Stataではじめるベイズ統計

株式会社ライトストーン

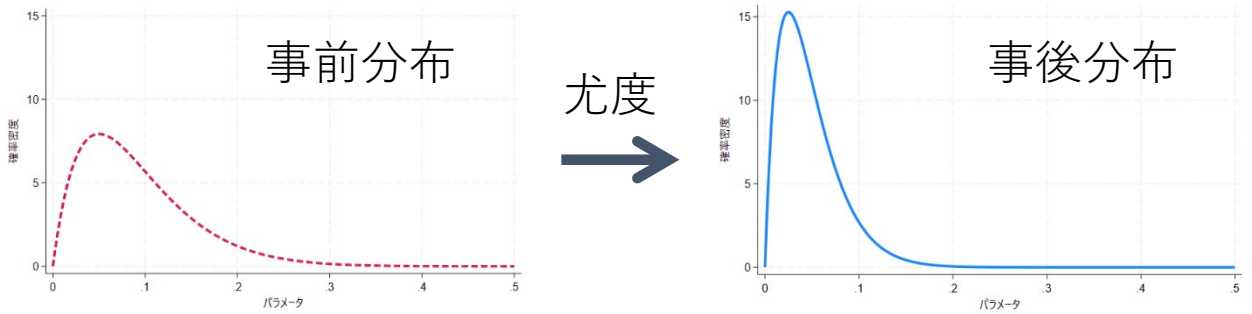
目次

目次

- ベイズ統計とは
- 分析例：感染症の有病率
- Stataのベイズ推定コマンド

ベイズ統計とは

1. ベイズ統計とは
   - モデルのパラメータが何らかの確率分布（事前分布）に従う
   - 尤度と事前分布から事後分布を得る（事後分布 ∝ 尤度 × 事前分布）
2. 長所
   - 事前情報を事前分布に組み込むことができる
   - 事後分布を任意の精度で求めることができる
   - 信用区間を求めて確率的に解釈できる
3. 短所
   - 事前分布に主観が入る
   - 計算の負荷が高い



分析例：感染症の有病率

背景

ある町における感染症の有病率を推定する。

20名を抽出し、感染者数をyとする。

有病率を表すパラメータを$\theta \in [0,1]$とする。

$$y \mid \theta \sim Binominal(20, \theta)$$

抽出した20名の中に感染者はいなかった。$(y = 0)$

既存研究より、θは概ね0.05から0.20で平均値は0.10

尤度と事前分布

$$y \mid \theta \sim Binominal(20, \theta)$$

$$\theta \sim Beta(2, 20)$$

事後分布

$$\theta \mid y \sim Beta(2, 40)$$

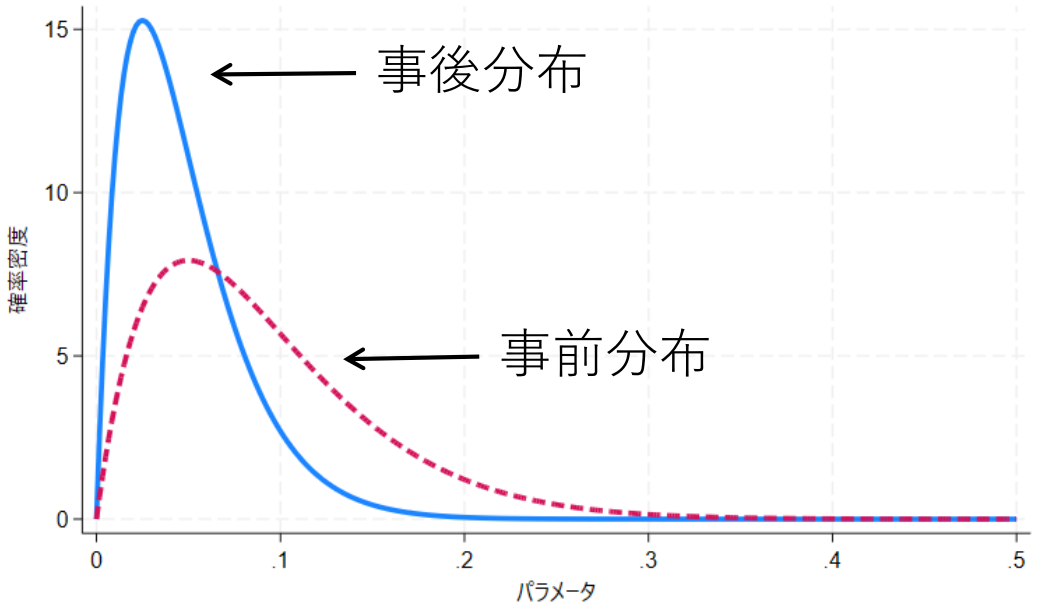
頻度統計の場合

推定値：

$$\bar{y} = \frac{y}{n} = \frac{0}{20} = 0$$

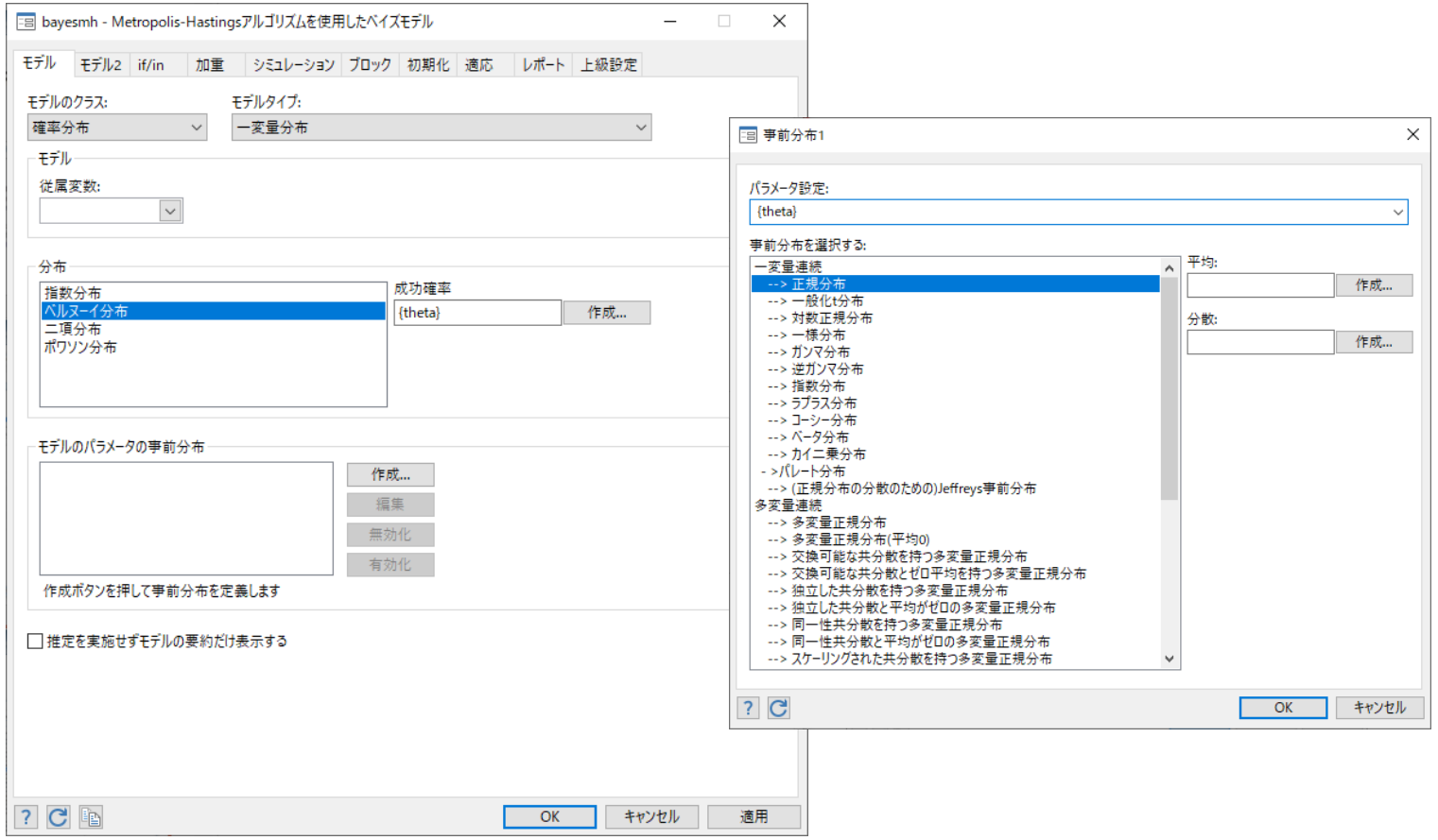
95%信頼区間：

$$\left(\bar{y} - 1.96\sqrt{\bar{y}(1-\bar{y})/n}, \bar{y} + 1.96\sqrt{\bar{y}(1-\bar{y})/n}\right) = 0$$



