Rで学ぶベイズ統計学入門

9.3　Zellnerのg事前分布によるモデル選択

9.4　生存モデル

茨城大学工学部
佐々木研究室
荒井悠有

9.3 Zellnerのg事前分布によるモデル選択

-はじめに

　これまでは(β,σ^2)に無情報事前分布を割り当ててきた。
　ここでは、g事前分布と呼ばれる事前情報を回帰モデルに組み込み、モデル選択を行う方法を紹介する。

・g事前分布

ここでは、σに条件付けた回帰係数のベクトルβは平均β^0で分散共分散行列$c\sigma^2(X'X)^{-1}$の多変量正規事前分布にしたがうと仮定する。
σ^2には、$1/\sigma^2$に比例する標準的な無情報事前分布を割り当てる。

-g事前分布を使うには次の２種類の数量を指定する

・回帰係数のベクトルβ^0の推定値
・事前分布に対するデータの情報量を反映させる定数c

＊cの値が
- 小さい・・・事前の確信に重きをおく
- 大きい・・・無情報事前分布を選んだ場合と同じ効果

・(β,σ^2)の結合事後密度

$$g(\beta,\sigma^2 \mid y) = g(\beta \mid y,\sigma^2)g(\sigma^2 \mid y)$$

平均β^1で分散共分散行列V^1の多変量正規分布

$$\beta^1 = \frac{c}{c+1}\left(\frac{\beta^0}{c} + \hat{\beta}\right) \qquad V_1 = \frac{\sigma^2 c}{c+1}\left(X'X\right)^{-1}$$

・σ^2の周辺事後分布

逆ガンマ分布 inverse gamma(a_1,b_1)

$$a_1 = n/2 \qquad b_1 = \frac{S}{2} + \frac{1}{2(c+1)}\left(\beta^0 - \hat{\beta}\right)' X'X\left(\beta^0 - \hat{\beta}\right)$$

・例

-ツノメドリ(puffin)繁殖の成功率を調べたデータを検討してみる

＜観測データ＞

営巣頻度NEST、草類被覆GRASS、平均土壌深度SOIL、
崖の傾斜角ANGLE、崖縁からの距離DISTANCE

NESTとDISTANCEの関係を次の単回帰モデルで説明する。

$$NEST_i = \beta_0 + \beta_1\left(DISTANCE_i - \overline{DISTANCE}\right) + \varepsilon_i$$

$NEST_i$、$DISTANCE_i$・・・それぞれi番目のツノメドリに関する
営巣頻度と草類被覆

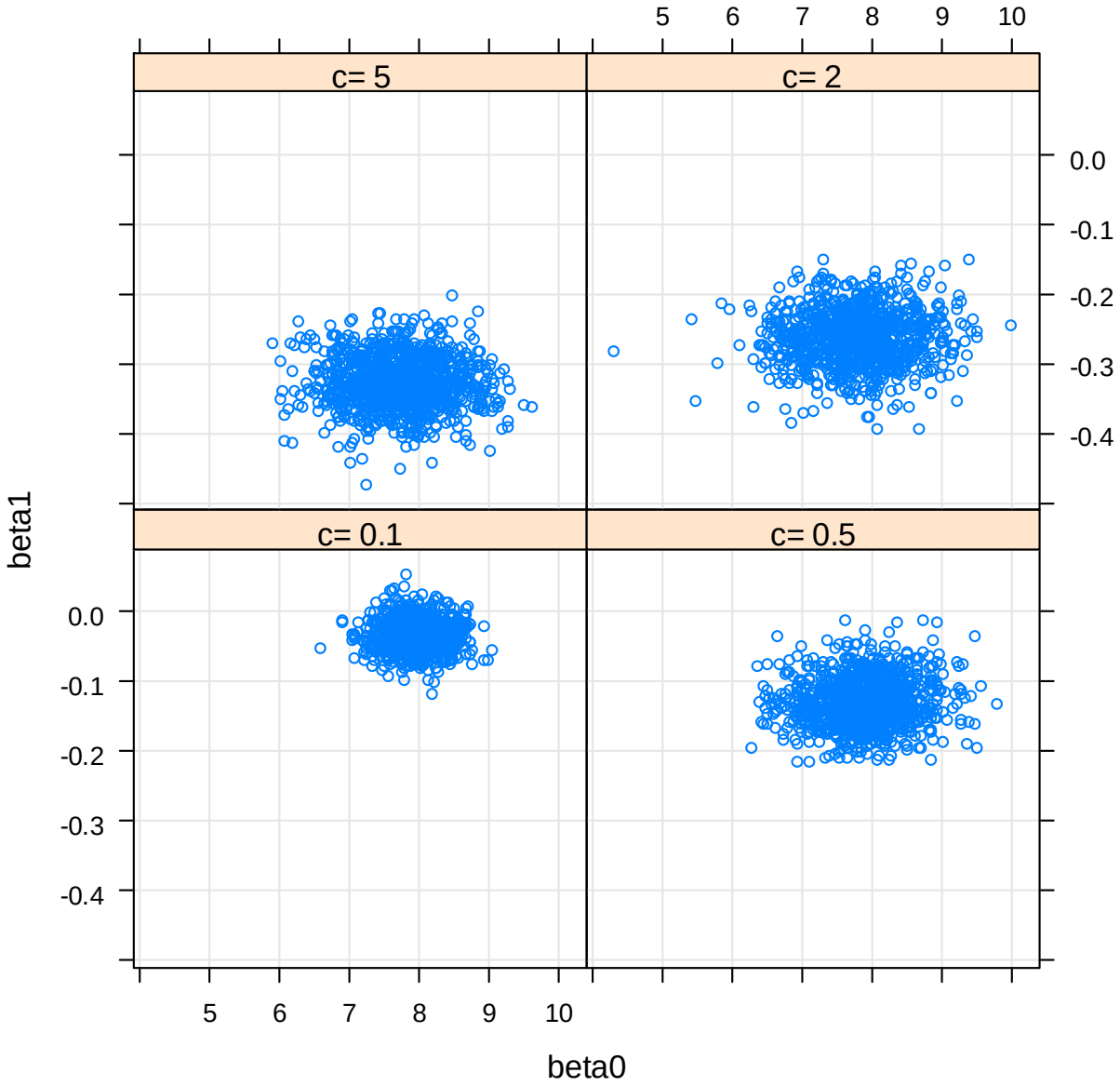
$\{\varepsilon_i\}$・・・正規分布$N(0,\sigma)$からのランダム標本

