ベイズ

ベイズ分析では、未知のパラメータを持つ対象について、確率的な分析を行うことができます。ベイズ分析の特徴は、分析の事前情報を取り込んで、立てた仮説の確率を頻度統計と比較してより直感的に求めることができるというところです。

- Kuehl (2000, 551)の酸素の取り込みと運動の研究例を考えます。
- 研究の目的は、2 種類の運動プログラム：12 週間のステップエアロビクスまたは 12 週間の平地ランニングを行った際の酸素取り込み量の変化を分析する事。
- 例題データセット、oxygen.dta をダウンロードしましょう。

```
use https://www.stata-press.com/data/r16/oxygen, clear
describe
```

- データセットには次の変数があります。

change	酸素取り込み量の変化(l/m)
group	運動グループ(0: ランニング、1:エアロビクス)
age	年齢
ageXgr	年齢とグループの交差項

- Kuehl (2000)では、このデータを分析するために共分散分析を使用しますが、ここでは以下の線形回帰モデルを考えます

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

例題 1: OLS

- まずは、OLS モデルにフィットさせてみましょう。

```
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

------------------------------------------------------------------------------
      change |      Coef.   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
------------------------------------------------------------------------------
```

- 帰無仮説H_o: $\beta_{group} = 0$は、p 値が 0.014 なので、有意水準 5%で棄却され、group と age はどちらもアウトカムに対して有意であることがわかります。

- しかし、p値は、使用したデータにおいてどの程度、帰無仮説が正しいのかという疑問に答えるものであり、帰無仮説そのものの確率ではありません。
- group の 95%信頼区間は[1.38 9.51]であり、間に 0 を含みません、このため group は change の有意な予測因子であると言えます。ただし、group の真の係数が 95%の確率で、1.38 と 9.51 の間に存在すると結論付けることは出来ません、
- パラメータの確率的な区間の解釈は、ベイズ統計でのみ可能なものです。

例題 2: 事前情報のないベイズ正規線形回帰

- 例題 1 の頻度統計ではパラメータは未知の固定量であるとして、パラメータの確率的要約を得ることはできないと述べました。
- ベイズ統計はパラメータの事後分布を様々な観点から推定することに焦点を置き観測されたデータの持つ情報で事前分布を更新します。つまり事後分布はパラメータの事前分布とパラメータを与えるデータの尤度関数で表されます。
- bayesmh コマンドで、oxygen.dta をベイズ線形回帰でフィットしましょう。
- ここのベイズ線形モデルでは 4 つのパラメータ（3 つの回帰係数とデータの分散）を利用します。アウトカム、change は正規分布に従うと仮定し、分散は Jeffreys の無情報事前分布で開始します。ジェフリーズの事前分布では、係数の同時事前分布と分散が分散の逆数に従います。Xはデザイン行列、$\beta = (\beta_0, \beta_{group}, \beta_{age})'$は係数のベクトルです。

$$\text{change} \sim N(X\beta, \sigma^2)$$

$$(\beta, \sigma^2) \sim \frac{1}{\sigma^2}$$

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

- コマンドの後に従属変数名、さらに共変量と続きます。アウトカムの分布は likelihood()オプションで指定し、事前分布は prior()オプションで指定します。
- 分散パラメータを指定しなければなりませんので、{var}として定義しています。3 つの回帰係数{change:group}、{change:age}、{change:_cons}は bayesmh によって自動的に作成されます。
- 分布を likelihood()オプション内で normal({var})と設定し、尤度関数の分散パラメータを{var}としています。この設定と回帰の設定を合わせて、アウトカム change の尤度モデルが定義されます。事前分布には prior({change:}, flat)で全ての回帰係数に密度 1 の一様分布を適用します。{change:}はモデル名 change の全てのパラメータの省略表現です。最後に事前分布 jeffreys で分散パラメータ{var}に密度$1/\sigma^2$を指定します。
- 乱数シードを設定して、分析結果を再現可能にします。以降の例題ではすべて乱数シー

