mwp-239

ベイズ分析系コマンドの紹介　【 評価版 】

本 whitepaper ではベイズ分析系コマンドに関する基本的な用法を説明します。ベイズ分析は未知のパラメータに関する不確かさを確率によって表現できる推定手法です。その場合、帰結変数のみならず、統計モデル中に含まれる未知のパラメータもすべてランダムで、ある事前分布に従う変量であるとする考え方が前提となる点に注意してください。

ベイズ論的推定を行うには本 whitepaper で記述する `bayesmh` コマンドを用いるのが正攻法と言えますが、既存の推定コマンドに `bayes` プリフィックスを付加するアプローチも用意されています。この `bayes` プリフィックス機能については `[BAYES] bayes` (*mwp-287*) を参照ください。

1. ベイズ分析系コマンド

bayesmh コマンドはベイズ分析用コマンド群の中核をなすコマンドです。それは種々のベイズ論的回帰モデルをフィットさせ、adaptive な MH (Metropolis-Hastings) MCMC (Markov chain Monte Carlo) 法を用いてパラメータの推定を行います。各種ベイズモデルの選択は likelihood() オプションと prior() オプションの指定によって行えますが、独自のプログラムを組み込む形での推定も可能です。そのために用意されているのが evaluator() オプションですが、その詳細については [BAYES] **bayesmh evaluators** を参照ください。

ベイズ論的変数選択 (Bayesian variable selection) を実行するには bayesselect コマンドを使用します。それは 2 クラスの縮小事前分布 (shrinkage priors) — global-local shrinkage priors と spike-and-slab priors — を用意しています。詳細については [BAYES] **bayesselect** を参照ください。

推定実行後 MCMC の収束を視覚的にチェックするには bayesgraph コマンドを使用します。多重連鎖 (multiple chains) のシミュレーションを行った場合には bayesstats grubin コマンドを用いることによって Gelman-Rubin 収束診断情報を算出することができます。また bayesstats ess コマンドを使用すると、有効標本サイズ (effective sample sizes) やその他の統計量を算定することができます。一旦収束が確認された段階では、bayesstats summary コマンドを使用することによってモデルパラメータに関する事後平均や標準偏差等を算出することができる他、bayesstats ic コマンドを使用すればベイズ情報量基準 (Bayesian information criteria) やベイズ因子 (Bayes factors) の算出が可能です。また事後確率を比較することによって仮説検定を行う bayestest model コマンドや区間に関する仮説検定のための bayestest interval コマンド等も用意されています。また bayespredict と bayesstats ppvalues コマンドを使うと事後予測チェッキングの機能を用いてモデルの評価を行うことができます。

以下においてはベイズ分析系コマンドに関する基本的な用例を紹介しますが、その他の用例については [BAYES] **bayesmh** (*mwp-240*) を参照ください。

2. 基本的な用例

ここでは Example データセット oxygen.dta を用いてベイズ分析系コマンドの用例を紹介します。このデータセットには最大酸素摂取量に関する 2 種類のエクササイズ — step aerobics 12 週間と平坦地での outdoor running 12 週間 — の効果が記録されています。調査に際しては 12 人の健康な男性がランダムに“aerobic”か“running”のいずれかにアサインされ、酸素の最大換気量（*l*/ 分）がエクササイズの前後でどう変化したかが計測されています。

```
. use https://www.stata-press.com/data/r19/oxygen.dta [*1]
(Oxygen uptake data)
```

[*1] メニュー操作：File ▷ Example Datasets ▷ Stata 19 manual datasets と操作、Bayesian Analysis Reference Manual [BAYES] の Bayesian commands の項よりダウンロードする。

```
. list, separator(6)
```

	change	group	age	ageXgr
1.	-.87	Running	23	0
2.	-10.74	Running	22	0
3.	-3.27	Running	22	0
4.	-1.97	Running	25	0
5.	7.5	Running	27	0
6.	-7.25	Running	20	0
7.	17.05	Aerobic	31	31
8.	4.96	Aerobic	23	23
9.	10.4	Aerobic	27	27
10.	11.05	Aerobic	28	28
11.	.26	Aerobic	22	22
12.	2.51	Aerobic	24	24

キーとなる変数は change で最大換気量がどれだけ変化したかを表します。グループの種別は変数 group に (0: running, 1: aerobic)、被験者の年齢は変数 age に記録されています。4 番目の変数 ageXgr は交互作用 age#group を表しています。

