ベイズ統計データ分析
R&WinBUGS

古谷知之著

羽藤研究室
修士2年　松村草也

BehaviorInNetworks

目次

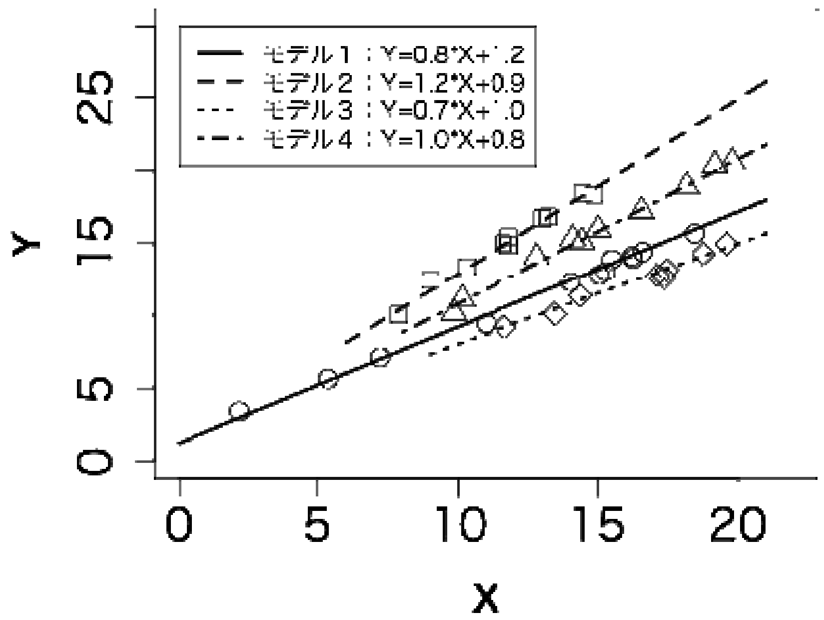
1

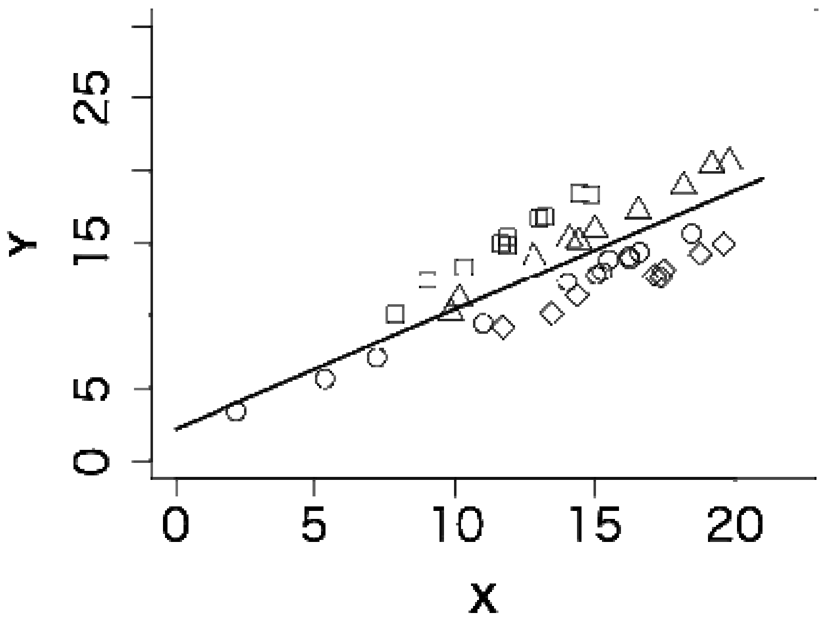
ベイズアプローチの基本

Basis for Bayesian Approach

ベイズアプローチの基本

複数のモデルでデータを再現 ⟷ 1つのモデルを求める





- モデル（パラメータセット）によって説明出来るデータが異なってくる.
- P(観測データ)は同時確率P(観測データ|モデル)P(モデル)の和であり，モデルを特定しない場合に全てのデータが観測される「周辺確率」である.
- データが観測されればそれは固定されることになる.（θの確率分布には影響しない）

ベイズの定理

$$
\begin{aligned}
P(\mathcal{D}|model1) &= 12/40 = 0.30 \\
P(\mathcal{D}|model2) &= 10/40 = 0.25 \\
P(\mathcal{D}|model3) &= 8/40 = 0.20 \\
P(\mathcal{D}|model4) &= 10/40 = 0.25
\end{aligned}
$$

↑モデルが与えられた状態でデータが観測される尤もらしさ.

$$
\begin{aligned}
P(\mathcal{D}) &= P(\mathcal{D}|model1)P(model1) \\
&\quad +P(\mathcal{D}|model2)P(model1) \\
&\quad +P(\mathcal{D}|model3)P(model1) \\
&\quad +P(\mathcal{D}|model4)P(model1) \\
&= 0.25
\end{aligned}
$$

↑モデルを特定しない場合に全てのデータが観測される確率

$$
\begin{aligned}
P(model1|\mathcal{D}) &= \frac{P(\mathcal{D}|model1)P(model1)}{P(\mathcal{D})} \\
&= 0.30 \\
P(model2|\mathcal{D}) &= 0.25 \\
P(model3|\mathcal{D}) &= 0.20 \\
P(model4|\mathcal{D}) &= 0.20
\end{aligned}
$$

↑データが与えられた時にモデルが成立する確率

$P(model1|\mathcal{D})$: 事後分布

$P(\mathcal{D}|model1)$: 尤度関数

$P(model1)$: 事前分布

$P(\mathcal{D})$: 周辺確率

従来のモデル推定方法とベイズアプローチの違い

- データを得る前のモデル(パラメタ)に関する事前の知識としての事前情報をあらかじめ用意し，実験や調査などから得られたデータが成立するかどうかという仮説の確からしさとを掛け合わせる.
- 頻度主義的立場からは，最も好ましいモデルを求める.
- ベイズ主義的立場からは，パラメータの尤もらしさの分布を求める.