Stataのベイズ推定コマンド（**bayesmh** コマンド）

メニュー操作：[統計] > [ベイズ分析] > [推定と回帰一般]



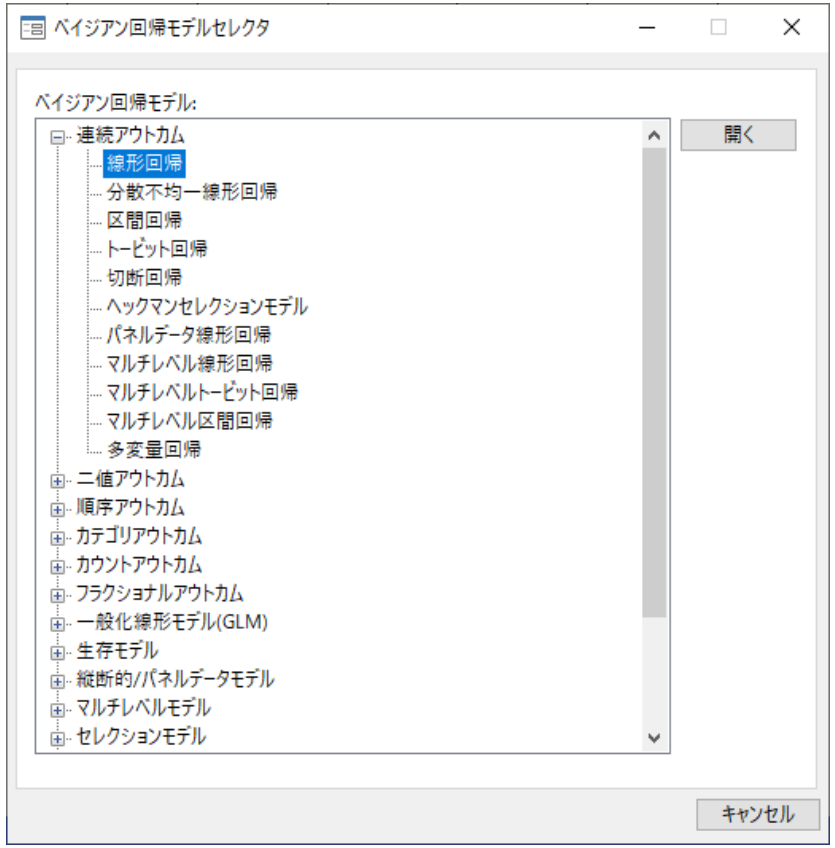
Stataのベイズ推定コマンド（**bayesmh**コマンド）

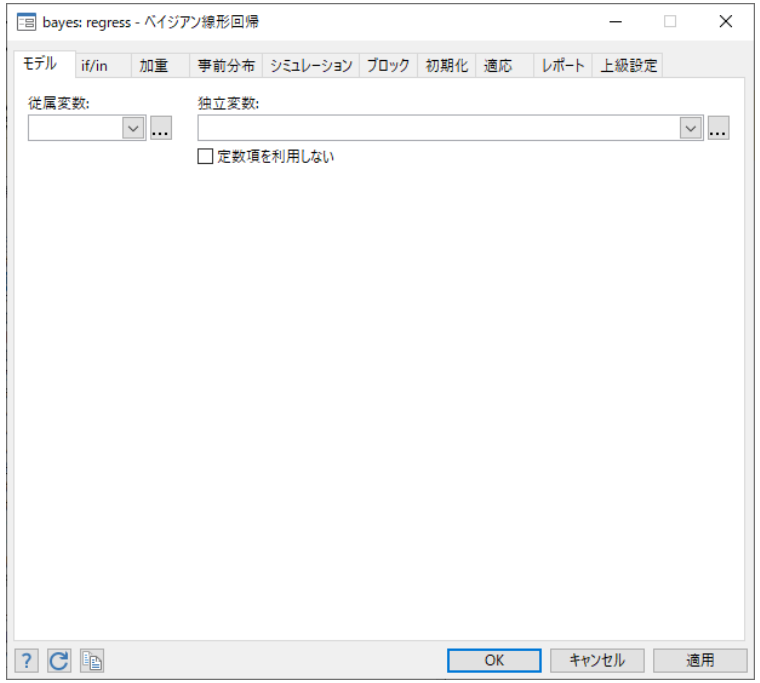
コマンド例

```
bayesmh sbp age sex bmi,                    ///
   likelihood(normal({sigma2}))             ///
   prior({sbp: _cons}, normal(0,100))       ///
   prior({sbp: age},   normal(0,100))       ///
   prior({sbp: sex},   normal(0,100))       ///
   prior({sbp: bmi},   normal(0,100))       ///
   prior({sigma2},      igamma(1,1))
```

Stataのベイズ推定コマンド（**bayes** プレフィクスコマンド）

メニュー操作：[統計] > [線形モデル他] > [ベイジアン回帰] > [線形回帰] 他





Stataのベイズ推定コマンド（**bayes** プレフィクスコマンド）

コマンド例

```
regress sbp age sex
bayes: regress sbp age sex

var inflation ogap fedfunds
bayes: var inflation ogap fedfunds
```

bayes プレフィクス対応の推定コマンド

Linear regression models

regress	[BAYES] bayes: regress
hetregress	[BAYES] bayes: hetregress
tobit	[BAYES] bayes: tobit
intreg	[BAYES] bayes: intreg
truncreg	[BAYES] bayes: truncreg
mvreg	[BAYES] bayes: mvreg

Binary-response regression models

logistic	[BAYES] bayes: logistic
logit	[BAYES] bayes: logit
probit	[BAYES] bayes: probit
cloglog	[BAYES] bayes: cloglog
hetprobit	[BAYES] bayes: hetprobit
binreg	[BAYES] bayes: binreg
biprobit	[BAYES] bayes: biprobit

Ordinal-response regression models

ologit	[BAYES] bayes: ologit
oprobit	[BAYES] bayes: oprobit
hetoprobit	[BAYES] bayes: hetoprobit
ziologit	[BAYES] bayes: ziologit
zioprobit	[BAYES] bayes: zioprobit

Categorical-response regression models

mlogit	[BAYES] bayes: mlogit
mprobit	[BAYES] bayes: mprobit
clogit	[BAYES] bayes: clogit

Count-response regression models

poisson	[BAYES] bayes: poisson
nbreg	[BAYES] bayes: nbreg
gnbreg	[BAYES] bayes: gnbreg
tpoisson	[BAYES] bayes: tpoisson
tnbreg	[BAYES] bayes: tnbreg
zip	[BAYES] bayes: zip
zinb	[BAYES] bayes: zinb

Generalized linear models

glm	[BAYES] bayes: glm

Fractional-response regression models

fracreg	[BAYES] bayes: fracreg
betareg	[BAYES] bayes: betareg

Survival regression models

streg	[BAYES] bayes: streg

Sample-selection regression models

heckman	[BAYES] bayes: heckman
heckprobit	[BAYES] bayes: heckprobit
heckoprobit	[BAYES] bayes: heckoprobit

Longitudinal/panel-data regression models

xtreg	[BAYES] bayes: xtreg
xtlogit	[BAYES] bayes: xtlogit
xtprobit	[BAYES] bayes: xtprobit
xtologit	[BAYES] bayes: xtologit
xtoprobit	[BAYES] bayes: xtoprobit
xtmlogit	[BAYES] bayes: xtmlogit
xtpoisson	[BAYES] bayes: xtpoisson
xtnbreg	[BAYES] bayes: xtnbreg

Multilevel regression models

mixed	[BAYES] bayes: mixed
metobit	[BAYES] bayes: metobit
meintreg	[BAYES] bayes: meintreg
melogit	[BAYES] bayes: melogit
meprobit	[BAYES] bayes: meprobit
mecloglog	[BAYES] bayes: mecloglog
meologit	[BAYES] bayes: meologit
meoprobit	[BAYES] bayes: meoprobit
mepoisson	[BAYES] bayes: mepoisson
menbreg	[BAYES] bayes: menbreg
meglm	[BAYES] bayes: meglm
mestreg	[BAYES] bayes: mestreg

Time-series models

var	[BAYES] bayes: var

DSGE models

dsge	[BAYES] bayes: dsge
dsgenl	[BAYES] bayes: dsgenl

参考：Stataマニュアル -> [BAYES] -> [Bayesian estimation]