Kuehl (2000) は ANCOVA (analysis of covariance) を用いて分析を行っていますが、ここでは線形回帰

$$\texttt{change} = \beta_0 + \beta_{\text{group}}\,\texttt{group} + \beta_{\text{age}}\,\texttt{age} + \epsilon$$

を使用することにします。ただし ϵ は平均が 0、分散が σ^2 のランダムな誤差を意味します。なお、このデータを用いたベイズ分析については Hoff (2009) も参考になります。

▷ Example 1: OLS

最初に OLS (ordinary least squares) 回帰モデルを regress コマンドによりフィットさせます。

```
. regress change group age [*2]
```

```
. regress change group age

      Source |       SS           df       MS      Number of obs   =        12
-------------+----------------------------------   F(2, 9)         =     41.42
       Model |  647.874893         2  323.937446   Prob > F        =    0.0000
    Residual |   70.388768         9  7.82097423   R-squared       =    0.9020
-------------+----------------------------------   Adj R-squared   =    0.8802
       Total |  718.263661        11  65.2966964   Root MSE        =    2.7966
```

[*2] メニュー操作： Statistics ▷ Linear models and related ▷ Linear regression

change	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
group	5.442621	1.796453	3.03	0.014	1.378763	9.506479
age	1.885892	.295335	6.39	0.000	1.217798	2.553986
_cons	-46.4565	6.936531	-6.70	0.000	-62.14803	-30.76498

結果からすると group, age 共にアウトカムに対する有意な予測変数であると言えます。

例えば group について言うなら、p 値が 0.014 ですから、有意水準を 5% としたときに $H_0 : \beta_{\text{group}} = 0$ という帰無仮説は棄却されるわけです。この p 値の意味するところは、同じ手続きでの実験を多数繰り返したときに、グループによる効果はないとする帰無仮説が正しいとするなら、この標本で観測されたような $t = 3.03$、あるいはそれ以上に極端な検定統計量が得られる率は 1.4% でしかないということです。良く誤解されるのですが、p 値を帰無仮説に対する確率と解釈するのは誤りです。実際、それは帰無仮説が真であるとしたときに、手元の標本がどの程度起りやすいかという問いに答えるものです。データが所与のときに、帰無仮説がどの程度もっともらしいかという問いに答えるものではありません。この後者の問いに対してはベイズ論的な仮説検定によって答えることができます（Example 9 参照）。

信頼区間 (CI: confidence intervals) は p 値の持つ弱点を補うものとして良く用いられます。例えば group に対する係数値について言うなら、その 95% CI は $[1.38, 9.51]$ とレポートされています。この区間値には 0 が含まれていないので、group は change に対する有意な予測変数であると判断されるわけです。この 95% CI の意味するところは、同一の実験を多数繰り返し、その都度 CI の計算を行ったとしたときに、それらの区間の 95% はパラメータの真の値を含むであろうとするものです。group に対する真の係数値が $[1.38, 9.51]$ に入る確率が 0.95 であるとする解釈は誤りです。この確率は 0 か 1 かのいずれかであり、それがどちらであるかを特定の CI について知ることはできません。言えることは group に対する真の係数値にとって $[1.38, 9.51]$ がもっともらしい範囲であるということだけです。これに対しベイズ分析の考え方に立つなら、真のパラメータ値をある確率で含むと解釈できる区間を求めることが可能です（Example 9 参照）。 ◁

▷ Example 2: **ベイズ論的線形回帰** (1)

Example 1 で述べたように、頻度論的手法ではパラメータに関する確率論的言明は得られません。頻度論的手法の場合、パラメータは未知ではあるが固定的な量であるという考え方が基盤にあるからです。頻度論的モデルにおいてランダムな量は帰結（アウトカム）変数のみです。これに対しベイズ統計においては帰結変数のみならず、すべてのモデルパラメータもランダムな量とみなされます。この点が頻度論的統計と一線を画するところであり、パラメータ値や仮説についての確率的言明を可能にする点でもあります。

ベイズ統計では事後分布の持つ種々の側面の推定に焦点が当てられますが、その際、観測データ中に含まれる情報によって更新された事前分布が前提となります。すなわち事後分布は、あるパラメータの事前分布、及びそのパラメータが与えられたときのデータの尤度関数とによって記述されるものと言えます。