2

ベイズ推論

Baysian Estimation

事前分布の与え方とその特徴

- ベイズ的手法を用いたデータ分析では，何らかの事前情報確率p(θ)を与える事により，事後確率密度分布を計算している．
- その際，事前情報については何ら情報を持たないままに，あてずっぽで事前情報確率を与えている．
- そこで，「与える事前情報確率を変えたらθの事後確率も変化するのか？」という疑問がわいてきます。。

事前分布の与え方とその特徴

$$p(\theta|y) \propto \mathbf{l}(y|\theta)p(\theta)$$

事後情報　　尤度関数　事前情報

■事前分布の例

自然共役事前分布 natural conjugate prior distribution	事後分布と同じ分布が得られる
非正則事前分布 improper prior distribution	θに対する確率分布を標本分布に関して積分しても収束しない分布.
無情報事前分布 noninformative prior distribution	事前情報を定数と置く事で事前情報からの依存を無くすことが出来る分布.
階層事前分布 hierarchical prior distribution	未知パラメータθを標本iごとに推定するモデル. ハイパーパラメータλを用いてθを推定する.

■共役事前分布の例

変数Xの従う分布(尤度関数)	事前分布	事後分布
正規分布	正規分布	正規分布
指数分布	ガンマ分布	ガンマ分布
ポアソン分布	ガンマ分布	ガンマ分布

線形回帰モデル - 定義

■線形回帰モデルの定義

$$
\begin{aligned}
\mathbf{y} &= X\beta + \epsilon, \epsilon\, \mathcal{N}(0, \sigma^2) \\
y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \\
\mathbf{y} &= (y_1, \cdots, y_i, \cdots y_n)^\top \\
X &= \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 1 & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{nk} \end{pmatrix} \\
\beta &= (\beta_0, \cdots, \beta_k), \epsilon = (\epsilon_1, \cdots, \epsilon_n)^\top
\end{aligned}
$$

■計算例のデータについて
Rに組み込まれている，スイスの1888年の社会経済データ.
地域ごとの出生力/男性の農業従事者割合/教育水準/カソリックの割合/乳児死亡率など

```
> swiss
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6
Broye	83.8	70.2	16	7	92.85	23.6
Glane	92.4	67.8	14	8	97.16	24.9
Gruyere	82.4	53.3	12	7	97.67	21.0
Sarine	82.9	45.2	16	13	91.38	24.4
Veveyse	87.1	64.5	14	6	98.61	24.5
Aigle	64.1	62.0	21	12	8.52	16.5
Aubonne	66.9	67.5	14	7	2.27	19.1
Avenches	68.9	60.7	19	12	4.43	22.7
Cossonay	61.7	69.3	22	5	2.82	18.7
Echallens	68.3	72.6	18	2	24.20	21.2
Grandson	71.7	34.0	17	8	3.30	20.0
Lausanne	55.7	19.4	26	28	12.11	20.2
La Vallee	54.3	15.2	31	20	2.15	10.8
Lavaux	65.1	73.0	19	9	2.84	20.0
Morges	65.5	59.8	22	10	5.23	18.0
Moudon	65.0	55.1	14	3	4.52	22.4
Nyone	56.6	50.9	22	12	15.14	16.7
Orbe	57.4	54.1	20	6	4.20	15.3
Oron	72.5	71.2	12	1	2.40	21.0
Payerne	74.2	58.1	14	8	5.23	23.8
Paysd'enhaut	72.0	63.5	6	3	2.56	18.0
Rolle	60.5	60.8	16	10	7.72	16.3
Vevey	58.3	26.8	25	19	18.46	20.9
Yverdon	65.4	49.5	15	8	6.10	22.5
Conthey	75.5	85.9	3	2	99.71	15.1
Entremont	69.3	84.9	7	6	99.68	19.8
Herens	77.3	89.7	5	2	100.00	18.3
Martigwy	70.5	78.2	12	6	98.96	19.4
Monthey	79.4	64.9	7	3	98.22	20.2
St Maurice	65.0	75.9	9	9	99.06	17.8
Sierre	92.2	84.6	3	3	99.46	16.3
Sion	79.3	63.1	13	13	96.83	18.1
Boudry	70.4	38.4	26	12	5.62	20.3

線形回帰モデル - 最小二乗法による解法

■Rのlmによる計算例

```
# swissデータの読み込み
data(swiss)
summary(swiss)
# lm(被説明変数~説明変数, data=データ名)
# .はその他のデータを省略することを意味する
summary(lm(Fertility~.,data=swiss))
```

■β, σに着目した計算

```
k<- ncol(swiss)
y<-swiss[,1]
X<-cbind(1,as.matrix(swiss[,2:k]))
n<-nrow(X)
k<-ncol(X)