目次

ベイズの定理と事後分布

A, Bを確率変数、$p(\cdot)$を確率密度(質量)関数とし、条件付き確率を $p(A|B) = \frac{p(A,B)}{p(B)}$ とする。

ベイズの定理： $p(B|A) = \frac{p(A|B)\,p(B)}{p(A)}$

統計モデルにおいて、未知パラメータ θ をもつ確率モデルからのサンプルとしてデータベクトル y が与えられているとする。ベイズ統計では、θ を確率変数と考える。$f(y_i|\theta)$を θ が与えられたもとでの y_i の確率密度関数として、尤度関数 $L(\theta; y) = f(y; \theta) = \Pi_{i=1}^{n} f(y_i|\theta)$ を使用する。

また、θ は確率分布$p(\theta) = \pi(\theta)$に従うとし、この分布を事前分布と呼ぶ。

このとき、ベイズの定理よりデータ y が与えられた下での θ の事後分布は次式となる。

$$p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{p(y)} = \frac{f(y;\theta)\,\pi(\theta)}{\int f(y;\theta)\,\pi(\theta)\,d\theta} \propto L(\theta; y)\,\pi(\theta)$$

事後分布　周辺尤度　尤度　事前分布

事後分布を求める方法

0. 事後分布: $p(\theta|y) \propto L(\theta; y)\,\pi(\theta)$

1. 事後分布から**サンプリング**する（MCMC法）

2. サンプルの**ヒストグラム**を描く

3. 描いた事後分布が適切か**診断**する

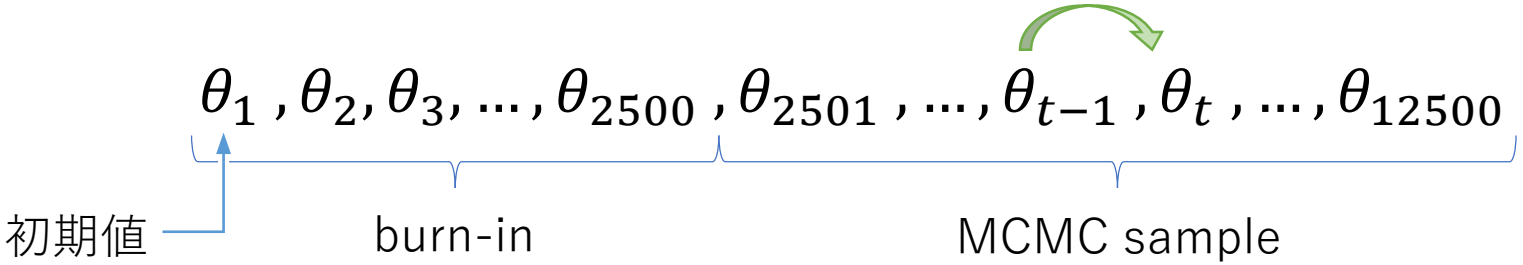
ポイント

・事後分布を関数として把握しているが、
分布の形状や統計量は不明。
・どのようにして分布を理解するか。

事後分布を求める方法

0. 事後分布: $p(\theta|y) \propto L(\theta; y)\,\pi(\theta)$

1. 事後分布から**サンプリング**する（MCMC法）



2. サンプルの**ヒストグラム**を描く

3. 描いた事後分布が適切か**診断**する

ポイント

- burn-inは初期値の影響を受けるため捨てられるサンプル。
- 初期値をもとにして2番目のサンプルが生成され、t-1番目のサンプルをもとにt番目のサンプルが生成される。
- 事後分布のサンプルとしてMCMC sampleが使用される。

事後分布を求める方法

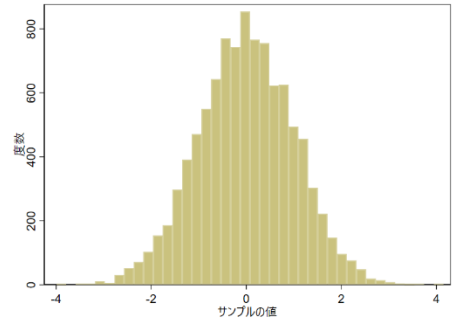
0. 事後分布: $p(\theta|y) \propto L(\theta; y)\,\pi(\theta)$

1. 事後分布から**サンプリング**する（MCMC法）

$$\theta_1, \theta_2, \theta_3, \ldots, \theta_{2500}, \theta_{2501}, \ldots, \theta_{t-1}, \theta_t, \ldots, \theta_{12500}$$

初期値　burn-in　MCMC sample

2. サンプルの**ヒストグラム**を描く



3. 描いた事後分布が適切か**診断**する

ポイント

・ヒストグラムを描いて形状を把握できた。
・得られた分布は適切な事後分布といえるか診断を行う。

事後分布を求める方法

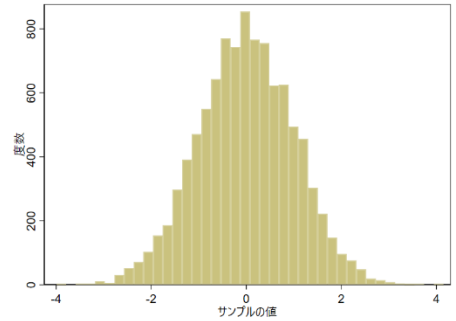
0. 事後分布: $p(\theta|y) \propto L(\theta; y)\,\pi(\theta)$

1. 事後分布から**サンプリング**する（MCMC法）

$$\theta_1, \theta_2, \theta_3, \ldots, \theta_{2500}, \theta_{2501}, \ldots, \theta_{t-1}, \theta_t, \ldots, \theta_{12500}$$

初期値　burn-in　MCMC sample

2. サンプルの**ヒストグラム**を描く



3. 描いた事後分布が適切か**診断**する

- 採択確率（t-1時点に留まらず候補点に移れているか。トレースプロットを確認）
- 自己相関（自己相関を持つため間引く必要有り。自己相関プロットを確認）
- 疑似収束（多峰性のある分布の場合に生じる問題。nchainsオプションを指定）

マルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo methods, MCMC法）

メトロポリス・ヘイスティング法（Metropolis–Hastings algorithm, MH法）

$q(\cdot)$を提案分布、θ_0を初期値とし、$t = 0,1,\dots,T-1$まで以下を繰り返す。

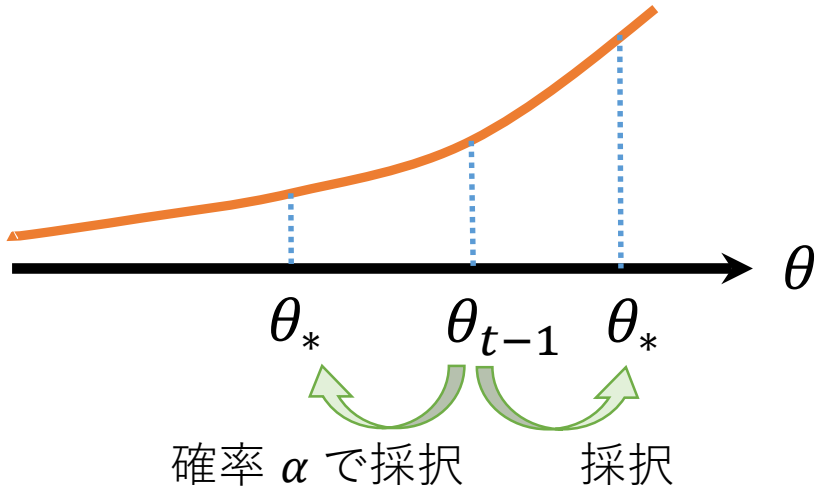
1. 提案分布 $q(\cdot \mid \theta_{t-1})$ から次の状態の候補 θ_*を生成する。
2. 採択確率 $\alpha(\theta_* \mid \theta_{t-1}) = min\{\frac{p(\theta_* \mid y)}{p(\theta_{t-1} \mid y)}, 1\}$ を算出する。（提案分布$q(\cdot)$ が対称な分布の場合）
3. 一様分布 $U(0,1)$ から u を生成する。
4. $u < \alpha(\theta_* \mid \theta_{t-1})$ なら、$\theta_t = \theta_*$ とする。満たさなければ、$\theta_t = \theta_{t-1}$ とする。

ギブスサンプリング（Gibbs sampling）

ギブスサンプリングは、MH法の特殊ケースとして考えられる。

パラメータにおける完全条件付き分布からのサンプリングが

得られるとき、 $q(\cdot \mid \theta_{t-1}^{\{-j\}})$ からの候補 θ_t^j は常に採択される。



共役な事前分布

未知パラメータθ	事前分布 $\pi(\theta)$	事後分布 $p(\theta\|y)$
ベルヌーイ分布の成功確率	ベータ分布	ベータ分布
正規分布の分散	逆ガンマ分布	逆ガンマ分布
ポアソン分布の期待値	ガンマ分布	ガンマ分布