それでは oxygen.dta に対してベイズ論的線形回帰モデルをフィットさせてみましょう。ベイズ論的パラメトリックモデルをフィットさせるためにはデータの分布を表す尤度関数と、すべてのモデルパラメータに対する事前分布とを指定する必要があります。Example 1 で想定した線形モデルにおいては 4 つのパラメータ — 3 つの回帰係数、及びデータの分散 — が存在します。ここでは帰結変数 change の分布として正規分布を、

その他の係数については noninformative な Jeffreys prior を仮定することにします。この Jeffreys prior の場合、係数と分散の結合事前分布 (joint prior distribution) は分散の逆数に比例することになります。

このモデルを数式で表記すると次のようになります。

$$\texttt{change} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$$
$$(\boldsymbol{\beta}, \sigma^2) \sim \frac{1}{\sigma^2}$$

ただし $\mathbf{X}$ は計画行列 (design matrix) を、$\boldsymbol{\beta}$ は係数ベクトル $(\beta_0, \beta_{\text{group}}, \beta_{\text{age}})'$ を意味します。

このモデルをフィットさせるには bayesmh コマンドを使用するわけですが、最初にモデル式の記述方法について見て行くことにします。

```
. bayesmh change group age, likelihood(normal({var}))    ///
        prior({change:}, flat) prior({var}, jeffreys)
```

回帰式の指定方法は他の回帰系コマンドの場合と変りません。すなわち従属変数 change と共変量 group, age の並びをコマンド名 bayesmh に続く形で指定します。一方、帰結変数の分布を意味する尤度は likelihood() オプションにより、事前分布は複数の prior() オプションにより指定します。

すべてのモデルパラメータが {} の中で指定されなくてはなりません。ただし回帰係数については bayesmh が自動生成するので明示は不要ですが、残りのモデルパラメータについてはユーザによる明示が必要となります。ここでの例で言うなら分散パラメータが {var} という形で規定されているわけです。3 つの回帰変数である {change:group}, {change:age}, {change:_cons} については省略形での指定が用いられています。

最後に必要となるのが尤度と事前分布の指定です。通常は bayesmh が用意している built-in の分布を使用するわけですが、事後分布を評価するための独自プログラム (evaluators) を利用することも可能です。詳細は [BAYES] **bayesmh evaluators** を参照ください。上記のコマンドでは likelihood() オプションの中で normal({var}) という指定を行っていますが、これによって分散パラメータを {var} とする形の正規分布用尤度関数の使用が指示されたことになります。この指定と回帰式の記述の双方により帰結変数 change に対する尤度モデルが規定されます。3 つの回帰係数に対しては prior({change:}, flat) という指定があるので、密度 1 の flat prior が仮定されることになります。{change:} というのは省略形で方程式名しか指定されていないわけですが、その場合にはそれを構成するすべての回帰変数が指定されたものとして解釈されます。分散パラメータ {var} については prior jeffreys という指定が行われていますが、これによって密度 $1/\sigma^2$ が仮定されることになります。

それでは実際に bayesmh コマンドを実行させてみます。コマンドは上記の通りですが、ここではダイアログインタフェースでの操作を紹介します。なお、bayesmh コマンドは MCMC 法を用いて事後分布の推定を行うため、最初にまず乱数発生用の seed を設定します。結果の再現性にこだわらないのであれば seed の設定はスキップできます。

```
. set seed 14
```

- Statistics ▷ Bayesian analysis ▷ General estimation and regression と操作
- `bayesmh` ダイアログ: Model タブ: Class of models: Linear models
  Model type: Univariate linear regression
  Model: Dependent variable: `change`
  Independent variables: `group age`
  Likelihood model: Continuous –> Normal regression
  Variance: {`var`}
  Priors for model parameters: ⇒ Create...