ド値に 14 を使用します。

```
set seed 14
bayesmh change group age, likelihood(normal({var})) prior({change:}, flat) prior({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 はモデルの要約のヘッダの右側では、MCMC の反復回数が 12,500 であり、これは MCMC サンプルから除外されるデフォルト 2,500 回のバーンインとデフォルトが 10,000 回の MCMC サンプルまたは MCMC サンプルサイズの合計です。
- 受容率は推定されたパラメータ値をアルゴリズムが採択する割合です。この例の受容率 0.14 は 10,000 のパラメータの 14%がアルゴリズムで採択されることを示します。この数値は通常 30%に満たないことが大半です。受容率が低い(10%未満)場合は、収束に問題が生じていることがあります。
- 最後に、bayesmh は推定結果の要約表を報告します。Mean は事後平均の推定結果、パラメータの周辺事後分布の平均を報告しています。事後平均推定は例題 1 の OLS 推定の結果に非常に近くなります。この結果は、無情報事前分布を使用しているためです。
- Std. Dev.は、推定された事後標準誤差を報告します、これは事後分布の標準誤差です。この値はパラメータの事後分布の変動性を示し、OLS の標準誤差に相当します。

- Median は事後分布の中央値を報告し、分布の対称性を判断することができます。事後平均と中央値は回帰係数に近い値を取っていますので、事後分布が対称であることが考えられます。
- 最後の 2 列は、パラメータの信用区間を表します。信頼区間とは異なり、これらの区間は直感的に確率で解釈できます。例えば、group の係数が 1.16 と 9.25 の間にある確率はおよそ 0.95 です。区間の下限は 0 より大きいので、運動プログラムは酸素取り込みの変化に影響を与えていると結論付けることができます。

例題 3: 事前情報のあるベイズ線形回帰

- 単純化するためにすべての係数が独立で平均 0、分散σ^2の正規分布に従い、分散パラメータは上記のように逆ガンマ分布に従うものとします。

$$(\beta|\sigma^2)\sim \text{i.i.d } N(0,\sigma^2)$$

$$\sigma^2\sim \text{InvGamma}(2.5,2.5)$$

- このモデルを bayesmh コマンドでフィットします。normal(0, {var})で係数の事前分布を、igamma(2.5, 2.5)で分散の事前分布を指定します。

```
set seed 14
bayesmh change group age, likelihood(normal({var})) prior({change:},
normal(0, {var})) prior({var}, igamma(2.5, 2.5))
```

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

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

Priors:
  {change:group age _cons} ~ normal(0,{var})                            (1)
                     {var} ~ igamma(2.5,2.5)
------------------------------------------------------------------------------
(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  =      .1984
                                                Efficiency:  min =     .03732
                                                             avg =     .04997
Log marginal-likelihood = -49.744054                         max =     .06264

------------------------------------------------------------------------------
             |                                                 Equal-tailed
             |      Mean   Std. Dev.     MCSE     Median  [95% Cred. Interval]
-------------+----------------------------------------------------------------
change       |
       group |  6.510807   2.812828   .129931   6.50829    .9605561   12.23164
         age |  .2710499   .2167863   .009413   .2657002   -.1556194   .7173697
       _cons | -6.838302   4.780343   .191005  -6.683556   -16.53356   2.495631
-------------+----------------------------------------------------------------
         var |  28.83438   10.53573   .545382   26.81462    14.75695    54.1965
------------------------------------------------------------------------------
```