- 事前分布と完全条件付き分布が同じ分布であるとき、準共役な事前分布という。
- パラメータが（準）共役な事前分布に従うときギブスサンプリングを適用できる。
- ギブスサンプリングにより低い計算負荷で、効率的にサンプリングできる。
- 上記の表は共役な事前分布の一例であり、その他にも複数ある。

例題

例題：bayesによる回帰モデル（1/3）

```
. bayes: regress wage age

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  wage ~ regress(xb_wage,{sigma2})

Priors:
  {wage:age _cons} ~ normal(0,10000)
          {sigma2} ~ igamma(.01,.01)                                        (1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_wage.
```

尤度

事前分布

例題：bayesによる回帰モデル（2/3）

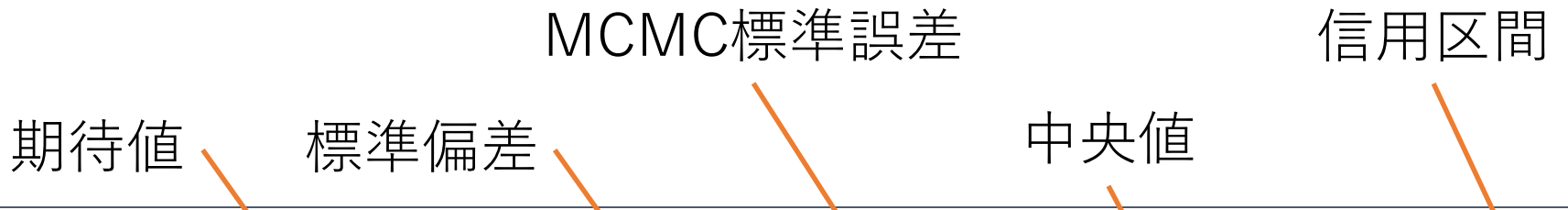
適用されたMCMC法の手法

サンプルサイズ

```
Bayesian linear regression                      MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling        Burn-in          =      2,500
                                                MCMC sample size =     10,000
                                                Number of obs    =        488
                                                Acceptance rate  =      .3739
                                                Efficiency:  min =      .1411
                                                             avg =      .1766
Log marginal-likelihood = -1810.1432                         max =      .2271
```

サンプリングの効率性を示す指標

例題：bayesによる回帰モデル（3/3）



	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred. interval]	
wage						
age	**.4008591**	**.0595579**	**.001586**	**.4005088**	**.2798807**	**.5183574**
_cons	**5.969069**	**1.737247**	**.043218**	**5.997571**	**2.60753**	**9.396475**
sigma2	**90.76252**	**5.891887**	**.123626**	**90.43802**	**79.71145**	**102.8558**

Note: Default priors are used for model parameters.

例題：bayesによる回帰モデル（ギブスサンプリング）（1/3）

```
. bayes, gibbs: regress wage age

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  wage ~ normal(xb_wage,{sigma2})

Priors:
  {wage:age _cons} ~ normal(0,10000)                                      (1)
          {sigma2} ~ igamma(.01,.01)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_wage.
```

ギブスサンプリングを指定するオプション

例題：bayesによる回帰モデル（ギブスサンプリング） (2/3)

```
Bayesian linear regression                      MCMC iterations  =     12,500
Gibbs sampling                                  Burn-in          =      2,500
                                                MCMC sample size =     10,000
                                                Number of obs    =        488
                                                Acceptance rate  =          1
                                                Efficiency:  min =          1
                                                             avg =          1
Log marginal-likelihood =  -1810.087            max =          1
```

効率的なサンプリング

例題：bayesによる回帰モデル（ギブスサンプリング） (3/3)

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred. interval]	
wage						
age	**.3999669**	**.0611328**	**.000611**	**.4005838**	**.2787908**	**.518693**
_cons	**6.012074**	**1.804246**	**.018042**	**6.000808**	**2.488816**	**9.549921**
sigma2	**90.84221**	**5.939535**	**.059395**	**90.54834**	**79.8132**	**103.0164**

Note: Default priors are used for model parameters.

効率が良いためサンプル数が多く、MH法と比べてMCMC標準誤差が小さい。

例題：bayesmhコマンドによる回帰モデル

```
. set seed 14

. bayesmh mpg weight, likelihood(normal({var})) ///
>         prior({mpg:}, normal(0,100)) ///
>         prior({var}, igamma(10,10)) blocksummary

Burn-in ...
Simulation ...
```

likelihoodオプションやpriorオプションを指定して、
尤度と事前分布を定める。
blocksummaryオプションを指定して、
ブロックに関する情報を表示する。