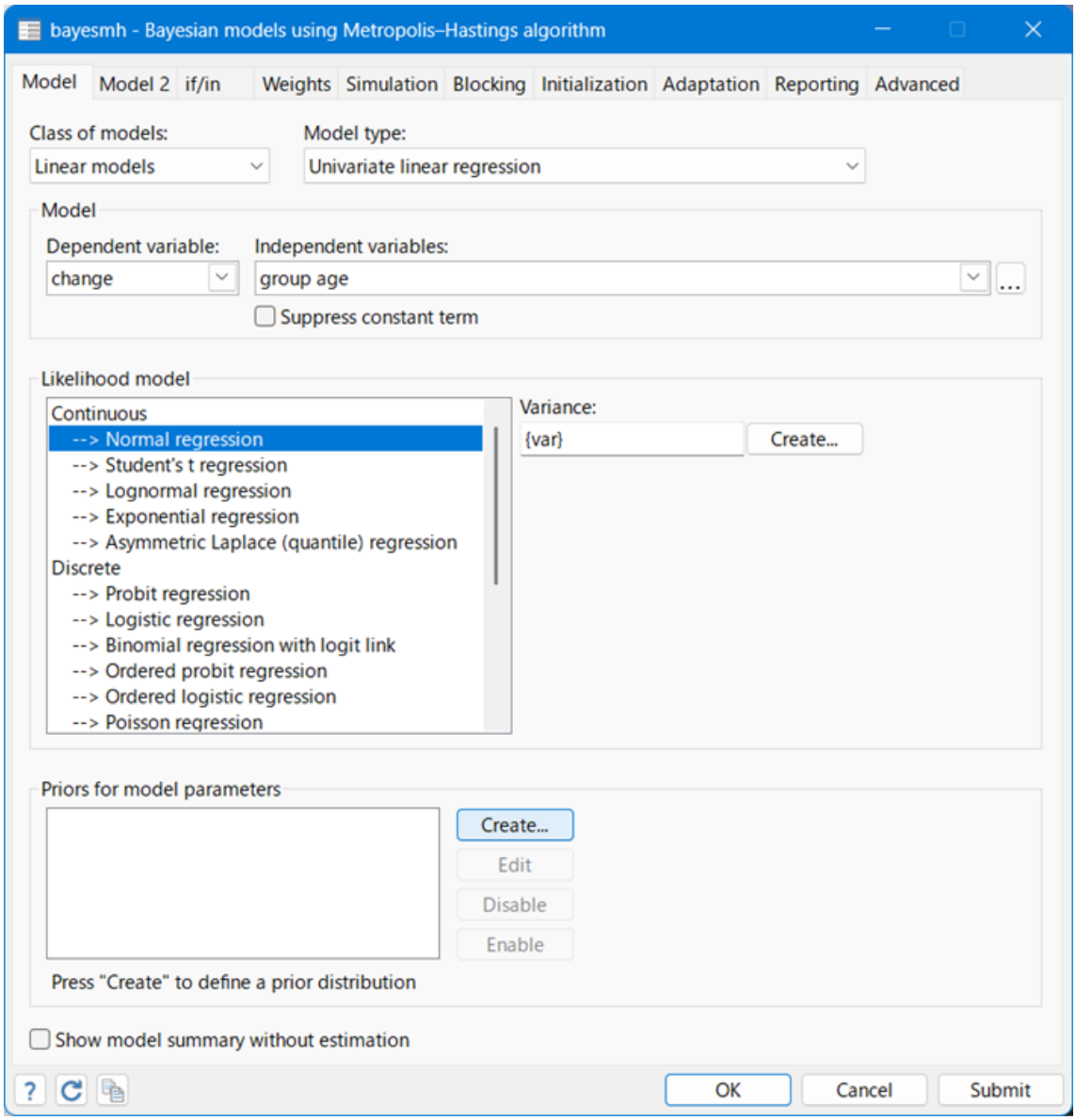

図 1 `bayesmh` ダイアログ - Model タブ

- Prior 1 ダイアログ: Parameters specification: {change:}
  Choose a prior distribution: Generic –> Flat prior