# βの不偏推定量
# solve() は逆行列を計算する
betahat <- solve(t(X)%*%(X))%*%t(X)%*%y
betahat
# σ^2の不偏推定量
S2=t(y-X%*%betahat)%*%(y-X%*%betahat)
sig2hat <- S2/(n-k)
sig2hat
# βの不偏分散
diag(as.real(sig2hat)*solve(t(X)%*%X))
```

■尤度関数としてとらえてみると？

$$l(\beta,\sigma^2|y,X) = -\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_i x_{i1}-\cdots-\beta_n x_{in})^2\}$$

$$\boxed{\begin{aligned} l(\beta,\sigma^2|y,X) = -\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(\mathbf{y}-X\hat{\beta})^\top(\mathbf{y}-X\hat{\beta}) \\ -\frac{1}{2\sigma^2}(\beta-\hat{\beta})^\top(X^\top X)(\beta-\hat{\beta}))\} \end{aligned}}$$

→これはσ^2に関する逆ガンマ関数.

$$\begin{aligned} \Gamma^{-1}(a,b) &= \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}\exp\{-b(\sigma^2)^{-1}\} \\ &\propto (\sigma^2)^{-a-1}\exp\{-b(\sigma^2)^{-1}\} \end{aligned}$$

$$\boxed{\begin{aligned} \beta|\sigma^2,X &\sim \mathcal{N}(\tilde{\beta},\sigma^2M^{-1}) \\ \sigma^2|X &\sim \mathcal{IG}(a,b), a>0, b>0 \end{aligned}}$$

Mはk*kの正定値対象行列
JGは逆ガンマ関数

線形回帰モデル - 最小二乗法による解法

```
Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2743  -5.2617   0.5032   4.1198  15.3213 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *  
Examination      -0.25801    0.25388  -1.016  0.31546    
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 ** 
Infant.Mortality  1.07705    0.38172   2.822  0.00734 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared: 0.7067,	Adjusted R-squared: 0.671 
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

- βとσの事前分布について何も情報を持たない状況で、適当にパラメータを設定して得られたもの。

線形回帰モデルのベイズ推定 - 尤度関数として扱う

β(ガウス型関数),σ(逆ガンマ関数) の事後平均
β(ガウス型関数)の事後分散を計算し，比較を行う.

$$\beta|\sigma^2, X \sim \mathcal{N}(\tilde{\beta}, \sigma^2 M^{-1})$$
$$\sigma^2|X \sim \mathcal{IG}(a, b), a > 0, b > 0$$

```
#### 2.2.2 ####
M <- diag(k)

a <- 2.1
b <- 2
c <- 0.1
M=(1/c)*diag(k)
T=solve(solve(M)+solve(t(X)%*%X))
# βの事後平均
betavar <- solve(M+t(X)%*%X)%*%t(X)%*%y
betavar
# σ^2の事後平均
S2=t(y-X%*%betahat)%*%(y-X%*%betahat)
S2var <- (2*b+S2+t(betahat)%*%T%*%betahat)/(n+2*a-
2)
S2var
# βの事後分散
betasig <-
diag(as.real(2*b+S2+t(betahat)%*%T%*%betahat)/(n+2
*a-2)*solve(M+t(X)%*%X))
betasig
```

$\tilde{\beta}, M, a, b$ については情報がないので適当に設定.

→ 事前情報への依存度が高い.

信頼区間と最高事後密度

ところで，得られた事後分布はどの程度信頼出来る結果なのか？

■信頼区間

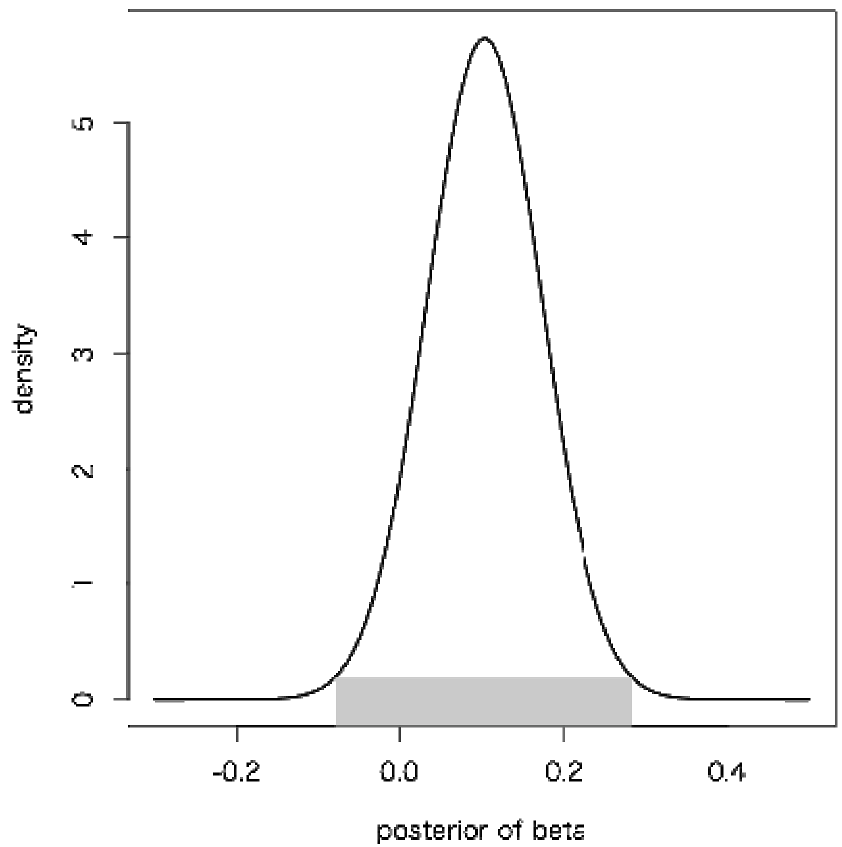
$$p(\frac{\alpha}{2} \leq \theta \leq 1-\frac{\alpha}{2}) = \int_{\alpha/2}^{1-\alpha/2} p(\theta)d\theta = 1-\alpha$$

この区間内にθが入る確率が1-αであるため，
100(1-a)%信頼区間と呼ぶ.