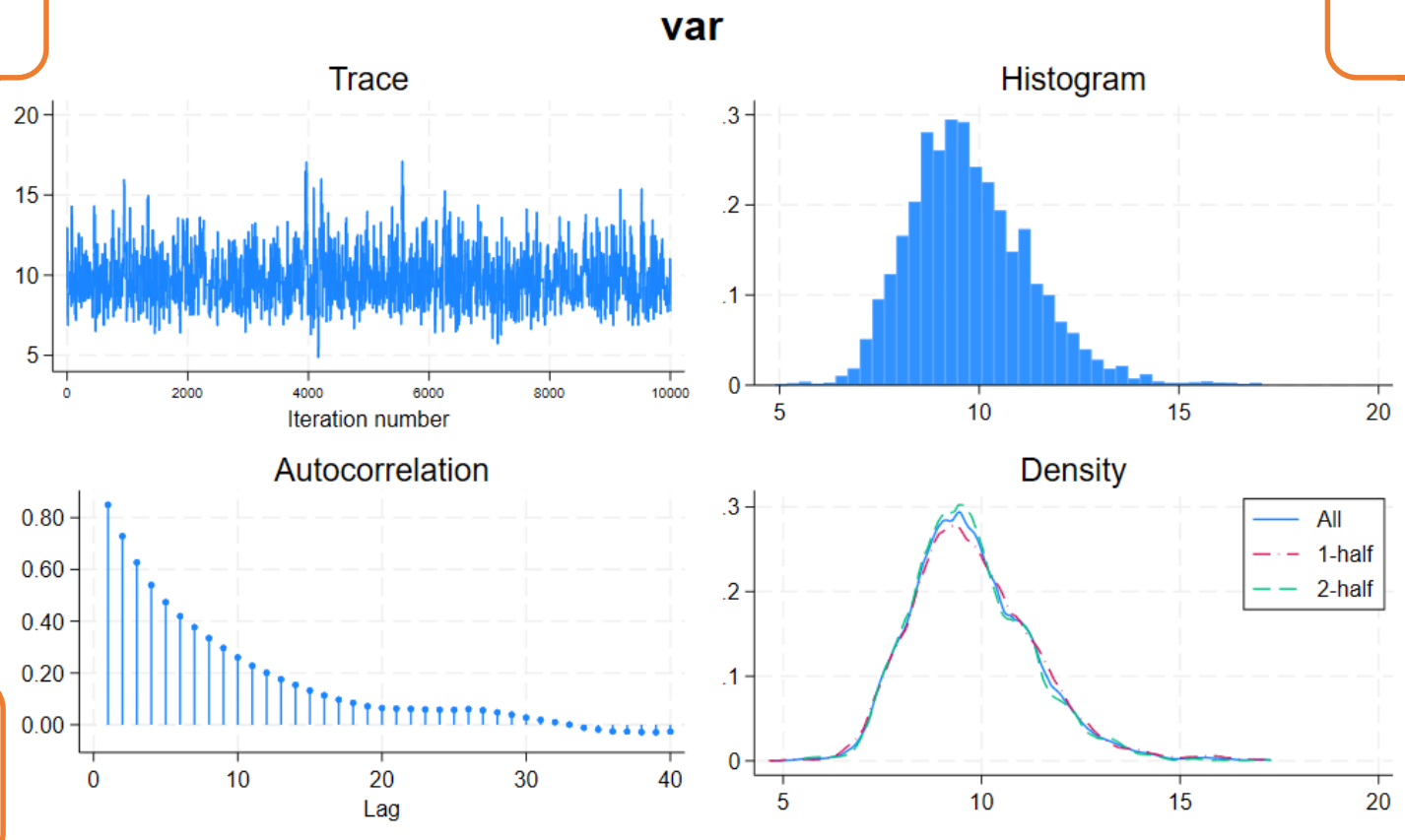
目次

診断プロット

```
bayesgraph diagnostics {var}
```

トレースプロット
trace

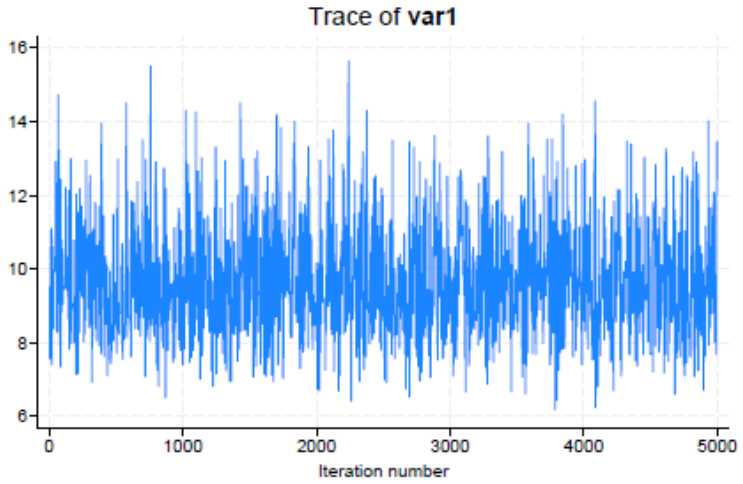
ヒストグラム
histogram

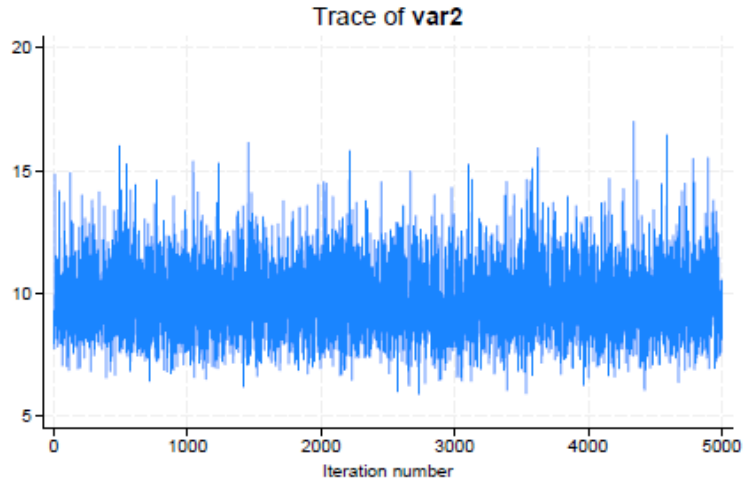


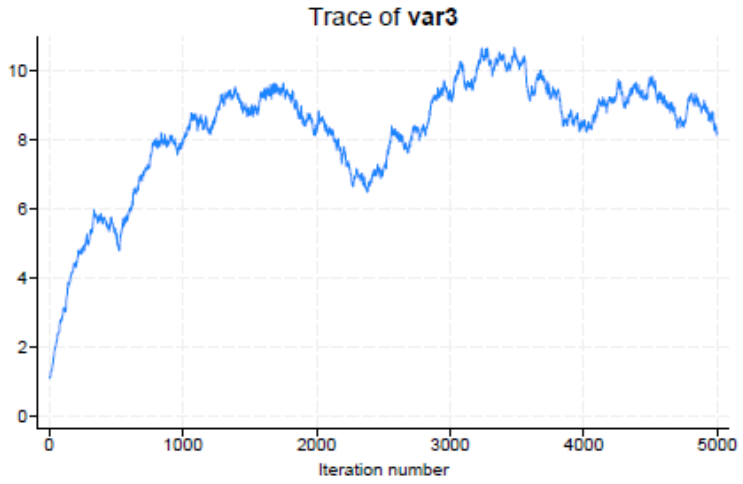
自己相関プロット
ac

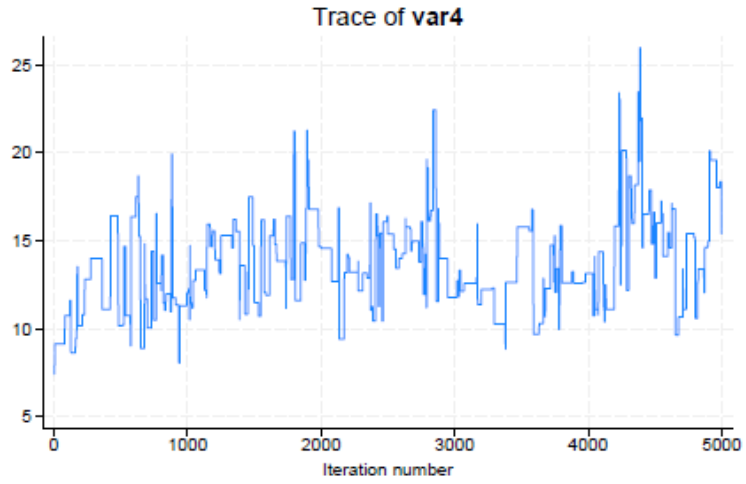
密度プロット
kdensity

トレースプロットの例



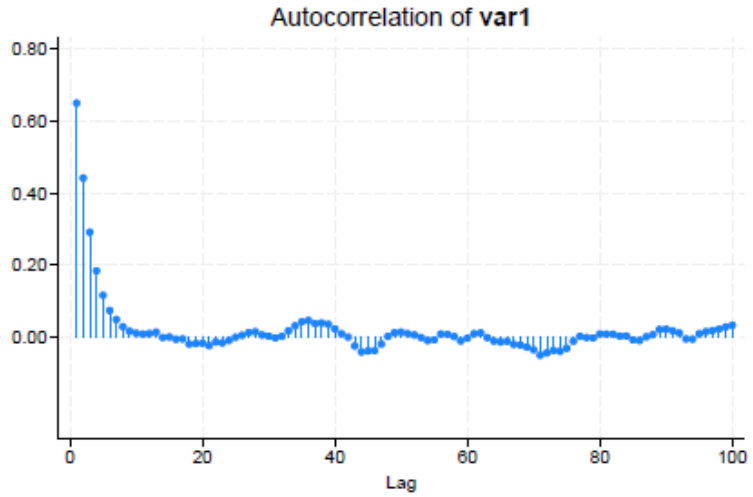


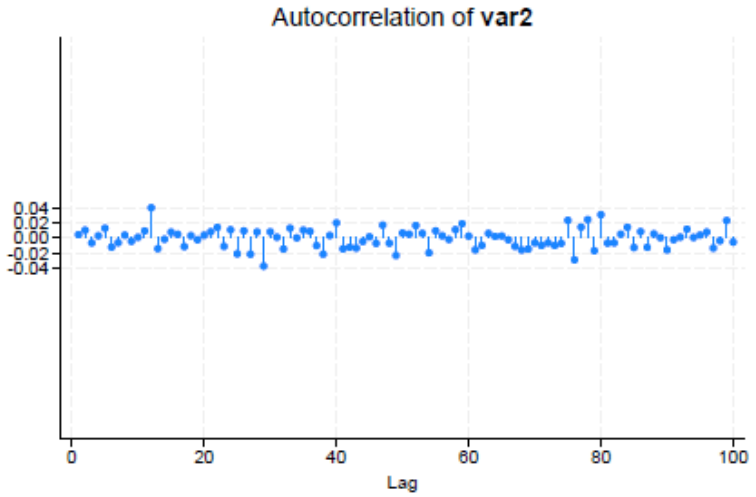


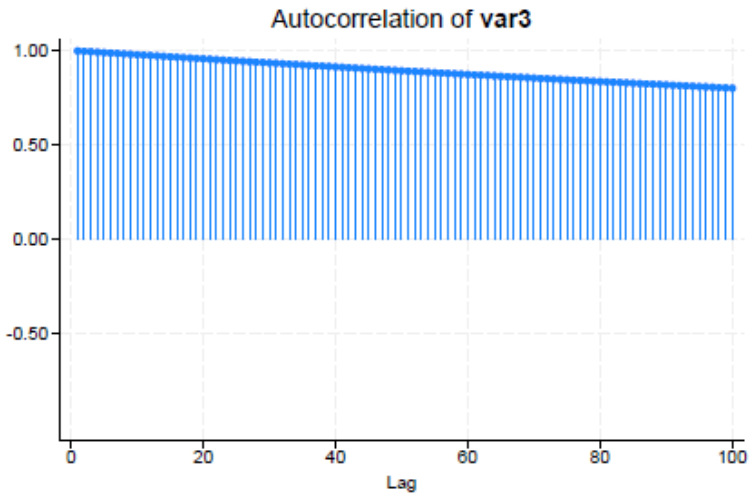


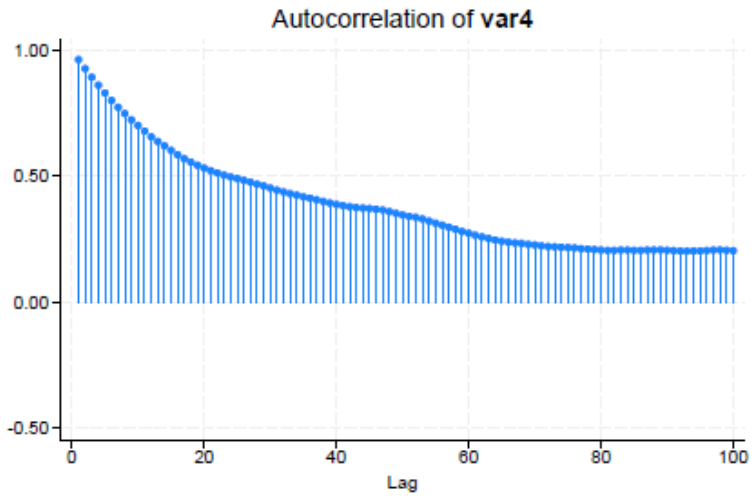
- [左上] 収束の**問題無し**
  - ランダムウォークMH法
  - 適度な採択率（約35%）
  - 効率性は10%から20%
- [右上] 収束の**問題無し**
  - ギブスサンプリング
  - 採択率は1に近い値
  - 効率性は95%以上
- [左下] 収束の**問題有り**
  - 混合が不十分
  - ランダムウォークMH法
  - 提案分布の分散が過小
- [右下] 収束の**問題有り**
  - 採択率が3%以下と過小
  - 提案分布の分散が過大

自己相関プロットの例







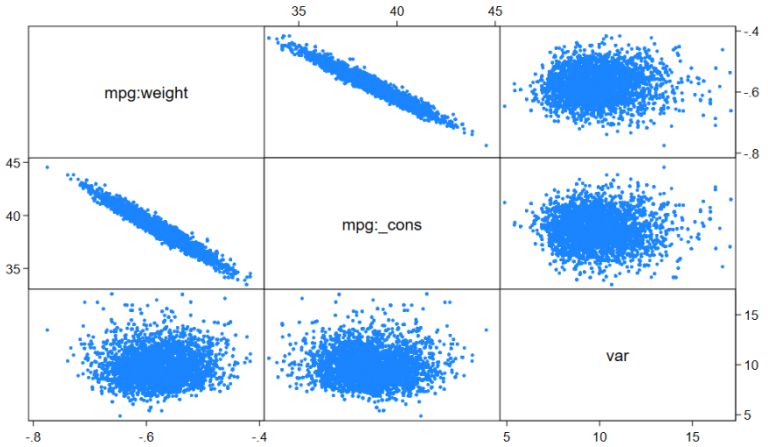


- [左上] 相関の**問題無し**
  - ラグ10で無視できる程度に
- [右上] 相関の**問題無し**
  - ラグに関係なく相関は低い
- [左下] 相関の**問題有り**
