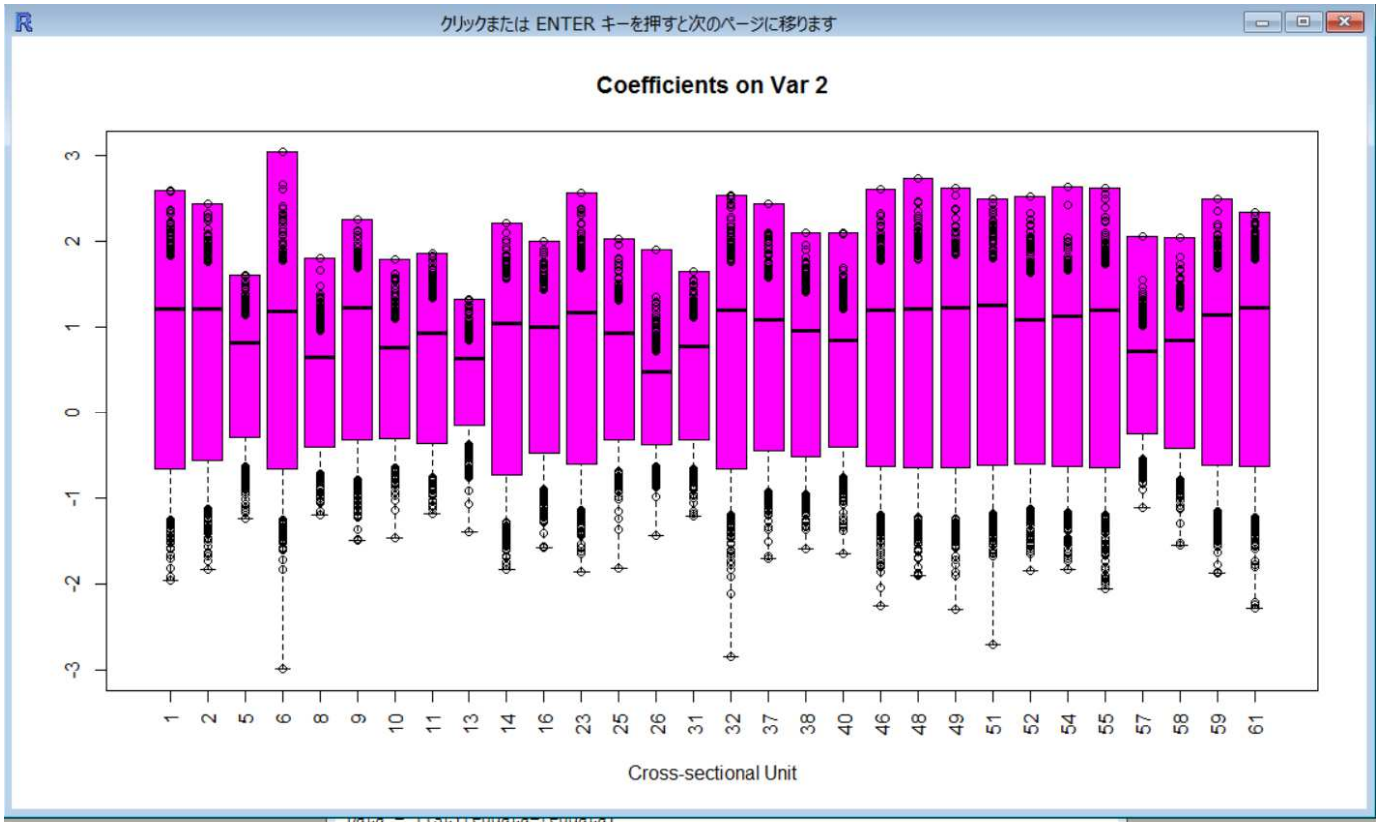


- MCMCによるパラメータ推定
  - R：MCMCの繰り返し推定数



階層ベイズとMCMC

- MCMCで得られr多、それぞれ個体の係数（尤度）を表示
- plot(result$betadraw)



