Stataで学ぶベイズ統計

LightStone Corp

2020年4月

第5回 ベイズプリフィックス

- Stata 英文PDFマニュアルの例題を使って、bayes:(ベイズプリフィックス)の利用方法を説明する
- bayes:+推定コマンド。推定コマンドに応じて事前分布を自動設定する
- もちろん、常に最適な事前分布が自動設定される訳ではないので、調整する必要がある
- 具体的に事前分布を変更する場合はpriorオプションを利用する
- 線形モデルのパラメータの事前分布はデフォルトで平均0,分散10000の正規分布(実質的には無情報分布)

構成

- 無情報事前分布
  - Example 1:OLSによる推定
  - Example 2:ベイズ推定後の予測
  - Example 3:ギブスサンプリングの利用
- 有用な事前分布
  - Example 4:事前分布の変更
  - Example 5:一様分布
  - Example 6:Zellnerのg
- ロジスティック回帰における完全予測
  - Example 7:完全予測の問題点
  - Example 8:有用な事前分布
- 切断型のポアソン回帰
  - Example 9:デフォルトの事前分布
  - Example 10:ハイパーパラメータの利用

無情報事前分布

- 事前分布として無情報事前分布を利用した重回帰分析を行い、ベイズ推定の効果を考察する
- Example 1:通常のOLSによる推定とbayes:のデフォルト設定によるベイズ推定の結果を比較する
- Example 2:ベイズ推定後にフィット値の計算を行う
- Example 3:MCMCのアリゴリズムとして、より効率的なギブス・サンプリングを利用する

Example 1

- 女性の賃金に関するモデル推定

```
. use https://www.stata-press.com/data/r16/womenwage,clear
. reg wage age
```

Source	SS	df	MS
Model	3939.49247	1	3939.49247
Residual	43984.4891	486	90.503064
Total	47923.9816	487	98.406533

A

Number of obs	=	488
F(1, 486)	=	43.53
Prob > F	=	0.0000
R-squared	=	0.0822
Adj R-squared	=	0.0803
Root MSE	=	9.5133

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.399348	.0605289	6.60	0.000	.2804173	.5182787
_cons	6.033077	1.791497	3.37	0.001	2.513041	9.553112

ベイズ推定

2つの係数の事前分布は平均0, 分散10000の正規分布。これは実質的に無情報分布

```
. set seed 15
. bayes: regress wage age
```

	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

B

残差の分散には逆ガンマ分布igamma(0.01, 0.01)を利用。これも実質的に無情報分布(ただし、非負)。前頁のAと、このページのBはほぼ同じ値になる

収束

```
MCMC iterations  =     12,500
Burn-in          =      2,500
MCMC sample size =     10,000
Number of obs    =        488
Acceptance rate  =      .3739
Efficiency:  min =      .1411
             avg =      .1766
             max =      .2271
```

効率性は10%を越えており、問題なし(参照:第4回資料)

Example 2

- ベイズ推定後の統計量の計算には主にbayesstatsコマンドを利用する
- ここでは40歳の女性の賃金のフィット値wage40を求める

```
. bayesstats summary (wage40: {wage:_cons} + {wage:age}*40)
```

計算値の名前は適当に決める

```
.
```

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
wage40	22.00343	.81679	.024045	21.99231	20.39435	23.6718

Example 3

- bayes:を利用したベイズ推定のデフォルトのMCMCアルゴリズムは適応型(adaptive)MH。
- 効率性に優れたギブス・サンプリングを利用する

```
. set seed 15
. bayes, gibbs: regress wage age
```

ギブス・サンプリング

- MCSE以外はM-H法とほぼ同じ
- 効率性は格段に優れており、1になっている
- つまり、効率性が高くなると、MCSEはより小さくなる

	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

ほとんど変わらない

ギブス・サンプリングとは

- M-Hと同じくMCMCのアルゴリズムのひとつ
- Geman and Geman (1984)…画像処理の方法として提案
- Gelfand and Smith (1990)…事後分布を求める一般的な方法として提案し、ここから幅広く利用され始めた