  - ラグ100でも高い自己相関
- [右下] 相関の**問題有り**
  - 左下ほどではないが、高い自己相関

ブロックオプションによるサンプリング効率の改善

例：3つのパラメータを持つ統計モデル

{param1}　{param2}　{param3}



パラメータの散布図行列
（bayesgraph matrix）を描画して
パラメータ間の相関関係を把握

ブロック 1

{param1}　{param2}

ブロック 2

{param3}

相関のあるパラメータ毎に
サンプリングすることで効率改善
（ブロックオプションで設定）

参考：Stataマニュアル -> [BAYES] -> [bayesmh] -> [Remarks and examples] -> [Improving efficiency of the MH sampling]

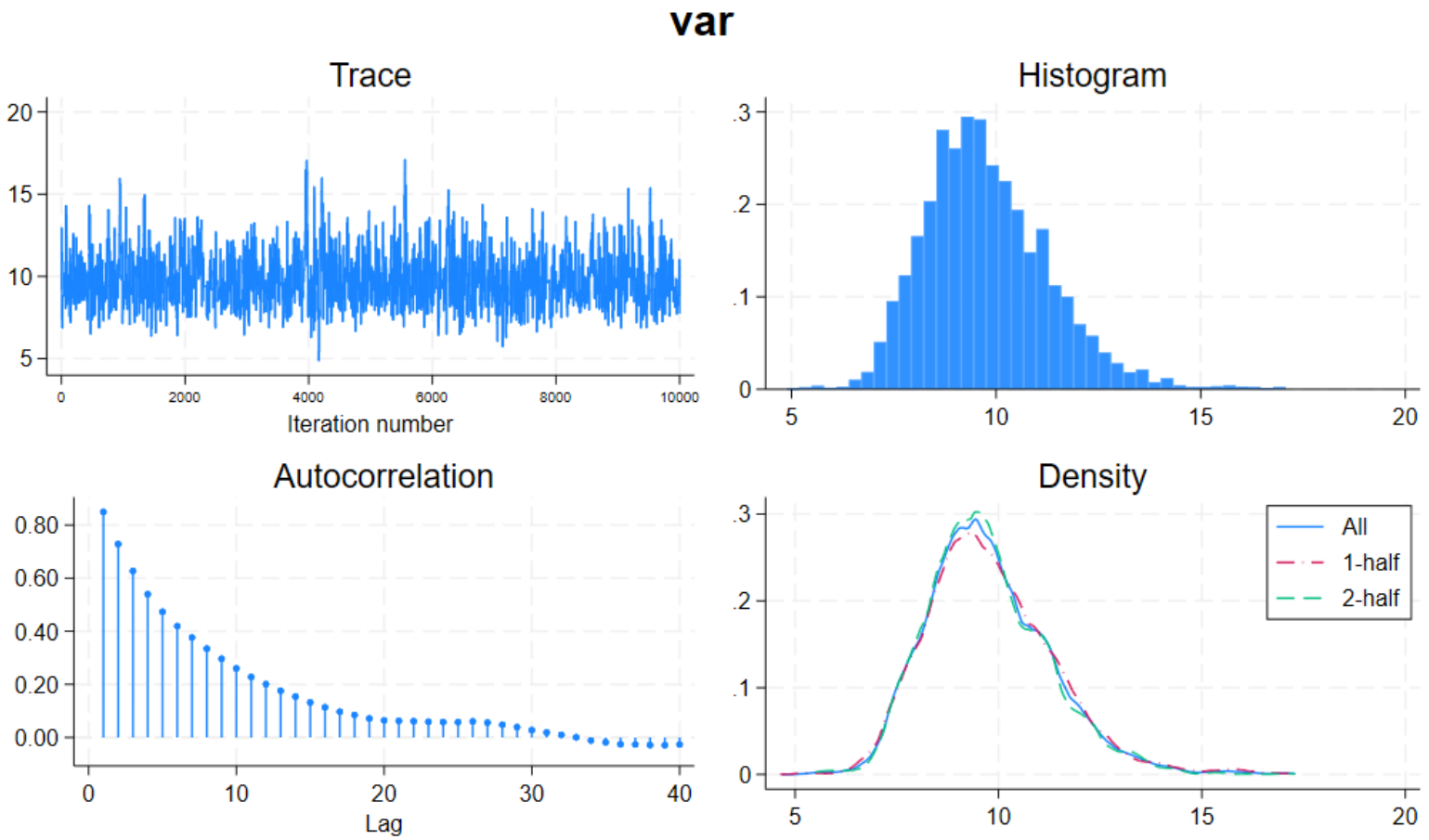
例題

ベイズ推定コマンドの実行

```
. set seed 14

. bayesmh mpg weight, likelihood(normal({var})) ///
>         prior({mpg:}, normal(0,100)) ///
>         prior({var}, igamma(10,10)) blocksummary

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  mpg ~ normal(xb_mpg,{var})

Priors:
  {mpg:weight _cons} ~ normal(0,100)                                      (1)
                {var} ~ igamma(10,10)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_mpg.
```