■最高事後密度(HPD:Highest Posterior Density)

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 \omega_{ii}}} \sim \mathcal{T}_{n-k-1}$$

t検定量を定義．T_{n-k-1}は標本数n，自由度n-k-1のt分布である．ωiiは$X^{\top}X$の(i,i)要素を意味する.



$$\hat{\beta} + \sqrt{\hat{\sigma}^2 X^{\top} X F^{-1}_{n-k-1}}(\frac{\alpha}{2}), \hat{\beta} + \sqrt{\hat{\sigma}^2 X^{\top} X F^{-1}_{n-k-1}}(1-\frac{\alpha}{2})$$

経験ベイズによる推定

$$
\begin{aligned}
\mathbf{y} &= X\beta + \epsilon, \epsilon\,\mathcal{N}(0, \sigma^2) \\
y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \\
\mathbf{y} &= (y_1, \cdots, y_i, \cdots y_n)^\top \\
X &= \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 1 & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{nk} \end{pmatrix} \\
\beta &= (\beta_0, \cdots, \beta_k), \epsilon = (\epsilon_1, \cdots, \epsilon_n)^\top
\end{aligned}
$$

$$\tau = \frac{1}{\sigma^2}$$

$$p(\epsilon|\tau) \propto \tau^{n/2} \exp\{-(\frac{\tau}{2})\epsilon^\top \epsilon\} \quad （1）$$

$$p(\beta|\tau) \propto \frac{1}{\tau} \quad （2）$$

- 誤差項の分散σ^2を精度τ＝1/σ^2に置き換える。
- τのように、事前分布の形状を決定するようなパラメタのことを超パラメタと呼ぶ。
- モデルパラメータを求めるためには、βとτの事前分布が必要である。単純な例として（２）のような非正則一様分布を与える。

経験ベイズによる推定

- 線形回帰モデルのパラメータを，予測分布をシミュレーションする方法を用いてベイズ推定する.（この方法はMCMCの基礎となるものである．）
- 同じ線形回帰モデルを使用.

```
#### 2.4節 ####
ndraw <- 1000   # 繰り返し計算回数
#  τの事前情報を計算
tauval <- rgamma(ndraw,(n-k)/2,(n-k)*sig2hat/2)
bval <- matrix(0,nrow=ndraw,ncol=k)
# 多変量正規分布の関数mvrmormを使うためライブラリMASSを呼び出す
library(MASS)
for(i in 1:ndraw){
V <- (1/tauval[i])*solve(t(X)%*%(X))
bval[i,] <-mvrnorm(1,betahat, V)}
# パラメータの事後平均・分散・標準偏差を計算
bval.mean <- matrix(0,nrow=k,ncol=1)
for(i in 1:k){bval.mean[i,] <- mean(bval[,i])}
bval.mean
bval.var <- matrix(0,nrow=k,ncol=1)
for(i in 1:k){bval.var[i,] <- var(bval[,i])}
bval.var
bval.sd<-sqrt(bval.var)
```

τ→ε→βを繰り返し求める.

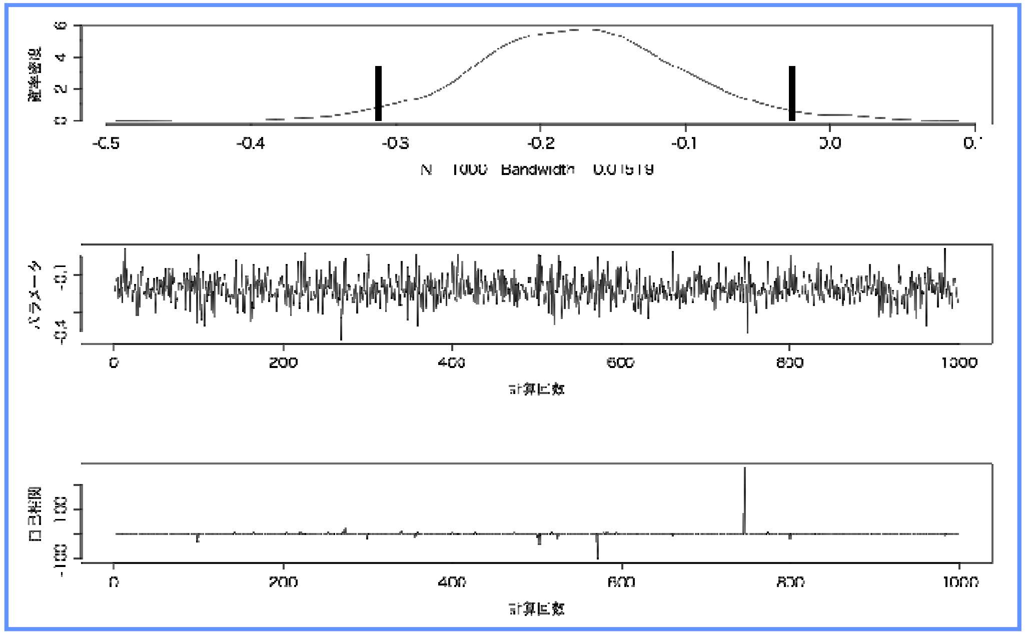
βとτの事前分布はどうするか？
→簡単な下記の非正則一様分布を与える.