ギブス・サンプラー

- ギブス・サンプリングを使って確率標本を作成する方法
- 基本となるsystematic scanによる単一連鎖(single chain)の用法を簡単に説明する
- 今、回帰式にp個のパラメータθがあるものとする

Step 1.適当な分布を使ってθの初期値を設定する

$$\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \ldots \theta_p^{(0)}\right)$$

Step 2.初期値の(0)を一般化して(i)番目のパラメータセットを次のように考える。*i*はモンテカルロ計算の繰り返し回数

$$\boldsymbol{\theta}^{(i)} = \left(\theta_1^{(i)}, \theta_2^{(i)}, \ldots \theta_p^{(i)}\right)$$

ギブス・サンプラー

Step 2a.観測したデータyを使って先頭のパラメータθ1のi+1番目求める。右辺のθは*i*番目の値

$$\theta_1^{(i+1)} = \pi\left(\theta_1 | \theta_2^{(i)}, \ldots \theta_p^{(i)}, \mathbf{y}\right)$$

Step 2b.2番目のパラメータをStep2aの値を使って求める

$$\theta_2^{(i+1)} = \pi\left(\theta_2 | \theta_1^{(i+1)}, \theta_3^{(i)} \ldots \theta_p^{(i)}, \mathbf{y}\right)$$

この作業を繰り返してN+M個のパラメータセットを作成する。ここでNは稼働検査期間、Mは平均や標準偏差の計算に利用する標本数

多重連鎖

- 初期値をkセット用意して、それぞれのセットを使ってギブス・サンプリングを行う方法を多重連鎖(multiple chain)と呼ぶ
- 多重連鎖から得られる標本は独立となる割合が大きい。しかし、初期値が適切でないと、不変分布に収束しない可能性が生じる
- Stata v16から多重連鎖をサポート

有用な事前分布

- 事前分布として有用な事前分布を利用した重回帰分析を行う
- Example 4:デフォルトの事前分布
- Example 5:一様分布
- Example 6:Zellnerのg

Example 4

- デフォルトの事前分布
  - 線形回帰モデルの推定値の大きさ(絶対値)に注目する

```
. use https://www.stata-press.com/data/r16/auto, clear
. reg price length i.foreign
```

.

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
length	90.21239	15.83368	5.70	0.000	58.64092	121.7839
foreign						
Foreign	2801.143	766.117	3.66	0.000	1273.549	4328.737
_cons	-11621.35	3124.436	-3.72	0.000	-17851.3	-5391.401

稼働検査期間の調整

- 係数推定値が2桁以上ある(大きな推定値)場合、稼働検査期間を長くとって、MCMCによるサンプリングの範囲が適切な範囲に収まるように調整する
- デフォルトの稼働検査期間2,500を5,000に変更する

```
. set seed 15
. bayes, burnin(5000): regress price length i.foreign
```

稼働検査期間

絶対値ですべて小さくなっている

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
price						
length	33.03301	1.80186	.060848	33.07952	29.36325	36.41022
foreign						
Foreign	32.77011	98.97104	4.07922	34.3237	-164.1978	222.0855
_cons	-8.063175	102.9479	3.34161	-9.110308	-205.9497	196.9341
sigma2	7538628	1297955	29334.9	7414320	5379756	1.04e+07

- 回帰係数のデフォルトの事前分布はN(0,10000)
- 係数の大きさが2桁程度だと、本来、無情報であるものが、情報として価値を持つようになる
- 定数項は-11621から驚くほど縮小(シュリンケージ)している

Example 5

- 事前分布として一様分布を利用する
- これは結果としてOLSの推定値と同じような値になることが多い

```
. set seed 15
. bayes, prior({price:},flat) burnin(5000): regress price length i.foreign
```