分散パラメータに注目

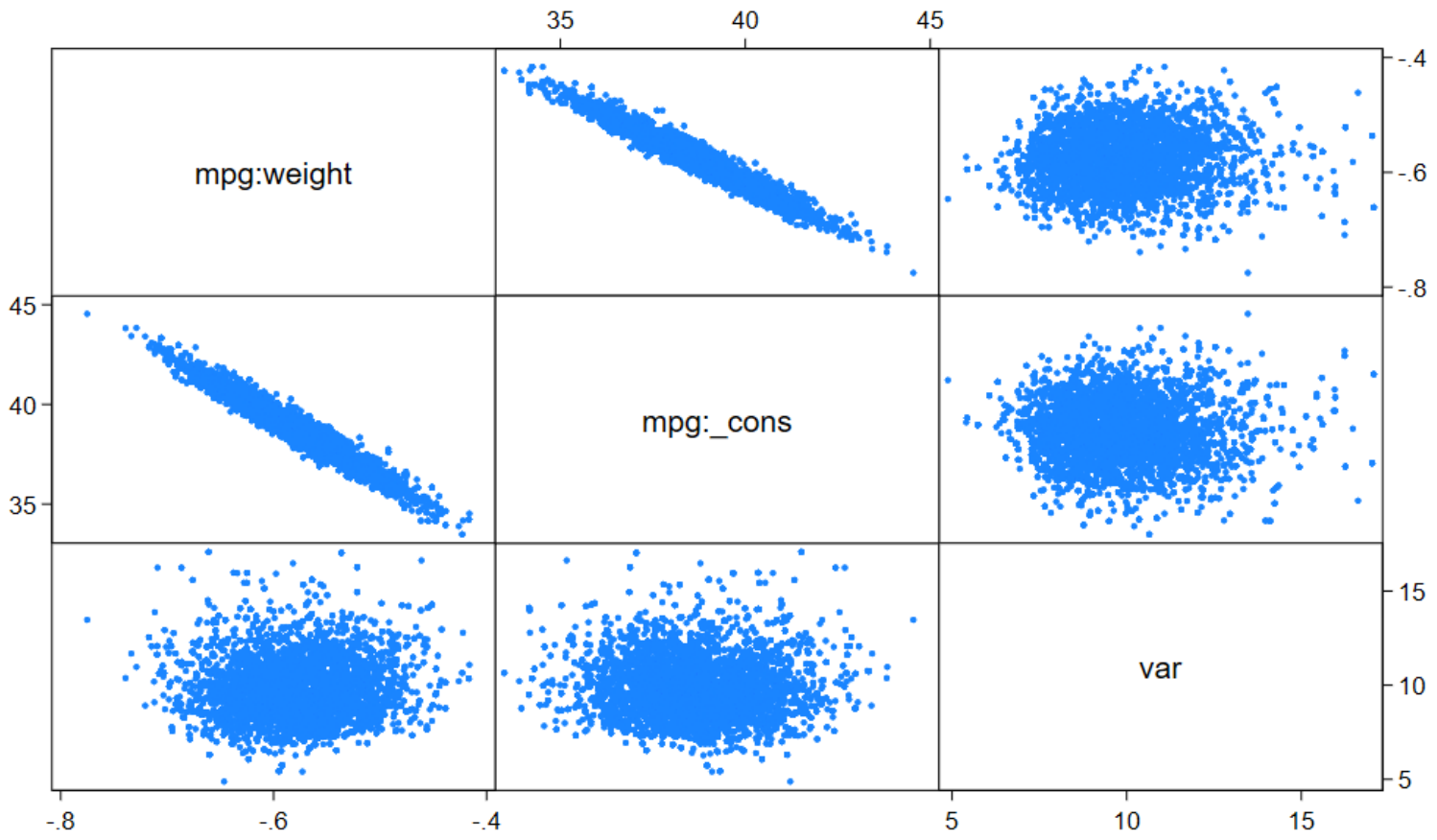
分散パラメータの診断プロット



- [左上] トレースプロット
  - 定義域を探索できている
- [右上]ヒストグラム
  - 逆ガンマ分布の形状
- [左下]自己相関プロット
  - **ラグ20**以降で低い相関
- [右下] 密度プロット
  - 逆ガンマ分布の形状

⇒ **収束の問題無し**

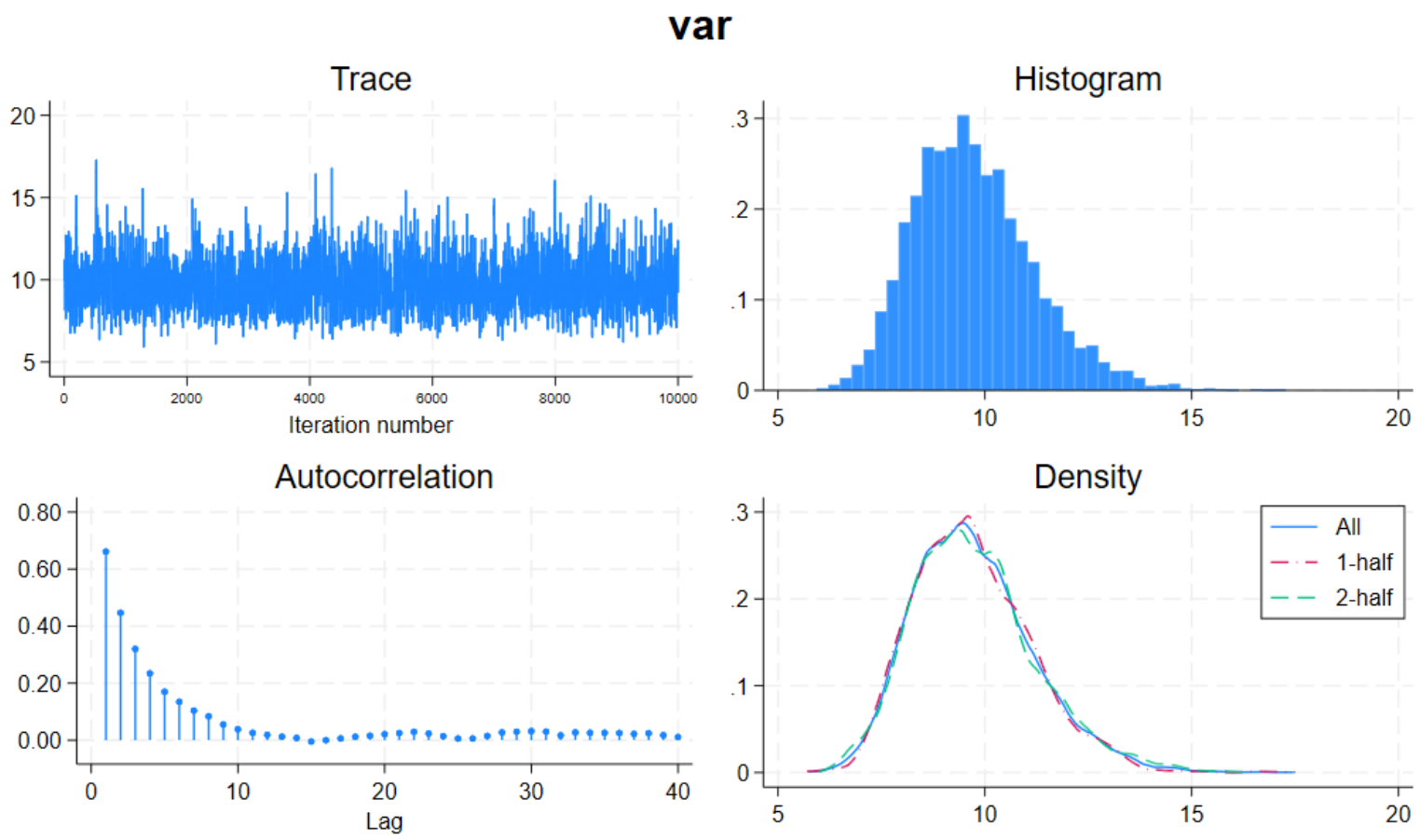
パラメータの散布図行列



- {mpg: weight} と {mpg: _cons} は相関関係がある
- {var} と 他2つのパラメータの間には相関が無い
- 相関の低いパラメータを分けてサンプリングを効率化できる

⇒ **ブロックオプションを指定**

ブロックオプションによる改善



- [左上] トレースプロット
  - 定義域を探索できている
- [右上]ヒストグラム
  - 逆ガンマ分布の形状
- [左下]自己相関プロット
  - **ラグ10**以降で低い相関
- [右下] 密度プロット
  - 逆ガンマ分布の形状

