Bays 本輪読ゼミ　第４章　資料

Chap. 4　Multiparameter Models　－多変量モデル－

2008/1/10(Thr.)　倉本惠生＠森林総研北海道支所
shkura@affrc.go.jp

4.1　この章でやること　（私が思うに・・・）

A）パラメトリックでも分布がパラメータ記述できない状況*とか

B）ノンパラメトリックな状況で、R を使ってベイズ推定してみよう

→こんなとき

A)正規分布で平均と分散の両方が未知な状況（4.2）

B)多項モデル(4.3)

ロジスティック回帰（生物検定）(4.4)

２つの比率の比較（分割表）(4.5)

4.2　（平均と分散が未知の）正規分布をベイズ推定！

例．マラソンの完走タイム

20 人の走者のタイムを計ってデータをとる

（ベイズ推定でなければこのサンプルから平均と分散をとっておしまい）　$N(\mu,\sigma)$

（ベイズ推定だったら）

無情報の事前分布 $g(\mu,\sigma^2) \propto 1/\sigma^2$

→平均と分散の事後密度は文中式で与えられる

n ：サンプル数

$\bar{y}$：サンプル平均

$S=\Sigma^{n}_{i=1}(y_i-\bar{y})^2$

この結合事後密度（平均と分散が一緒になった密度分布ってことか）は、

正規（平均のとる分布？）／逆カイ自乗（分散のとる分布？）の結合した格好になる

・分散 σ^2 のもとでとる平均 μ の事後密度は $N(\bar{y},\sigma/\sqrt{n})$

・分散の周辺事後分布は $S\chi^{-2}_{n-1}$　χ^{-2}_{ν}：逆カイ自乗分布　自由度 ν

［式を書く］

```
library(LearnBayes)
 data(marathontimes)
 d = mycontour(normchi2post, c(220, 330, 500, 9000), time)
 title(xlab="mean", ylab="variance")
```

R 関数 normchi2post　(μ,σ^2) の結合事後密度の対数を計算する
mycontour　contour コマンドを使うためのもの
文法 mycontour(A,B,C)
　用例　mycontour(normchi2post, c(220, 330, 500, 9000), time)
 ３つを定義してやる A　対数密度を定義する関数の名前
　　B　密度を図示する際の範囲　c （x 下限， x 上限，y 下限，y 上限）
　　C　対数密度関数で用いたデータ

平均と分散の結合事後分布は、まず分散を逆カイ自乗分布からシミュレートし、次に平均をサンプルの正規分布からシミュレートして求める。→関数 rchisq を使ってカイ自乗分布をもとに分散の 1000 回のシミュレーション。次に、分散のシミュレート（点）を逆カイ自乗分布に変換する。最後に、平均のシミュレート点を関数 rnorm で求める。

```
#結合事後分布を求める
S = sum((time – mean(time))^2)
n = length(time)
sigma2 = S/rchisq(1000, n-1)
mu = rnorm(1000, mean=mean(time), sd= sqrt(sigma2)/sqrt(n))

#シミュレート結果の点をコンタープロットの上に表示する
points (mu, sigma2)

#区間推定　quantile を使う。
quantile(mu, c(0.025, 0.975)) #平均の 95%信頼区間
quantile(sqrt(sigma2), c(0.025, 0.975) #標準偏差（分散の平方根）の 95%信頼区間
```

4.3 多項モデル

内容はノンパラメトリックをベイズでというとこか？

多項モデルとは　・テキスト中の例に補足しながら説明します

意思決定問題の対象が２つ（２ブランドや Yes or No）の場合、二項モデルで消費者や有権者の意思を理解することができるが、対象が３つ以上の場合に複数の二項モデルを適用しても全体を統一的に正しく理解することができない。そのとき使うのが多項モデル。

テキストの例のように候補者が３名以上存在している場合、他の候補者を考慮せずに候補者１と２の支持数の差（ここでは Bush が Dukakis に差をつけているかが焦点）をサンプル調査から比較しても有権者の意向を正しく認識できているとはいえません。有権者が支持候補に投票するかどうかは、支持候補者と他の複数候補を同時に考慮して、支持候補者が最善か否かを判断しているからです。ほかにも、どの業者から原材料を調達するべきか？どの場所に店舗を新規出店するべきか？などなど、ほとんどの意思決定問題の対象は２つではなく、3 つ以上存在している。

○多項モデルの数学的記述はテキスト参照

［式を書く］

もとになる分布は Dirichlet 分布。R のパッケージでは Drichlet 分布を直接シミュレートできないが、ガンマ分布と Drichlet 分布の間に成り立つ関係（*）を利用して Drichlet 分布のランダムシミュレート点集合を求められる→rdirichlet 関数。