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
price						
length	89.51576	16.27187	.586237	89.60969	57.96996	122.7961
foreign						
Foreign	2795.683	770.6359	26.0589	2787.139	1305.773	4298.785
_cons	-11478.83	3202.027	113.271	-11504.65	-17845.87	-5244.189
sigma2	6270294	1089331	25002.1	6147758	4504695	8803268

Example 6

- 多変量正規分布を用いるZellnerのg

```
. set seed 15
. bayes, prior({price:}, zellnersg0(3, 50, {sigma2})) burnin(5000):
        regress price length i.foreign
```

3　:線形モデルのパラメータの個数
50　:自由度(事前分布の影響力。サンプルサイズに近い値を利用する)
{sigma2}:回帰モデルの誤差分散のパラメータ

Zellnerのg

.

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
price						
length	87.53039	16.24762	.569888	87.72965	55.5177	119.9915
foreign						
Foreign	2759.267	794.043	31.3829	2793.241	1096.567	4202.283
_cons	-11223.95	3211.553	113.34	-11308.39	-17534.25	-4898.139
sigma2	6845242	1159035	26286.9	6716739	4978729	9521252

OLSの推定結果に近い

有用な事前分布

- 事前分布として有用な情報事前分布を利用した様々な回帰分析例
- Example 7:ロジスティック回帰の完全予測
- Example 8:有用な事前分布の利用
- Example 9:切断型ポアソン回帰のベイズ推定
- Example10:ハイパーパラメータの利用

Example 7

- 心臓病のリスク要因をモデル化する

```
. use https://www.stata-press.com/data/r16/heartswitz, clear
. des
```

- disease:心臓病の有無
- restecg:心電図による診断結果(0は正常)
- isfbs:空腹時血糖のインジケータ変数
- 通常のロジットモデルの推定

```
. logit disease restecg isfbs age male
```

完全予測

```
Logistic regression                               Number of obs     =         26
                                                  LR chi2(1)        =       0.01
                                                  Prob > chi2       =     0.9403
Log likelihood = -4.2358076                       Pseudo R2         =     0.0007
```

disease	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
restecg	0	(omitted)				
isfbs	0	(omitted)				
age	-.0097846	.1313502	-0.07	0.941	-.2672263	.2476572
male	0	(omitted)				
_cons	3.763893	7.423076	0.51	0.612	-10.78507	18.31285

- 3つのカテゴリカル変数とアウトカムが完全予測の状態になっているので、係数を推定できない

デフォルト設定によるベイズ推定

```
. set seed 15
. bayes: logit disease restecg isfbs age male
```

disease	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
restecg	(omitted)					
isfbs	(omitted)					
age	-.0361322	.1663327	.004759	-.0284239	-.4043079	.2714653
male	(omitted)					
_cons	6.377579	9.57045	.270646	5.547644	-9.432267	28.88536

- デフォルトの事前分布を利用したベイズ推定では問題を解決できない

Example 8

```
. set seed 15
. bayes, prior({disease:restecg age}, normal(0,10))
         prior({disease:isfbs male}, normal(1,10))
         prior({disease:_cons}, normal(-4,10)) nomleinitial:
         logit disease restecg isfbs age male, asis
```

- 変数のセットを3つに分けて、それぞれ異なる正規分布を事前分布として設定する
- nomleinitial:デフォルトでは最尤法推定の推定値をベイズ推定の初期値に利用するが、ここでは利用しない
- asis:完全予測の場合に脱落する変数を脱落させない。このオプションだけでは問題解決はできないことに注意

有用な事前分布

- 事前分布と初期値の計算を工夫することでパラメータを推定できた

.

disease	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
restecg	1.965122	2.315475	.115615	1.655961	-2.029873	6.789415
isfbs	1.708631	2.726071	.113734	1.607439	-3.306837	7.334592
age	.1258811	.0707431	.003621	.1245266	-.0016807	.2719748
male	.2671381	2.237349	.162967	.3318061	-4.106425	4.609955
_cons	-2.441911	2.750613	.110611	-2.538183	-7.596747	3.185172

bayes:の初期値

- ベイズ推定のパラメータの初期値には最尤推定の値を利用する

.bayes: logit y x

.logit y x (最初にML推定を実行して初期値を得る)