例題 4: 多変量事前情報のベイズ正規線形回帰

- 正規線形回帰の回帰係数でよく利用される Zellner の g 事前分布(Zellner 1986)を使用します。gは事前標本サイズ、ν_0逆ガンマ分布の事前自由度、σ_0^2は逆ガンマ分布の事前分散です。この事前情報は係数間で依存関係があります。ここでは、Hoff (2009)の値に近いパラメータ：$g = 12, \nu_0 = 1, \sigma_0^2 = 8$を使用します。

$$(\beta|\sigma^2) \sim \text{MVN}(0, g\sigma^2(X'X)^{-1})$$

$$\sigma^2 \sim \text{InvGamma}(\nu_0/2, \nu_0\sigma_0^2/2)$$

- bayesmh コマンドは zellnersg0()を使用して上記の事前情報を設定します。はじめに、分散の次元を記述します。この例では 3 とします。次に事前分布の自由度として 12 を使用します。最後に分散のパラメータを記述します。この例では{var}です。

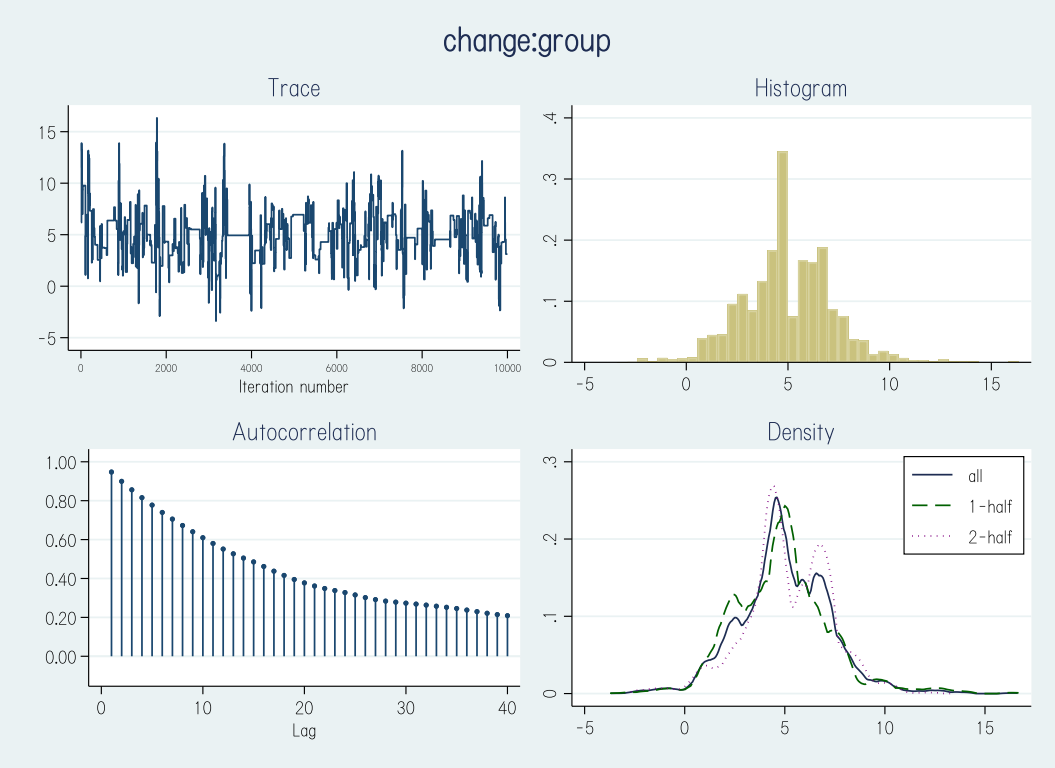
```
set seed 14
bayesmh change group age, likelihood(normal({var})) prior({change:},
zellnersg0(3,12,{var})) prior({var}, igamma(0.5, 4))
```

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

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

Priors:
  {change:group age _cons} ~ zellnersg(3,12,0,{var})                    (1)
                     {var} ~ igamma(0.5,4)
------------------------------------------------------------------------------
(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  =     .06169
                                                Efficiency:  min =      .0165
                                                             avg =     .02018
Log marginal-likelihood = -35.356501                         max =     .02159