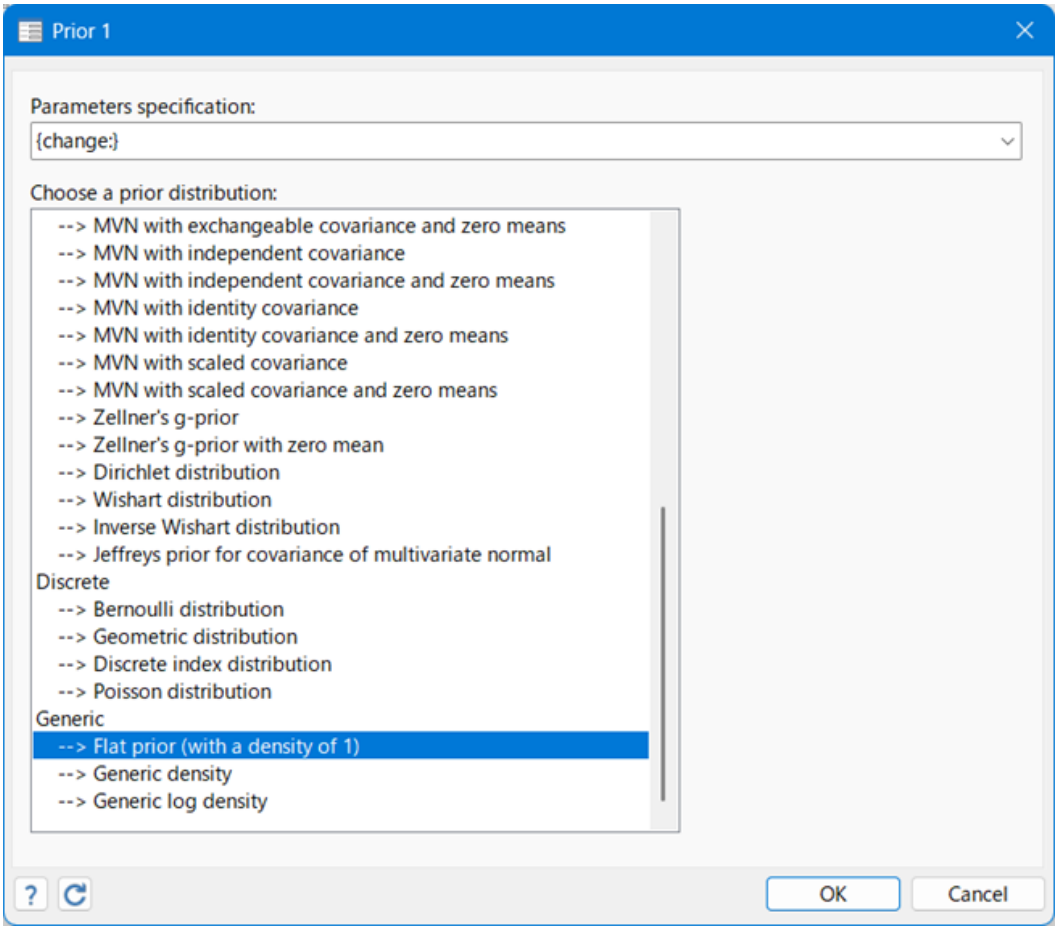

図 2 Prior 1 ダイアログ

- bayesmh ダイアログ: Model タブ:
  Prior 2: Parameters specification: {var}
  Choose a prior distribution:
  Univariate continuous –> Jefreys prior for variance of normal distribution

```
. bayesmh change group age, likelihood(normal({var})) prior({change:}, flat) prio
> r({var}, jeffreys)

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  change ~ normal(xb_change,{var})

Priors:
  {change:group age _cons} ~ 1 (flat)                                   (1)
                     {var} ~ jeffreys
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_change.

Bayesian normal regression                      MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling        Burn-in          =      2,500
                                                MCMC sample size =     10,000
                                                Number of obs    =         12
                                                Acceptance rate  =      .1371
                                                Efficiency:  min =     .02687
                                                             avg =     .03765
Log marginal-likelihood = -24.703776                         max =     .05724

------------------------------------------------------------------------------
             |                                                Equal-tailed
             |      Mean   Std. dev.     MCSE     Median  [95% cred. interval]
-------------+----------------------------------------------------------------
change       |
       group |   5.429677   2.007889   .083928   5.533821   1.157584   9.249262
         age |     1.8873   .3514983   .019534   1.887856   1.184714   2.567883
       _cons |  -46.49866    8.32077   .450432   -46.8483  -62.48236  -30.22105
-------------+----------------------------------------------------------------
         var |   10.27946   5.541467   .338079   9.023905   3.980325   25.43771
------------------------------------------------------------------------------
```

bayesmh からの出力の先頭部には設定されたモデルの要旨が示されています。モデルが複雑化した場合には、この内容をチェックしておくことが重要となります。実際、dryrun オプションの機能を利用すれば、推定を行うことなく、このモデルサマリだけを出力させることができます。モデルの確認が済んだ段階では nomodelsummary オプションを指定することによって、モデルサマリを抑止した形での推定結果を得ることも可能です。