Example 9

- 切断ポアソン回帰

```
. use https://www.stata-press.com/data/r16/probe, clear
. des
```

- シリコンウェーハーの製造において4つのプローブ(故障検出装置)が異常を検出する
- 異常の件数が10を超える場合に、その件数が変数failuresに記録される
- failuresには11以上の値が入っており、10以下の数は存在しない
- probeはカテゴリな変数(1-4)

切断ポアソン回帰

- カテゴリな変数の係数をすべて表示したいので、定数項を利用しない切断ポアソン回帰モデルを推定する

```
. tpoisson failures ibn.probe depth width,noconstant ll(10)
```

failures	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
probe						
1	2.714025	.0752617	36.06	0.000	2.566515	2.861536
2	2.602722	.0692732	37.57	0.000	2.466949	2.738495
3	2.725459	.0721299	37.79	0.000	2.584087	2.866831
4	3.139437	.0377137	83.24	0.000	3.065519	3.213354
depth	-.0005034	.0033375	-0.15	0.880	-.0070447	.006038
width	.0330225	.015573	2.12	0.034	.0025001	.063545

デフォルトの設定を利用

```
. set seed 15
. bayes, dots: tpoisson failures ibn.probe depth width, noconstant ll(10)
```

failures	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
probe						
1	2.689072	.0696122	.008596	2.688881	2.557394	2.833737
2	2.581567	.0644141	.00966	2.588534	2.436973	2.701187
3	2.712054	.0695932	.006415	2.717959	2.55837	2.844429
4	3.13308	.0397521	.004592	3.133433	3.055979	3.208954
depth	-.000404	.0033313	.000165	-.000504	-.0067928	.0061168
width	.036127	.0165308	.001821	.0360637	.001239	.067552

Note: Default priors are used for model parameters.

Note: There is a high autocorrelation after 500 lags.

ラグ500で自己相関が0.01を超える場合に表示する

*効率性が0.013で悪いことに注意

自己相関

- デフォルトの設定ではベイズ推定と、標準コマンド(最尤推定)の推定結果はほぼ同じ
- ベイズ推定では自己相関に関する警告が表示されている
- MCMCの標本における自己相関が500以上のラグでも存在する場合、稼働検査期間をより長く設定する方法もある
- ここでは事前分布を変更してデータへのフィットを向上し、自己相関の解消を狙う

Example 10

- ハイパーパラメータという考え方を利用する
- probeの4つのダミー変数の係数{failures:ibn.probe}に平均のパラメータ{probe_mean}を設定し、分散は10000とする
- {probe_mean}はgamma(2,1)のハイパー事前分布に従うものとする。この平均は正値とする
- {probe_mean}の初期値は1とする

```
. set seed 15
. bayes, prior({failures:ibn.probe}, normal({probe_mean}, 10000)) ///
  prior({probe_mean}, gamma(2, 1)) initial({probe_mean} 1) dots: ///
  tpoisson failures ibn.probe depth width, noconstant ll(10)
```

ハイパーパラメータ

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
failures						
probe						
1	2.703599	.0770656	.003757	2.704613	2.551404	2.848774
2	2.592738	.0711972	.002796	2.594628	2.446274	2.728821
3	2.716223	.0755001	.003549	2.719622	2.568376	2.863064
4	3.137069	.0388127	.001317	3.136773	3.062074	3.211616
depth	-.000461	.0033562	.000109	-.0004457	-.0067607	.0062698
width	.0337508	.0152654	.000532	.0337798	.003008	.0622191
probe_mean	2.051072	1.462867	.041051	1.71286	.2211973	5.809428

*効率性(avg)は約8%に向上して、収束に問題はない
*推定値は最尤法のそれに近い
*MCSEはデオフォルトの推定に比べ小さくなっている(モデルの当てはまりが良いことを示している)