------------------------------------------------------------------------------
             |                                                Equal-tailed
             |      Mean   Std. Dev.     MCSE     Median  [95% Cred. Interval]
-------------+----------------------------------------------------------------
change       |
       group |  4.988881   2.260571   .153837   4.919351   .7793098   9.775568
         age |  1.713159   .3545698   .024216   1.695671   1.053206   2.458556
       _cons | -42.31891   8.239571   .565879  -41.45385  -59.30145  -27.83421
-------------+----------------------------------------------------------------
         var |  12.29575   6.570879   .511475   10.3609   5.636195   30.93576
------------------------------------------------------------------------------
```

例題 5: 収束の確認

- bayesgraph コマンドでパラメータ推定における MCMC の収束を視覚的に確認できます。診断で良く用いられるグラフは、bayesgraph diagnostics でまとめて表示できます。

```
bayesgraph diagnostics {change:group}
```



- 表示される診断結果には、トレースプロット、自己相関プロット、ヒストグラム、MCMC サンプルの全体、前半、後半の密度とカーネル密度推定のオーバーレイです。
- トレースプロットと自己相関プロットは高い自己相関を示しています。ヒストグラムは単峰分布ではありません。これらから、収束に問題があることがわかります。
- 有効なサンプルサイズと収束に関する統計量を `bayesstats ess` コマンドで表示します。

```
bayesstats ess
```

```
Efficiency summaries     MCMC sample size =     10,000
                         Efficiency:  min =      .0165
                                      avg =     .02018
                                      max =     .02159
```

	ESS	Corr. time	Efficiency
change			
group	215.93	46.31	0.0216
age	214.39	46.64	0.0214
_cons	212.01	47.17	0.0212
var	165.04	60.59	0.0165

- MCMC サンプルの自己相関が低いほど、ESS 推定量は MCMC サンプルサイズに近くなり、推定したパラメータはより正確なものになります。
- このような場合、分散パラメータを独立したブロックに分割して、回帰係数とは別にパ

ラメータのアップデート行うことができます。bayesmh コマンドでは、block()オプションで指定します。

```
set seed 14
bayesmh change group age, likelihood(normal({var})) prior({change:},
zellnersg0(3,12,{var})) prior({var}, igamma(0.5, 4)) block({var})
saving(agegroup_simdata)
```

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

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

Priors:
  {change:group age _cons} ~ zellnersg(3,12,0,{var})                    (1)
                     {var} ~ igamma(0.5,4)
------------------------------------------------------------------------------
(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  =      .3232
                                                Efficiency:  min =     .06694
                                                             avg =      .1056
Log marginal-likelihood = -35.460606                         max =      .1443

------------------------------------------------------------------------------
             |                                                Equal-tailed
             |      Mean   Std. Dev.     MCSE     Median  [95% Cred. Interval]
-------------+----------------------------------------------------------------
change       |
       group |   5.080653   2.110911   .080507   5.039834   .8564619   9.399672
         age |   1.748516   .3347172   .008875   1.753897   1.128348   2.400989
       _cons |  -43.12425   7.865979   .207051   -43.2883  -58.64107  -27.79122
-------------+----------------------------------------------------------------
         var |   12.09916   5.971454   .230798   10.67555   5.375774   27.32451
------------------------------------------------------------------------------

file agegroup_simdata.dta saved
```

- estimates store agegroup を使用して、推定結果を agegroup に保存します。bayesmh の後に estimates store を使用するには、saving()オプションで bayesmh のシミュレーション結果を Stata のデータセットとして保存する必要があります。

例題 6: 推定後要約

- モデルパラメータとパラメータの関数について、bayesstats summary コマンドで推定後に要約して表示することができます。

```
summarize group
scalar sd_x = r(sd)
summarize change
scalar sd_y = r(sd)
bayesstats summary (group_std:{change:group}*sd_x/sd_y)
```

```
Posterior summary statistics                     MCMC sample size =    10,000

   group_std : {change:group}*sd_x/sd_y
```