回帰係数のベクトル$\beta=(\beta_0,\beta_1)$の事前推定が$\beta^0=(8,0)$に等しい
とする。

(β,σ)の事後標本を、以下の値を使ったg事前分布によってシミュレーションする。

事前の推定β^0=(8,0)、事前の定数c=0.1,0.5,2,5

```
> data(puffin)
> X <- cbind(1,puffin$Distance - mean(puffin$Distance))
> c.prior <- c(0.1, 0.5, 5, 2)
> fit <- vector("list", 4)
> for (j in 1:4)
+ {
+ prior <- list(b0 = c(8,0), c0 = c.prior[j])
+ fit[[j]] <- blinreg(puffin$Nest, X, 1000, prior)
+ }
> BETA <- NULL
> for (j in 1:4)
+ {
+ s <- data.frame(Prior = paste("c=", as.character(c.prior[j])),
+ beta0 = fit[[j]]$beta[,1], beta1 = fit[[j]]$beta[,2])
+ BETA <- rbind(BETA,s)
+ }
> library(lattice)
> with(BETA, xyplot(beta1 ~ beta0|Prior, type = c("p", "g")))
```



事前パラメータcと、それぞれによるβ_0,β_1の事後分布

・モデル選択

-g事前密度クラスを回帰分析の問題で最適なモデル選択に利用する

-反応変数yにk個の予測変数の候補があるとする。

→回帰モデルにおける予測変数の組み合わせは全部で2^k通り

yの事前予測密度は以下の積分で与えられる。

$$m(y)=\int f(y|\beta,\sigma^2)g(\beta,\sigma^2)d\beta d\sigma^2$$

反応変数のサンプリング密度に$f(y|\beta,\sigma^2)$

パラメータのベクトル(β,σ^2)に事前密度$g(\beta,\sigma^2)$

を割り当てている

・bayes.model.selection()関数
-g事前密度を使って、変量のすべての組み合わせについて回帰モデルの対数予測密度を計算する。

＜引数＞
・反応値のベクトルy
・共変量の行列X
・g事前密度のパラメータc
・定数項が行列Xに含まれるかを指示する論理ベクトルconstant

＜返り値＞
・モデル、対数予測密度、事後モデルの確率からなるデータフレームmod.prob
・モデルごとにラプラスのアルゴリズムが収束したかを示す論理ベクトルconverge

9.4 生存モデル

-生存分析において寿命のモデルを構築したいとする。
n個体の集合で、寿命$t_1,\ldots,t_n$を観測した。

p個の共変量$x_1,\ldots,x_p$を使って生存期間のばらつきを記述したいとする。
このモデルは以下の対数線形モデルで表現できる。

$$\log t_i = \mu + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \sigma\varepsilon_i$$

$x_{i1},\ldots,x_{ip}$はi番目の個体のp個の共変量
ε_iは密度$f(\varepsilon)=\exp(\varepsilon-e^{\varepsilon})$のガンベル分布にしたがうと仮定

・時間の対数$y_i=\log t_i$の密度

$$f_i(y_i)=\frac{1}{\sigma}\exp(z_i-e^{zi})$$

ただし、$z_i=(y_i-\mu-\beta_1 x_{i1}-\cdots-\beta_p x_{ip})/\sigma$

i番目の個体の生存関数は$S_i(y_i)=\exp(-e^{zi})$

・回帰係数のベクトル$\beta=(\beta_1,\ldots,\beta_p)$と$\mu,\sigma$の尤度関数

$$L(\beta,\mu,\sigma)=\prod_{i=1}^{n}\{f_i(y_i)\}^{\delta i}\{S_i(y_i)\}^{1-\delta i}$$

・事後密度（比例定数は除く）

μ、βには一様分布を割り当て、σには$1/\sigma$に比例する通常の無情報事前分布を割り当てる

$$g(\beta,\mu,\sigma\mid data)\propto\frac{1}{\sigma}L(\beta,\mu,\sigma)$$