ゼロ拡大-負の二項分布

- 国立公園で釣果countをモデル化する
- 釣果(匹数)はグループの人数やエサの種類に依存する
- 一方で、まったく釣果のない個人やグループをたくさん存在する
- グループ内の子供の人数や、キャンパーであるかによって釣果がゼロになるか否かは大きく左右されると考える

```
. use https://www.stata-press.com/data/r16/fish, clear
. zinb count persons livebait, inflate(child camper) nolog
```

負の二項分布

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
count						
persons	.9742984	.1034938	9.41	0.000	.7714543	1.177142
livebait	1.557523	.4124424	3.78	0.000	.7491503	2.365895
_cons	-2.730064	.476953	-5.72	0.000	-3.664874	-1.795253
inflate						
child	3.185999	.7468551	4.27	0.000	1.72219	4.649808
camper	-2.020951	.872054	-2.32	0.020	-3.730146	-.3117567
_cons	-2.695385	.8929071	-3.02	0.003	-4.44545	-.9453189
/lnalpha	.5110429	.1816816	2.81	0.005	.1549535	.8671323
alpha	1.667029	.3028685			1.167604	2.380076

ベイズ推定

ML推定値と似ている

```
. set seed 15
. bayes, dots: zinb count persons livebait, inflate(child camper)
```

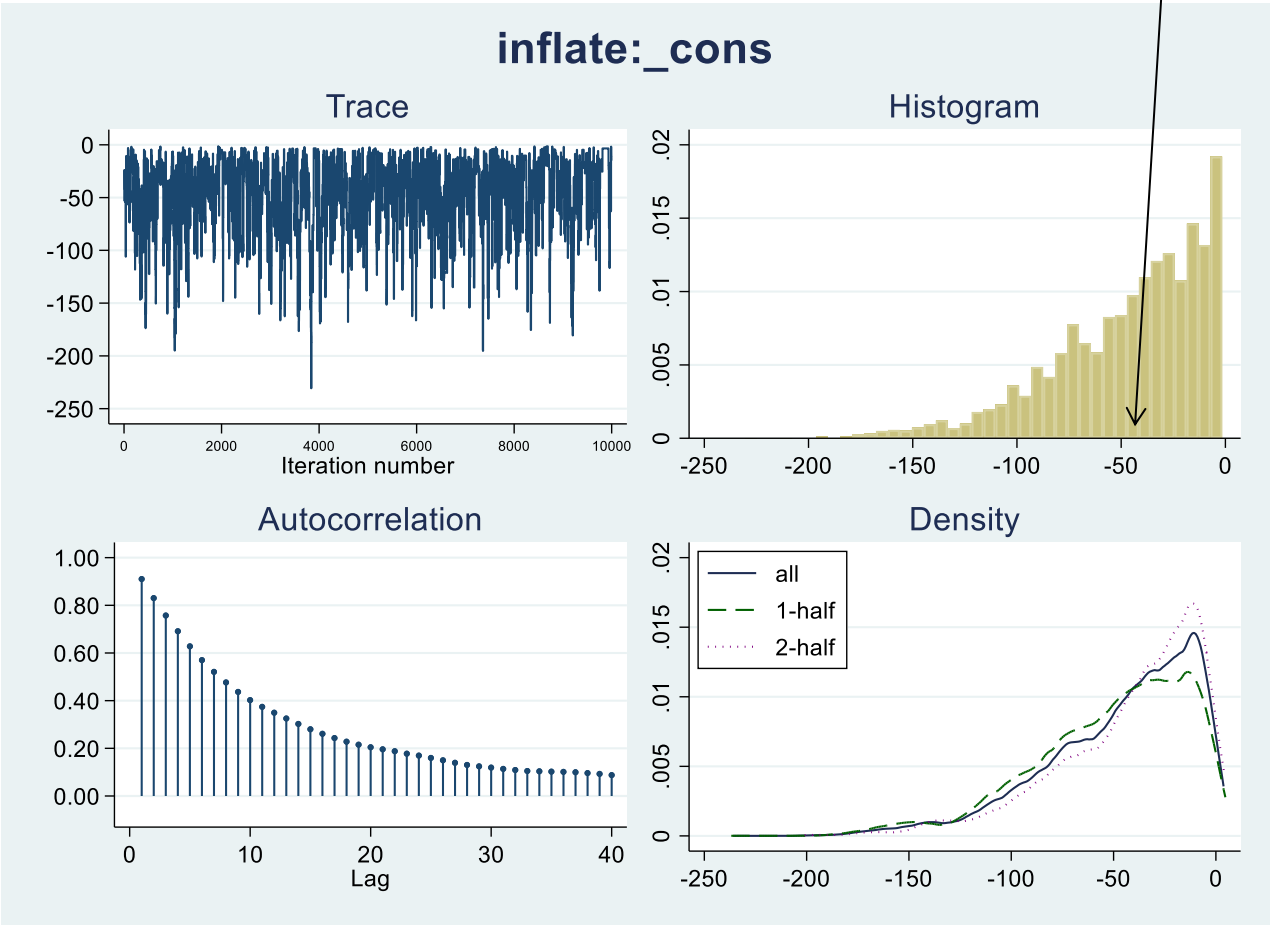
	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
count						
persons	.9851217	.1084239	.003601	.985452	.7641609	1.203561
livebait	1.536074	.4083865	.013509	1.515838	.753823	2.3539
_cons	-2.805915	.4700702	.014974	-2.795244	-3.73847	-1.89491
inflate						
child	46.95902	36.33974	1.87977	38.77997	3.612863	138.3652
camper	-46.123	36.34857	1.88567	-37.66796	-137.4568	-2.544566
_cons	-46.62439	36.36232	1.88355	-38.5171	-137.5522	-3.272469
lnalpha	.7055935	.1591234	.003962	.7048862	.3959316	1.025356