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred. interval]	
group_std	.3283509	.1364233	.005203	.3257128	.0553512	.6074792

例題 7: ベイズ予測

- ベイズ予測は、モデルの適合度を確認し、観測値の将来の値を予測するのに適しています。bayespredict を使って、アウトカム変数 change の反復標本を作成し、新しいデータセット change_pred.dta として保存します。

```
bayespredict {_ysim}, saving(change_pred) rseed(16)
```

```
Computing predictions ...

file change_pred.dta saved
file change_pred.ster saved
```

- bayesstats summary コマンドで予測された観測値の事後要約を計算させます。

```
bayesstats summary {_ysim} using change_pred
```

Posterior summary statistics MCMC sample size = 10,000

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred. Interval]	
_ysim1_1	-2.954378	3.763301	.060963	-2.930854	-10.39297	4.528522
_ysim1_2	-4.610688	3.771203	.059014	-4.660554	-11.9289	2.948378
_ysim1_3	-4.620784	3.758543	.057517	-4.645584	-12.03851	2.917013
_ysim1_4	.6417156	3.756645	.063162	.6019013	-6.83463	8.330498
_ysim1_5	4.069868	3.972042	.072874	4.065139	-3.780329	12.06363
_ysim1_6	-8.120147	3.832453	.061674	-8.096888	-15.54334	-.3579446
_ysim1_7	16.18539	4.076738	.072385	16.2033	8.105208	24.23569
_ysim1_8	2.156433	3.921	.072344	2.135557	-5.528265	10.00732
_ysim1_9	9.14268	3.780417	.071241	9.154486	1.571643	16.59816
_ysim1_10	10.91948	3.776916	.068083	10.92263	3.445305	18.59981
_ysim1_11	.3919052	3.969695	.079798	.344616	-7.389234	8.386358
_ysim1_12	3.902787	3.809399	.077872	3.884087	-3.530938	11.49579

- 1 列目は、事後平均、事後予測分布による期待されるアウトカムの MCMC 推定です。
- 複製標本と観測標本を比較します。これら 2 つの差異は、bayesstats ppvalues コマンドで事後予測 p 値として計測されます。

```
bayesstats ppvalues {_ysim} using change_pred
```

Posterior predictive summary MCMC sample size = 10,000

T	Mean	Std. Dev.	E(T_obs)	P(T>=T_obs)
_ysim1_1	-2.954378	3.763301	-.87	.2786
_ysim1_2	-4.610688	3.771203	-10.74	.9512
_ysim1_3	-4.620784	3.758543	-3.27	.3479
_ysim1_4	.6417156	3.756645	-1.97	.773
_ysim1_5	4.069868	3.972042	7.5	.1819
_ysim1_6	-8.120147	3.832453	-7.25	.4034
_ysim1_7	16.18539	4.076738	17.05	.4124
_ysim1_8	2.156433	3.921	4.96	.2198
_ysim1_9	9.14268	3.780417	10.4	.3644
_ysim1_10	10.91948	3.776916	11.05	.4858
_ysim1_11	.3919052	3.969695	.26	.5106
_ysim1_12	3.902787	3.809399	2.51	.6498

Note: P(T>=T_obs) close to 0 or 1 indicates lack of fit.

- 推定された事後予測 p 値はすべて 0.05 と 0.95 の間(_ysim1_2 を除いて)に収まっています。この結果から各観測値が適切にフィットされていることがわかります。
- 新しく 2 つの観測値を追加し、bayespredict で標本外予測を行います。

```
set obs 14
replace group = 1 in 13
replace group = 0 in 14
replace age = 26 in 13/14
bayespredict pmean, mean rseed(16)
```

```
list change age group pmean
```

	change	age	group	pmean
1.	-.87	23	Running	-2.914124
2.	-10.74	22	Running	-4.613421
3.	-3.27	22	Running	-4.701283
4.	-1.97	25	Running	.545417
5.	7.5	27	Running	4.060798
6.	-7.25	20	Running	-8.111091
7.	17.05	31	Aerobic	16.15393
8.	4.96	23	Aerobic	2.183771
9.	10.4	27	Aerobic	9.155602
10.	11.05	28	Aerobic	10.87576
11.	.26	22	Aerobic	.4234267
12.	2.51	24	Aerobic	3.937901
13.	.	26	Aerobic	7.380203
14.	.	26	Running	2.405744