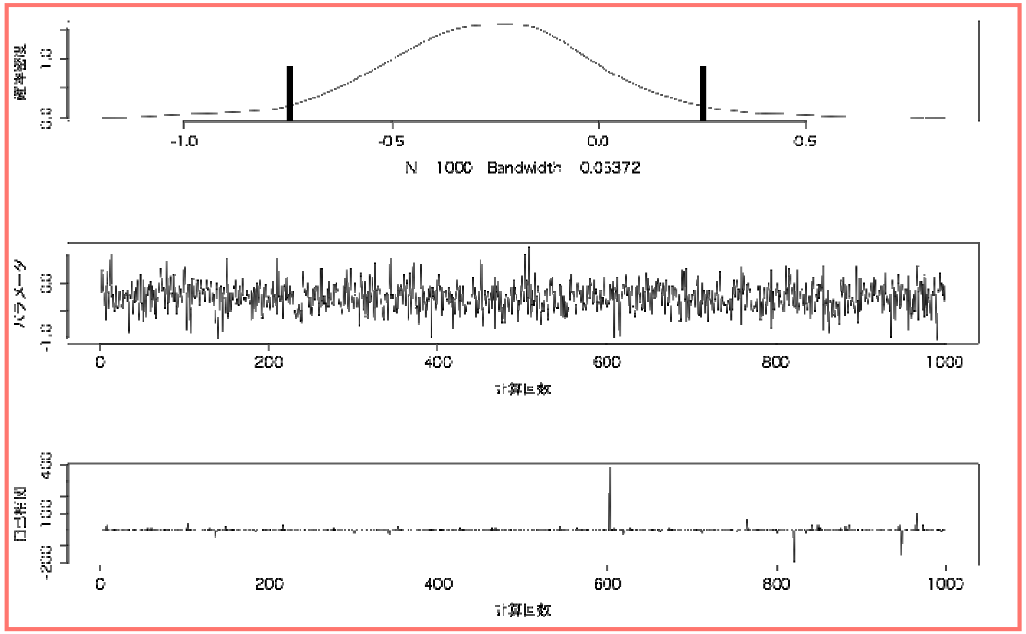
$$\tau = \frac{1}{\sigma^2}$$

$$p(\epsilon|\tau) \propto \tau^{n/2} \exp\{-(\frac{\tau}{2})\epsilon^\top \epsilon\}$$

$$p(\beta|\tau) \propto \frac{1}{\tau}$$

経験ベイズによる推定





swissデータの線形回帰モデルのベイズ推定結果

	平均	標準偏差	97.5%	2.5%
定数	67.2	1.29E+02	88.829	45.586
農業	-0.172	5.32E-03	-0.030	-0.314
出生力	-0.246	6.68E-02	0.266	-0.759
教育	-0.883	3.26E-02	-0.513	-1.252
カソリック	0.105	1.19E-03	0.177	0.034
幼児死亡率	1.059	1.52E-01	1.83	0.288

・「農業」はパラメタが負で有る可能性が高い.
・「出生力」はどちらかわからない.
　→推定精度が低い. 最小二乗法におけるパラメタ推定結果ではt値が5%水準に達しておらず.

3

マルコフ連鎖モンテカルロ法

Markov Chain Monte Carlo method

はじめに

- ベイズ推論では，与えられた事前情報に対して事後分布をサンプリングするシミュレーション計算により，予想事後分布を得る事が出来る.
- シミュレーションの計算回数が少ないと事前情報に強く依存する.
- 「何回程度計算すれば事前情報の影響を除去出来るか？」
- 「膨大なシミュレーション計算を効率的に実行するには？」

マルコフ連鎖について

■推移確率(transition probability)

$$\begin{aligned} P(X_{t+1}) \in A|x_1, \cdots, x_t) &= P(X_{t+1} \in A|x_t) \\ &= P(X_{t+1}|x_t)P(X_t|x_{t-1}) \cdots P(X_2|x_1)P(x_1) \end{aligned}$$

■推移核(transition kernel)