階層ベイズとMCMC

- 個々の尤度データ (betadraw)について、集計し、平均を求める

無題 - Rエディタ

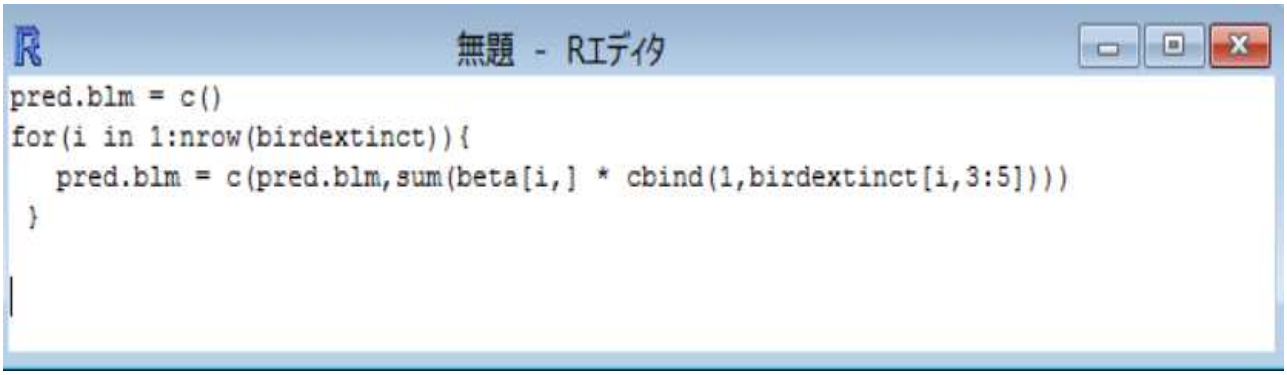
```
beta = data.frame()
for (i in 1:nrow(birdextinct)){
 tmp = rowMeans(result$betadraw[i,,seq(2000,10000)])
 beta = rbind(beta,tmp)
}
colnames(beta) = c("I",colnames(birdextinct)[3:5])
```

R Console

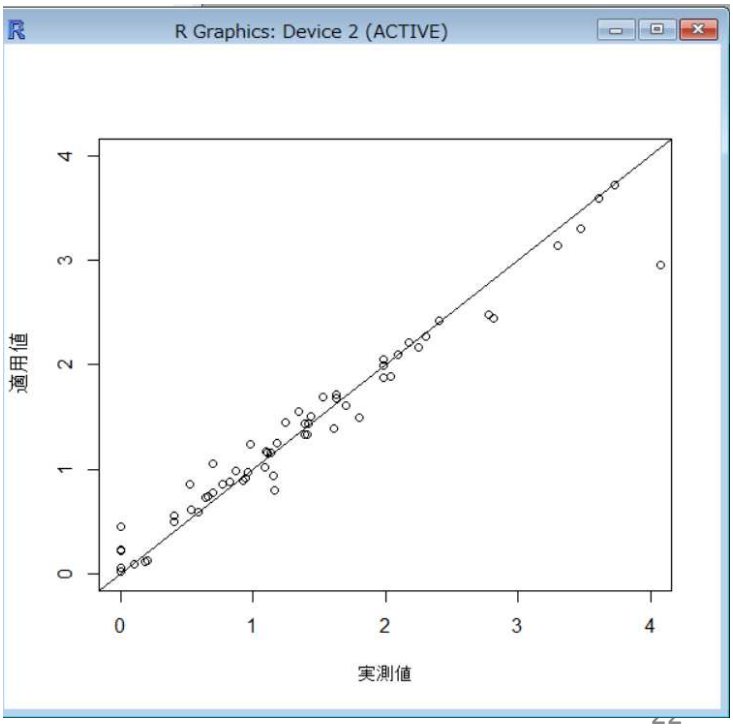
```
> 
> head(beta)
          I   nesting       size    status
1 0.4153440 0.2805352 -0.6754855 0.4578135
2 0.4643504 0.3153973 -0.7065854 0.5153436
3 0.4731368 0.2874442 -0.7010379 0.5151003
4 0.2850210 0.2565550 -0.6135515 0.2820504
5 0.4328852 0.2546172 -0.6611920 0.4704025
6 0.3067254 0.2518802 -0.6417374 0.5438449
> apply(beta,2,mean)
         I    nesting       size     status
 0.4307494  0.2794342 -0.6759149  0.4801149
> 
```