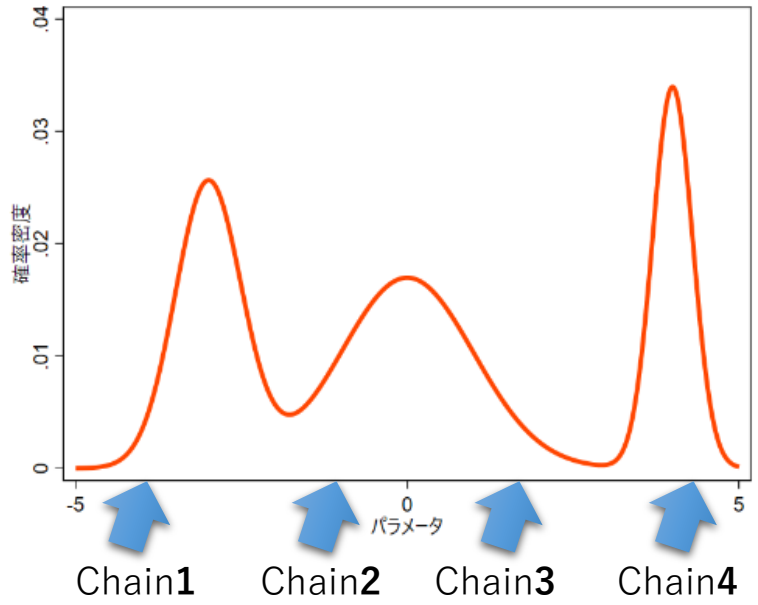
⇒ **ブロックオプションを追加して**
サンプリングを効率化

複数チェーンによる収束診断

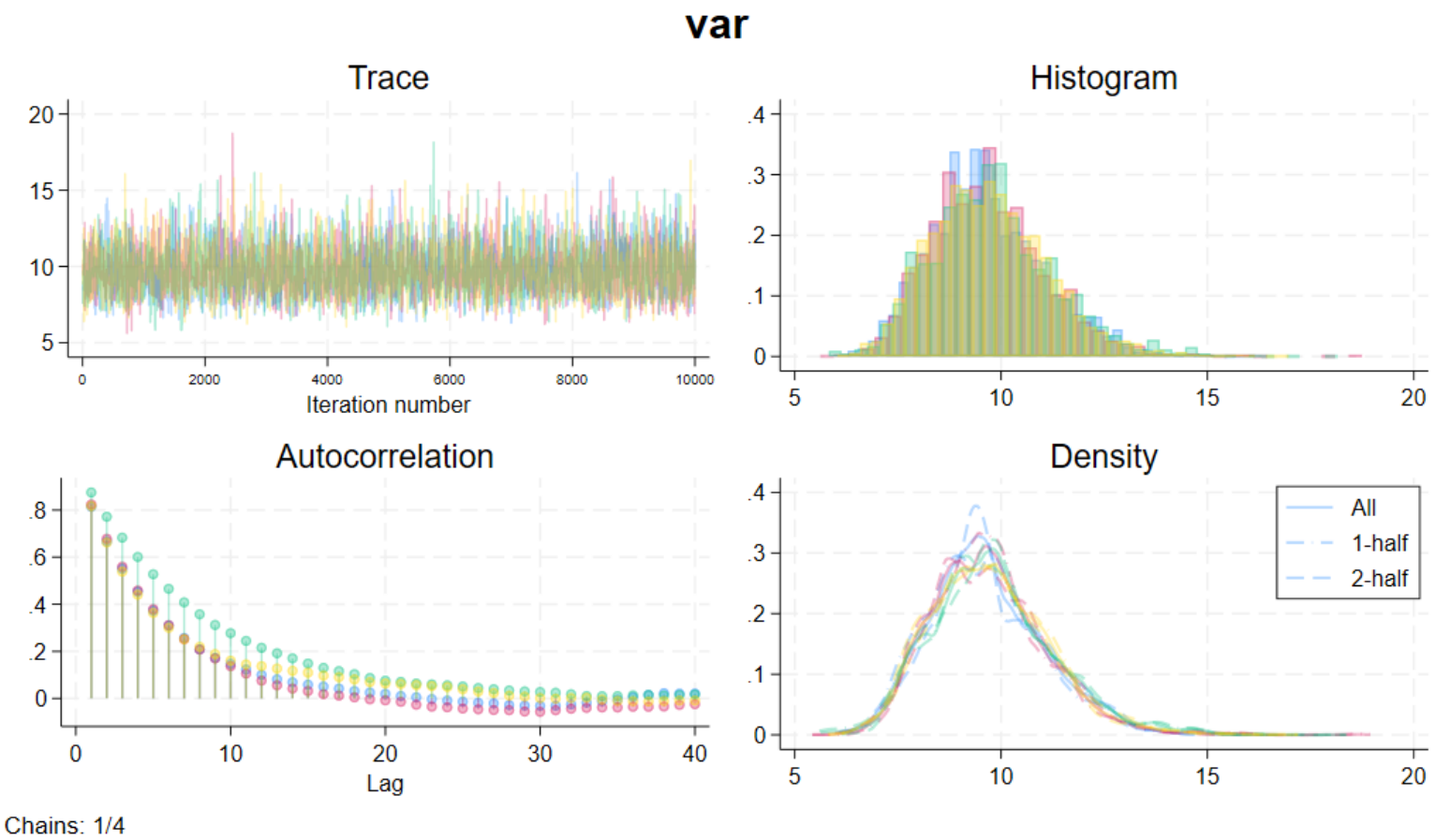
```
. bayesmh mpg weight, likelihood(normal({var})) ///
>                 prior({mpg:}, normal(0,100)) ///
>                 prior({var}, igamma(10,10)) ///
>                 nomodelsummary nchains(4) rseed(16)

Chain 1
  Burn-in ...
  Simulation ...

Chain 2
  Burn-in ...
  Simulation ...

Chain 3
  Burn-in ...
  Simulation ...

Chain 4
  Burn-in ...
  Simulation ...
```

分析の再現性を確保するため、
乱数のシードを設定する。

異なる初期値から
サンプリングを4回行う。



複数チェーンの診断プロット



- [左上] トレースプロット
  - 同じ範囲をトレースしている
- [右上]ヒストグラム
  - 重なり合っている
- [右下] 密度プロット
  - 重なり合っている

⇒ **疑似収束の問題は無い**

（視覚的に確認できる）

Gelman-Rubin 収束診断

```
Bayesian normal regression                      Number of chains     =          4
Random-walk Metropolis-Hastings sampling        Per MCMC chain:
                                                     Iterations      =     12,500
                                                     Burn-in         =      2,500
                                                     Sample size     =     10,000
                                                Number of obs        =         74
                                                Avg acceptance rate  =      .2275
                                                Avg efficiency: min  =      .07897
                                                                avg  =      .08265
                                                                max  =      .08827
Avg log marginal-likelihood = -226.73271        Max Gelman-Rubin Rc  =      1.002
```

bayesmhコマンドの実行結果では、
最大値のみ表示される。

bayesstats grubinコマンドで
詳細を表示する。

```
. bayesstats grubin

Gelman-Rubin convergence diagnostic

Number of chains      =          4
MCMC size, per chain  =     10,000
Max Gelman-Rubin Rc   =   1.002068
```

	Rc
mpg	
weight	**1.000783**
_cons	**1.000557**
var	**1.002068**

Convergence rule: Rc < 1.1

目次

Stataにおけるベイズ推定

推定コマンド

- bayesmh
- bayes

オプション

- likelihood
- prior
- gibbs
- block
- nchains
- rseed

事後分布 $\propto$ 尤度 × 事前分布

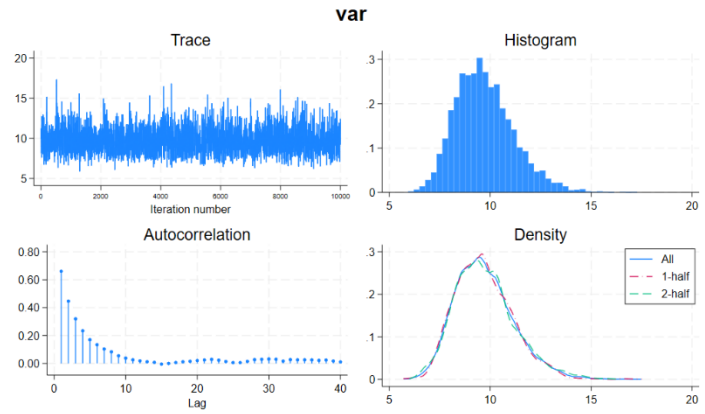
（MCMC法：MH , Gibbs）

$$\theta_1, \theta_2, \theta_3, \ldots, \theta_{t-1}, \theta_t, \ldots, \theta_T$$

burn-in　　MCMC sample

推定後コマンド

- bayesgraph diagnostics
- bayesgraph matrix
- bayesstats grubin



お知らせ

◆ Stataマニュアル

[ヘルプ] -> [英文PDFマニュアル]　または

https://www.stata.com/features/documentation/

◆ 評価版のご案内

弊社HPより評価版をご利用いただけます。

（学生の方は購入のご案内となります。）

https://www.lightstone.co.jp/stata/evaluate.html