$$K(x,y) = P(X_{t+1} = y|X_{t+1} = x), x, y \in A$$

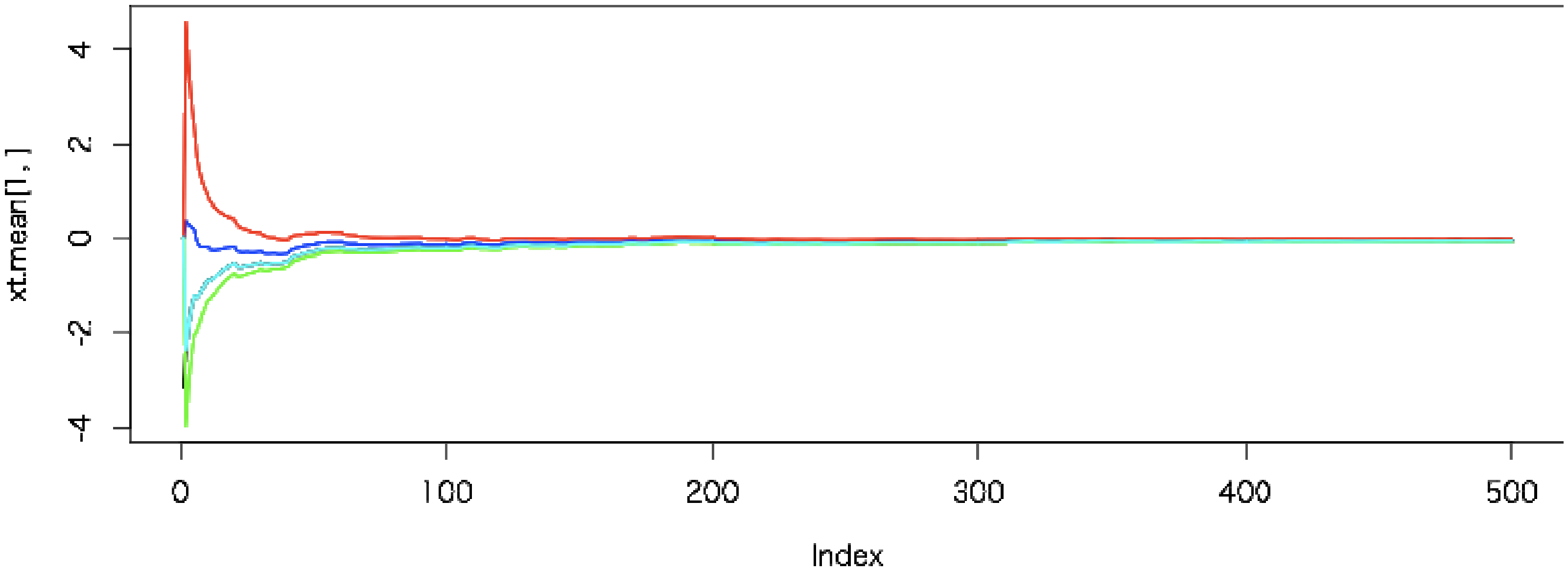
$$p(y) = \int K(x,y)p(x)dx$$

■時系列回帰モデル（マルコフ連鎖の代表例）

$$x_{t+1} = \rho x_t + \epsilon_{t+1}, \epsilon_{t+1} \sim \text{i.i.d.}\mathcal{N}(0,1), t = 0, 1, \cdots, T$$

$$K(x_{t+1}, x_t) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(x_{t+1} - \rho x_t)^2\}$$

時系列回帰モデル - 数値計算結果



- 5通りの乱数を生成し、標本平均をプロットしている.
- このような単純なマルコフ連鎖に限らず，MCMCでは事後分布の標本経路は初期値に依存する.
- しかし，初期値がどのような値でも，計算回数を重ねるごとに標本平均の値が定常状態に近づく事が図からみてとれる.
- 定常状態に至るまでの繰り返し計算を棄却し，定常状態にある推定値を用いて分析結果を議論した方が良いと考えられる.
- 繰り返し計算を始めてからサンプリングを棄却するまでの期間を「稼働検査期間」(burn-in-period)という.

ギブズ・サンプラー

マルコフ連鎖の推移核を構築するより一般的な方法.

■ギブズ・サンプラーのアルゴリズム

1.ドロー回数をsとする. s=(0,...,ndraw)
2.s=0に対して，初期値を決める.
3.x^sからx^{s+1}を生成する.

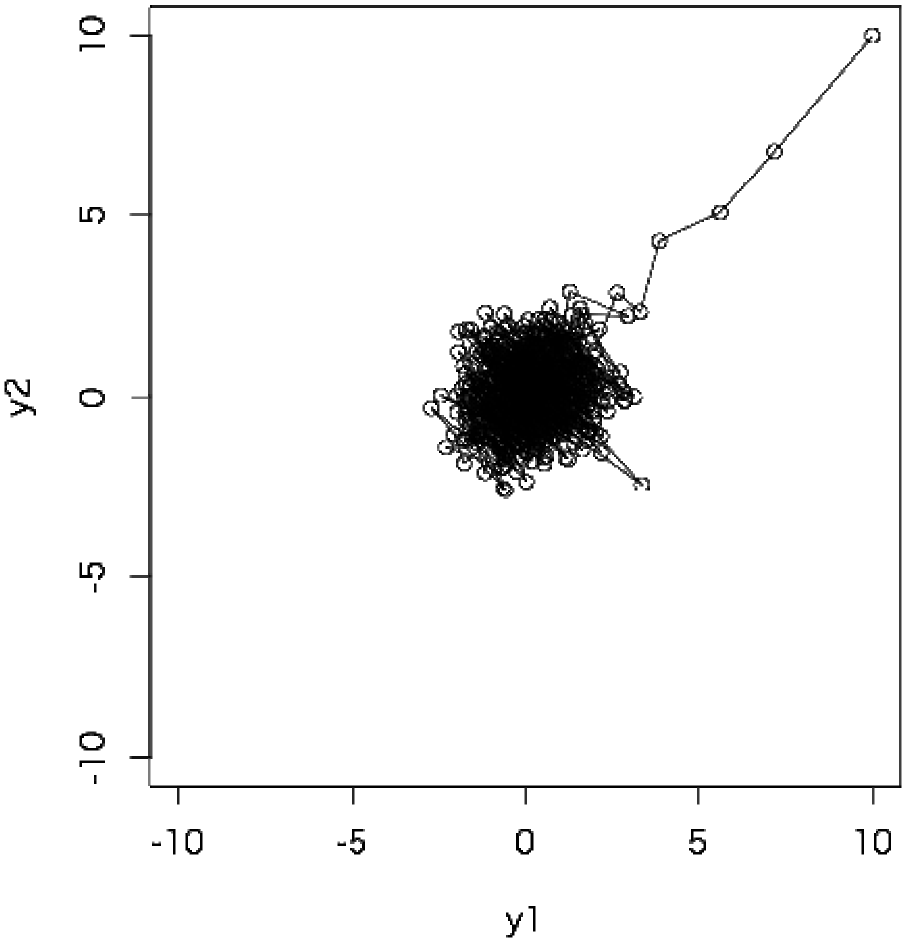
$$\mathbf{x}^s = (x_1^s, \cdots, x_l^s)$$

$$\begin{aligned} x_1^{s+1} &\sim p(x_1 | x_2^s, \cdots, x_t^s) \\ x_2^{s+1} &\sim p(x_2 | x_1^{s+1}, x_3^s \cdots, x_t^s) \\ &\vdots \\ x_t^{s+1} &\sim p(x_t | x_1^{s+1}, \cdots, x_{t-1}^{s+1}) \end{aligned}$$