モデルサマリの次には推定プロセスに関するヘッダ情報が出力されてきます。MCMC の反復が 12,500 回行われたこと、そのうち最初の 2,500 回は捨てられ (burn-in iterations)、残りの 10,000 回が MCMC サンプルとして利用されたことが示されています。これらのデフォルト値に基づく推定は初期推定ととらえるべきでしょう。MCMC の収束をより確実なものとするためにさらなる調整が加えられることもあります（Example 5 参照)。

ヘッダ部には受容率 (acceptance rate) と効率 (efficiency) に関する情報も含まれています。受容率というのは提案されたパラメータ値のうちアルゴリズムによって受容されたものの比率を言います。ここでの例では受容率が 0.14 と出力されていますが、それは 10,000 回の提案値のうちで 14% がアルゴリズムによって受容されたことを意味しています。MH アルゴリズムの場合、この値が 50% を越すことは稀で、通常は 30% 以下の値となります。受容率の値が低い場合（例えば 10% に満たない値の場合）には収束上の問題が示唆されている可能性があります。ここで得られた 14% という率は少々小さい値であるのでさらなる検討が必要と言えるでしょう。一方、効率[*3]について言うなら、MH アルゴリズムの効率は他の MCMC 手法に比べて低いものとなる傾向があります。例えば効率が 10% 以上の値であれば良好と言え、1% 以下であれば問題ありと考えられます。ここでの効率の値はやや低めであるので、MCMC sampler のチューニングを検討すべきかも知れません。詳細については `[BAYES]` **bayesmh** (*mwp-240*) のセクション“*Improving efficiency of the MH algorithm—blocking of parameters*”を参照ください。

ヘッダに続く形で推定結果を集約したテーブルが出力されてきます。`Mean` カラムには事後平均 (posterior means) の推定値 — 該当パラメータに関する周辺事後分布 (marginal posterior distributions) の平均値 — が表示されています。これらは Example 1 における OLS 推定値にかなり近い値となっています。これは noninformative な — パラメータ値に関しデータ中に含まれているもの以上の情報をもたらさない — prior を用いた場合に期待される結果であると言えます（ただし MCMC の収束が前提）。

次のカラムに示されているのは事後標準偏差 (posterior standard deviations) の推定値 — 周辺事後分布の標準偏差 — です。これらの値は当該パラメータに関する事後分布のばらつき度合を表現するもので、OLS における標準誤差に相当するものと言えます。

事後平均推定値の精度は MC 標準誤差 (MCSE: Monte Carlo standard errors) として表現されています。これらの値はパラメータ値との相対において小さいに越したことはありません。MCMC サンプルサイズを増やせばこれらの値はより小さなものとなります。

`Median` カラムには事後分布のメディアン推定値が示されているわけですが、これは事後分布の対称性を評価する上で参考になります。ここでの例で言えば、事後平均とメディアンの推定値は類似した値であるため、これら回帰係数に対する事後分布は対称形であると推察できます。

最後の 2 つのカラムには確信区間 (credible intervals) の情報が表示されています。Example 1 で述べたようにこれらの区間は信頼区間 (confidence intervals) とは異なり、直接的な確率に基づく解釈を可能にします。例えば `group` に対する係数値が $[1.16, 9.25]$ の範囲にある確率は約 95% であるということができるわけです。この場合、下限値は 0 よりも大きな値であるため、エクササイズプログラムの選択は酸素摂取量の変化に影響を持つと結論付けることができます。同様の検定はベイズ論的仮説検定の機能を用いても実行できます (Example 9 参照)。ただし結果の解釈に先立ち MCMC の収束を確認することを忘れてはいけません。これについては Example 5 を参照ください。

`bayes` prefix を用いるとベイズ論的線形回帰がより簡便に実行できるわけですが、それについては Example 11 を参照ください。 ◁

[*3] 混合の質 (mixing quality) に関係する評価尺度。

▷ Example 3: **ベイズ論的線形回帰** (2)

評価版では割愛しています。

▷ Example 4: **ベイズ論的線形回帰** (3)

評価版では割愛しています。

▷ Example 5: **収束の確認**

評価版では割愛しています。

▷ Example 6: bayesstats summary **コマンド**

評価版では割愛しています。

▷ Example 7: **ベイズ論的予測機能**

評価版では割愛しています。

▷ Example 8: **モデル比較**

評価版では割愛しています。

▷ Example 9: **仮説検定**

評価版では割愛しています。

▷ Example 10: **シミュレーションデータセットの消去**

評価版では割愛しています。

▷ Example 11: Bayes prefix **の使用**

評価版では割愛しています。

■