*［式を書く］

alpha= c(728, 584, 138)

theta= rdirichlet(1000, alpha)

問題は Bush が Dukakis をリードしているかなので、θ1－θ2 の分布が最終的に必要です。

hist(theta[, 1] – theta[,2], main="")

4.4 生物検定

（ロジスティック回帰分析の話だと理解しましたが・・・）

ロジスティック回帰分析とは？
疾患のリスクファクターや薬の投与量を分析するためによく用いられる多変量解析手法。
主として医学分野で用いられる
→数学的には説明変数 x が計量尺度のデータで、目的変数 y が名義尺度を計量尺度化したデータ（頻度だとか比率がそうです）の重回帰分析に相当します。 単純に回帰してしまうと、回帰直線は原理的には有り得ない出現率 0 以下の領域と出現率 1 以上の領域まで入りこんでしまいます*。つまり、予測値が 0 より小さくなったり、1 より大きくなったりするという、ありえないことがおきます。 そこで説明変数と出現率の関係を直線で回帰せず、出現率 0 から 1 の間で変化する曲線（ロジスティック曲線）で回帰します。
*もうひとつ問題があって、それぞれの頻度や比率を出すのに使ったサンプル数が考慮されない（→最尤推定で）

例　ある薬の投与量と死亡発生の試験データ（表４．１）
ｙｉ：ある投与レベル Xi におけるサンプル数 ni 中の死亡数　＃二項分布をとる
　log（ｐｉ／（１－ｐ ｉ））＝β0＋β１ xi
未知の回帰パラメータ β0 と β１の尤度関数は
　L(β０, β１)＝･･･

P

実際に非ベイズ的手法とベイズと両方でやってみる
まずはデータを用意する

```
 x = c(-0.86, -0.3, -0.05, 0.73) #
 n = c(5, 5, 5, 5)
 y = c(0, 1, 3, 5)
data = cbind( x, n, y)
```

1)最初に非ベイズ的やり方で→glm 関数を使う

```
  response = cbind(y, n – y)　#死亡数を目的変数にとっているわけです
  results = glm(response ~ x, family = binomial)
  summary(results)
```

2)ベイズでやってみよう

1.まず事後密度の範囲（長方形グリッド）を決める→logsticpost 関数を使います

```
mycontour(logsticpost, c(-4, 8, -5, 39), data)
title(xlab="beta0", ylab="beta1")
```

○範囲の決め方がいまいち分からないが・・・

最頻値の１％になるように？

2.次に範囲内での事後密度からβ０とβ１の組合せをシミュレートして点に落とす

→simcontour

```
s = simcontour(logsticpost, c(-4, 8, -5, 39), data, 1000)
points(s$x, s$y)
```

3.各パラメータの事後分布（ここでは傾きβ１）をみる

```
plot(density(s$y),xlab="beta1")
```

4.LD-50（薬の世界ではよく使う基準：半数死亡投与レベル）をもとめる

θ＝－β０／β１

個々の（β０，β１）の組合せからθを計算することでその周辺事後密度を求める

```
theta=-s$x/s$y
hist(theta, xlab="LD-50")
```

区間推定はこう

```
quantile(theta, c(.025,.975)) #95%信頼区間
```

4.5 ２つの割合を比較する

２つの二項分布からなる２つの割合を比較する

数学的に記述すると

y1：二項分布（n1,p1）

y2：二項分布（n2,p2）

仮説１：p1＞p2、仮説２：p1<p2

従属事前分布から検証する

```
sigma=c(2,1,.5,25)
plo=.0001;phi=.9999
par(mfrow=c(2,2))
for (i in 1:4)
+ {
+ mycontour(howardprior, c(plo,phi,plo,phi),c(1,1,1,1,sigma[i]))
+ title(main=paste("sigma=",as.character(sigma[i])),
+ xlab="p1",ylab="p2")
+ }

sigma=c(2,1,.5,.25)
par(mfrow=c(2,2))
for (i in 1:4)
+ {
+ mycontour(howardprior, c(plo,phi,plo,phi),
+ c(1+3,1+15,1+7,1+5,sigma[i]))
+ lines(c(0,1),c(0,1))
+ title(main=paste("sigma=",as.character(sigma[i])),
+ xlab="p1",ylab="p2")
+ }

 S= simcontour(howardprior, c(plo,phi,plo,phi),
+ c(1+3,1+15,1+7,1+5,sigma[i]))
Sum(s$x>s$y)/1000
```