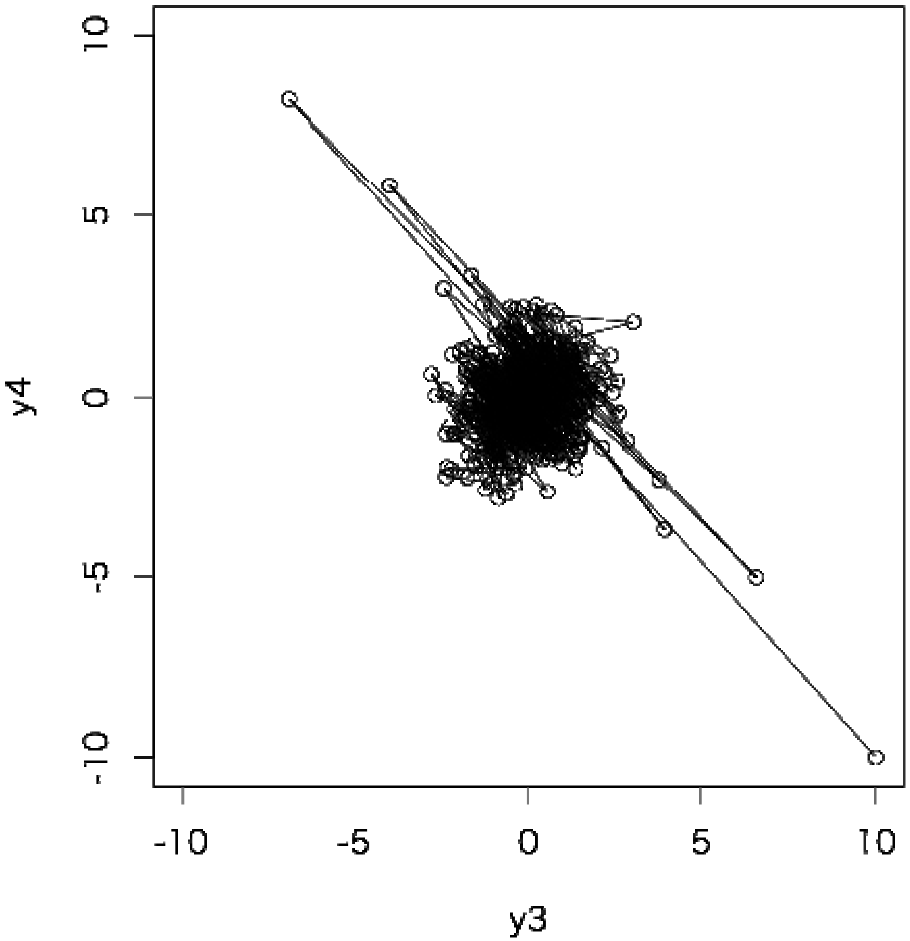
4.s<ndrawのとき，3に戻る. s=ndrawの時計算終了

```
# 初期値(10, 10)の場合
T<-500; ro<-0.7; y0<-10; mu1<-mu2<-0; sig1<-sig2<-1
y1 <- rep(y0,T); y2<-rep(y0,T)
for(i in 2:T){
y2[i] <- rnorm(1, (mu2+ro*sig2/sig1*(y1[i-1]-mu1)), sqrt(sig2*(1-ro^2)))
y1[i] <- rnorm(1, (mu1+ro*sig1/sig2*(y2[i-1]-mu1)), sqrt(sig1*(1-ro^2)))}
# 初期値(-10, 10)の場合
T<-500; ro<-0.7; y00<--10; mu1<-mu2<-0; sig1<-sig2<-1
y3 <- rep(y0,n); y4<-rep(y00,n)
for(i in 2:T){
y4[i] <- rnorm(1, (mu2+ro*sig2/sig1*(y3[i-1]-mu1)), sqrt(sig2*(1-ro^2)))
y3[i] <- rnorm(1, (mu1+ro*sig1/sig2*(y4[i-1]-mu1)), sqrt(sig1*(1-ro^2)))}
# 図3.3
par(mfrow=c(1,2))
plot(y1,y2,type="o",xlim=c(-10,10),ylim=c(-10,10))
plot(y3,y4,type="o",xlim=c(-10,10),ylim=c(-10,10))
```

ギブズ・サンプラー



初期値：(10,10)

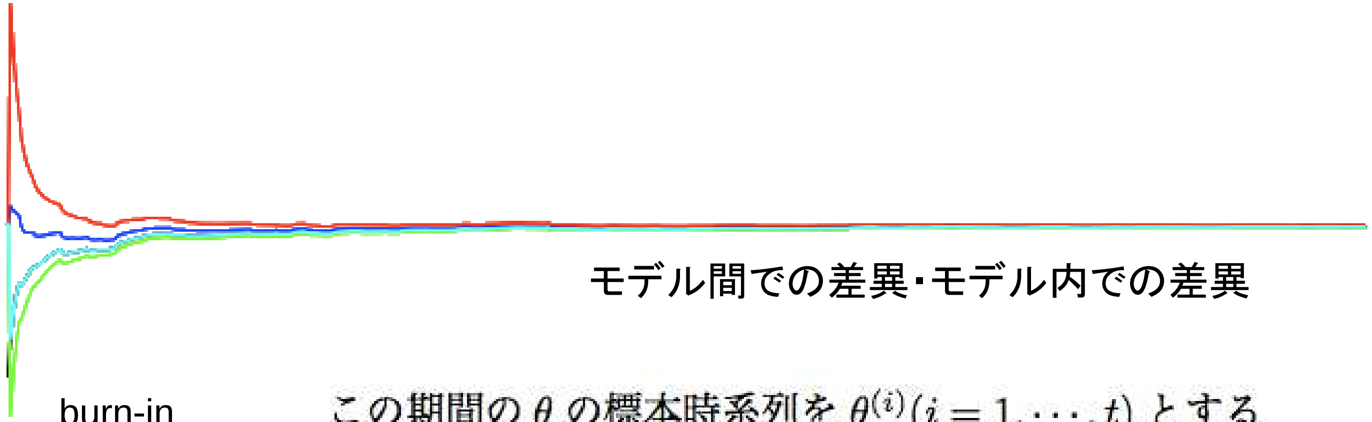


初期値：(10,-10)

↑初期値によらず定常状態に近づくことがわかる.