例題 8: モデルの比較

- ここまで推定モデルと誤差項を含む完全モデルを比較してみましょう。Zellner の g 事前分布と分散の逆ガンマ事前分布をもう一度使用します。事前パラメータの値は例題 4 のものを使用します。

```
set seed 14
bayesmh change group age ageXgr, likelihood(normal({var}))
prior({change:}, zellnersg0(4,12,{var})) prior({var}, igamma(0.5, 4))
block({var}) saving(full_simdata)
```

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

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

Priors:
  {change:group age ageXgr _cons} ~ zellnersg(4,12,0,{var})              (1)
                             {var} ~ igamma(0.5,4)
------------------------------------------------------------------------------
(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  =      .3113
                                                Efficiency:  min =      .0562
                                                             avg =      .06425
Log marginal-likelihood = -36.738363                         max =      .08478
```

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred. Interval]	
change						
group	11.94079	16.74992	.706542	12.13983	-22.31056	45.11963
age	1.939266	.5802772	.023359	1.938756	.7998007	3.091072
ageXgr	-.2838718	.6985226	.028732	-.285647	-1.671354	1.159183
_cons	-47.57742	13.4779	.55275	-47.44761	-74.64672	-20.78989
var	11.72886	5.08428	.174612	10.68098	5.302265	24.89543

```
file full_simdata.dta saved
```

```
estimates store full
```

- モデルを比較するには、bayesstats ic コマンドを使用します。

```
bayesstats ic full agegroup
```

Bayesian information criteria

	DIC	log(ML)	log(BF)
full	65.03326	-36.73836	.
agegroup	63.5884	-35.46061	1.277756

Note: Marginal likelihood (ML) is computed using Laplace-Metropolis approximation.

- DIC の値が小さく、log(ML)の値が大きいほど、良いと判断できます。この結果から、交差項の無いモデル agegroup が優れていることがわかります。

例題 9: 仮説検定

- 例題 8 を使用して、モデルの実際の確率を計算します。これには、bayestest model

コマンドを使用します。

```
bayestest model full agegroup
```

```
Bayesian model tests

------------------------------------------------
             |    log(ML)       P(M)      P(M|y)
-------------+----------------------------------
        full |  -36.7384     0.5000     0.2185
    agegroup |  -35.4642     0.5000     0.7815
------------------------------------------------
Note: Marginal likelihood (ML) is computed using
      Laplace-Metropolis approximation.
```

- どちらのモデルの事前確率が等しいという仮定の下では、交差項を含まないモデル、agegroup の確率は、0.78 で、full モデルでは 0.22 です。full モデルより優れていることをより強く裏付けるには、agegroup の確率がより大きく(0.9 以上)なければなりません。
- パラメータ区間の仮設検定を行うこともできます。交差項の無いモデルにおいて、group の係数が 4 と 8 の間にある確率を計算しましょう。

```
estimates restore agegroup
bayestest interval {change:group}, lower(4) upper(8)
```

```
Interval tests      MCMC sample size =     10,000

       prob1 : 4 < {change:group} < 8

------------------------------------------------
             |      Mean   Std. dev.       MCSE
-------------+----------------------------------
       prob1 |      .6159    0.48641   .0155788
------------------------------------------------
```

例題 10: シミュレーションデータセットの削除

- 目的の分析ができたら、bayesmh で作成したシミュレーション用のデータセットは不要なので削除するようにしましょう。

```
erase agegroup_simdata.dta
erase full_simdata.dta
```