階層ベイズとMCMC

- 元データをモデルに適用（実測値とモデルの差異を算出）



- 実測値とモデルをプロット
  - plot(logtime,pred.blm,xlim=c(0,4), ylim=c(0,4),xlab="実測値",ylab="適用値")
  - par(new=T)
  - abline(0,1)

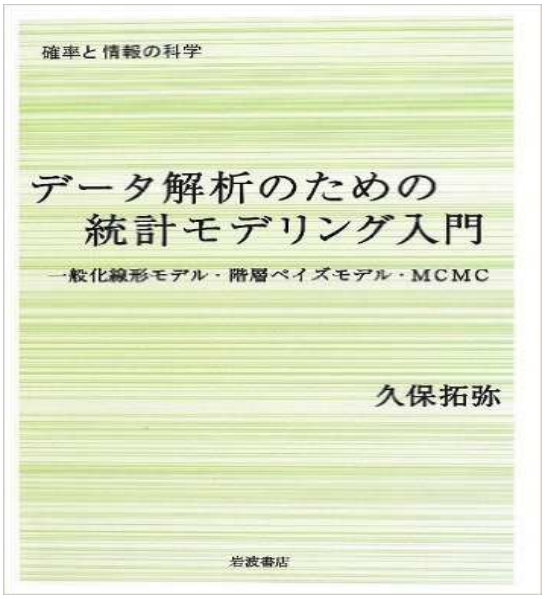


階層ベイズとMCMC

- 階層ベイズの運用
  - Rのパッケージのほかにも BUGSと呼ばれるツールあり
  - 開発も止まっており、GUIも使いにくいものの、製薬会社などでは結構利用されている。
  - 階層ベイズのウリは、個別のパラメータがわかること
  - ただし、演習のように60くらいであれば、判別できるものの、600, 6000、6万になると困難
  - 大手企業では、クラスタリングにわけて、そのクラスタリング内で階層ベイズをするアプローチをしているところもあり　→　様々な手法と組み合わせて使う方が一般的

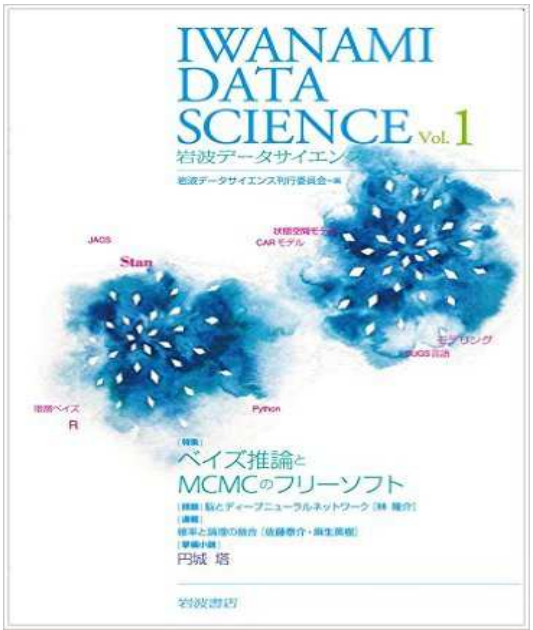
階層ベイズとMCMC

- さらに階層ベイズを勉強したい場合、



データ解析のための統計モデリング入門――一般化線形モデル・階層ベイズモデル・MCMC (確率と情報の科学) 単行本 – 2012/5/19
久保 拓弥 (著)

階層ベイズとMCMCについて、全般的にわかりやすく書かれておりおススメ



岩波データサイエンス Vol.1 単行本（ソフトカバー） – 2015/10/8
岩波データサイエンス刊行委員会 (編集)

いくつかの階層ベイズのトピックについて網羅的に触れられている

WinBUGの使い方（日本製薬協会）
http://www.jpma.or.jp/information/evaluation/allotment/winbugs.html