収束判定

MCMCでは，標本経路が初期値に依存しない定常状態であるかを見ることによって，収束判定を行う.



burn-in

この期間の θ の標本時系列を $\theta^{(i)}(i=1,\cdots,t)$ とする.

ndraw

名称	メソッド名	概要
Gelman-Rubin統計量	gelman.diag()	モデル間の標本時系列の平均，モデル内の平均を利用した指標R
Gewekeの判定方法	geweke.diag()	平均値の差に関するZ検定を応用
Raftery-Lewisの診断方法	raftery.diag()	マルコフ連鎖の四分位偏差を用いる

離散選択モデルのサンプルプログラム

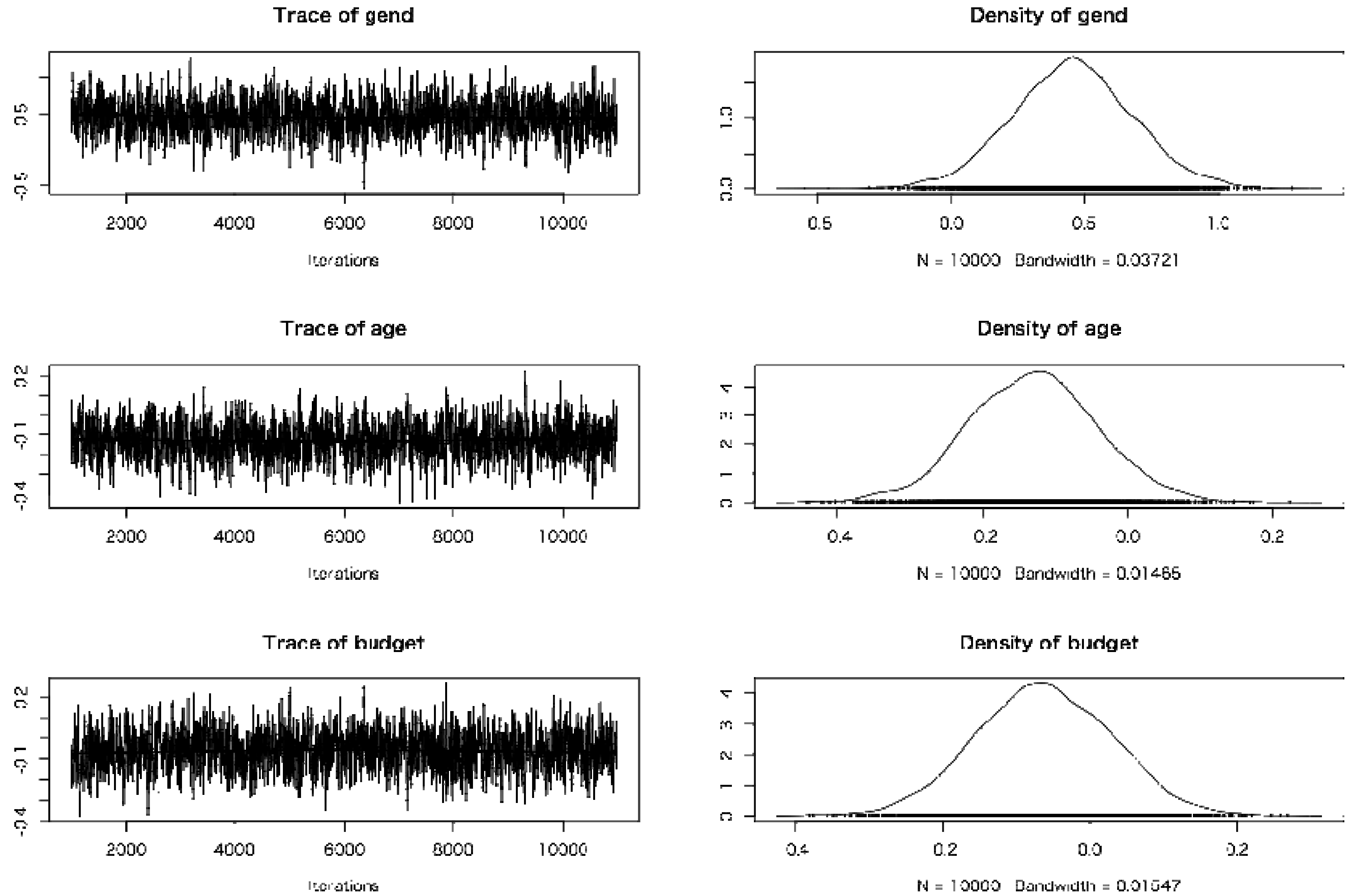
■サンプルデータについて

目的変数：満足度2
説明変数：性別、年齢、年収

のMNLモデルをMCMCにより、burnin1000回、ドロー10000回にて推定を行った結果を下記に示す。

```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

           Mean     SD Naive SE Time-series SE
gend    0.45403 0.2286 0.002286       0.006974
age    -0.13106 0.0886 0.000886       0.003121
budget -0.06319 0.0921 0.000921       0.002944

2. Quantiles for each variable:

            2.5%     25%      50%        75%   97.5%
gend    0.002459  0.3033  0.45453  0.6001149 0.91065
age    -0.306707 -0.1906 -0.13053 -0.0736864 0.04396
budget -0.243822 -0.1260 -0.06335  0.0005509 0.11695
```

推定結果



Trace of gend
Density of gend
Iterations
N = 10000 Bandwidth = 0.03721
Trace of age
Density of age
Iterations
N = 10000 Bandwidth = 0.01466
Trace of budget
Density of budget
Iterations
N = 10000 Bandwidth = 0.01547
2000
4000
6000
8000
10000