Rで学ぶベイズ統計学入門

第8章　モデル比較

８．５　二つのモデルを比較する
８．６　サッカーのゴールモデル
８．７　野球の打者は本当に調子の波があるのか？
８．８　2元分割表の独立性の検定

茨城大学情報工学科
倉持　辰洋

二つのモデルの比較（1）

- ベイズファクターによる仮説を比較する方法は、二つのモデルを比較するものに一般化できる
- y：データのベクトル

  θ：パラメータ　　とすると

  ベイズモデルは

  f(y | θ)：サンプリング密度

  g(θ)：事前密度　　を指定し構成される

二つのモデルの比較（2）

- 二つのベイズモデルを比較する

  モデル M_0：$y \sim f_1(y|\theta_0), \theta_0 \sim g_1(\theta_o)$

  モデル M_1：$y \sim f_2(y|\theta_1), \theta_1 \sim g_2(\theta_1)$

  ただし、θの定義はモデル間で異なることもありえる

- この時、M_0を支持するベイズファクターを求めたい

  ⇒BFは二つのモデルそれぞれのデータの周辺密度、または事前予測密度の比になる

二つのモデルの比較(3)

- モデルのデータの周辺密度、事前密度は次の式で求まる

$$m(y)=\int f(y|\theta)g(\theta)d\theta$$

- これよりベイズファクターは次の式で表される

$$BF=\frac{m_0(y)}{m_1(y)}$$

また事後確立はBFと二つのモデルのそれぞれの事前確立πより

$$P(M_0|y)=\frac{\pi_0 BF}{\pi_0 BF+\pi_1}$$

ラプラス法によるm(y)の近似（１）

- 周辺密度を近似する簡単な方法で５．３節のラプラス法がある。
- まず、事前予測密度は次の式で近似できる

$$m(y) \simeq (2\pi)^{\frac{d}{2}} g(\hat{\theta}) f(y|\hat{\theta}) |-H(\hat{\theta})|^{\frac{1}{2}}$$

$\hat{\theta}$ ：事後モード

$H(\theta)$：対数事後密度のヘシアン（行列の２階微分）

d ：パラメータの数

ラプラス法によるm(y)の近似（2）

- 対数尺度では次のようになる

$$\log m(y) \simeq \left(\frac{d}{2}\right)\log(2\pi) + \log\left(g(\hat{\theta})\, f(y|\hat{\theta})\right) + \left(\frac{1}{2}\right)\log\left|-H(\hat{\theta})\right|$$

- 積f（y|θ)g（θ）の対数を計算する関数を作成

  →laplace()関数が適用される

  出力intはlog m(y)の推定値となる

これよりm(y)が計算できBFを求められる

２元分割表の独立性の検定（１）

- 二つのカテゴリカルデータがあるとする

  それぞれの測定値に関連性があるかどうかを探りたい、そのための方法を二つ紹介する

  ⇒１．ピアソンのカイ二条統計量によるもの

  独立モデルの元での実測値ときたい数の差をはかる

  ２．ベイズ法によるもの

  独立であると仮定するモデル、依存関係があるとするモデル。二つのモデルから求める

２元分割表の独立性の検定（２）

- 具体例としてMoore(1995）に取り上げられている事例を検討する
- 課外活動に費やした時間と講義での成績との関連性について検討する

	課外活動（週当たりの時間）		
	< 2	2から12	> 12
C以上	11	68	3
DかF	9	23	5

Rでピアソンのカイ二乗等軽量を用いる

- 頻度表を読み込みchisq.test()関数を使って独立性の検定を行うことが出来る

```
> data <- matrix(c(11,9,68,23,3,5), c(2,3))
> data
     [,1] [,2] [,3]
[1,]   11   68    3
[2,]    9   23    5
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 6.9264, df = 2, p-value = 0.03133

 警告メッセージ:
In chisq.test(data) :  カイ自乗近似は不正確かもしれま
```

- 左が実行例となる

  ここでp値はほぼ0.03であり、これは成績が課外活動に費やした時間と関係を表す証拠となっている。

ベイズ法による
２元分割表の独立性の検定（１）

- 二つのモデルを仮定する。

  二つのカテゴリカル変数が独立と仮定：モデル　M_I

  依存関係があると仮定：モデル　M_D

- ベイズモデルの説明のためこれらのモデルのデータは対象の母集団からのランダムな標本
- 依存モデルでは多項割合がそれぞれ一様分布に従う。
- 独立モデルでは多項割合の構造は独立で、周辺割合に独立した一様事前分布が割り当てられる

依存モデルM_D

- 分割表の頻度は以下の表の割合の値による多項分布に従うとする
- 割合の値$p_{11},\ldots,p_{23}$はいかなる値もとりうる

  ただし、すべてのpを合計すると1となる

	課外活動（週あたりの時間）		
	< 2	2から12	> 12
C以上	p_{11}	p_{12}	p_{13}
D以上	p_{21}	p_{22}	p_{23}

独立モデルM_I

- 分割表の割合は以下の表のように、
  周辺確立　$\{p_{1+}, p_{2+}\}$, $\{p_{+1}, p_{+2}, p_{+3}\}$
  で定まる。
- 二つの割合集合は独立と仮定、それぞれのすべての値に一様な密度を割り当てる

	課外活動（週あたりの時間）			
	< 2	2から12	> 12	
C以上	$p_{1+}p_{+1}$	$p_{1+}p_{+2}$	$p_{1+}p_{+3}$	p_{1+}
D以上	$p_{2+}p_{+1}$	$p_{2+}p_{+2}$	$p_{2+}p_{+3}$	p_{2+}
	p_{+1}	p_{+2}	p_{+3}	

ベイズ法による
２元分割表の独立性の検定(2)

- 依存モデルを支持するベイズファクター

$$BF=\frac{D(y+1)D(1_R)D(1_C)}{D(1_{RC})D(y_R+1)D(y_C+1)}$$

y　:頻度の行列

y_R : 行合計のベクトル

y_C : 列合計のベクトル

1_R : R個の1を要素とするベクトル

D(v): ディリクレ関数　$D(v)=\prod \Gamma(v_i)/\Gamma(\sum v_i)$

ベイズ法による
２元分割表の独立性の検定（2）

- ctable()関数を用いる

  二元分割表に対してこのベイズファクターを計算
- 与える引数は

  行列a　：　確率行列を表す事前パラメータ

ベイズ法による
2元分割表の独立性の検定（3）

- すべての要素が1である行列aを指定する

  ⇒依存モデルの$\{p_{ij}\}$に一様事前分布

  $\{p_{i+}\}$, $\{p_{j+}\}$ のそれぞれにも一様分布

  これらを割り当てたことになる

```
> a <- matrix(rep(1,6), c(2,3))
> a
     [,1] [,2] [,3]
[1,]    1    1    1
[2,]    1    1    1
> ctable(data,a)
[1] 1.662173
```

- BF=1.66となり、

  控えめであるが独立モデルよりも依存モデルを支持する結果となった