ゼロを拡大するモデルには大きな違いがある

診断

. bayesgraph diagnostic {inflate:_cons}

期待値は-46



*定数項を示す標本の中央値を求めるてみる

MC標本の中央値

- モンテカルロ標本をファイル保存して中央値を調べる

```
. bayes, saving(sim_zinb, replace)
```

- モンテカルロ標本のファイルを開く

```
. use sim_zinb, clear
. des
```

MC標本の中央値

- inflateの式の3番目の係数が定数項

variable name	storage type	display format	value label	variable label
_chain	int	%8.0g		Chain identifier
_index	int	%8.0g		Iteration number
_loglikelihood	double	%10.0g		Log likelihood
_logposterior	double	%10.0g		Log posterior
eq1_p1	double	%10.0g		{count:persons}
eq1_p2	double	%10.0g		{count:livebait}
eq1_p3	double	%10.0g		{count:_cons}
eq2_p1	double	%10.0g		{inflate:child}
eq2_p2	double	%10.0g		{inflate:camper}
eq2_p3	double	%10.0g		{inflate:_cons}
eq0_p1	double	%10.0g		{lnalpha}
_frequency	int	%8.0g		Frequency weight

MC標本の中央値

```
. egen mode = mode(eq2_p3)
. display mode[1]
```

- 中央値が-3.42と表示される
- これは最尤推定における定数項-2.70に近い
- 拡大パラメータαを比較する。ベイズ推定のパラメータは対数変換されているので、指数計算を行い比較する

```
. bayesstats summary (alpha: exp({lnalpha}))
.
```

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
alpha	2.050889	.3292052	.008191	2.023616	1.485768	2.788087

ML推定よりも、やや大きい

bayes:の事前分布に関する補足

- 分散など、正値に限定されるパラメータの場合、逆ガンマ分布が利用される(デフォルト)

$$\sigma^2 \sim \mathrm{InvGamma}(0.01, 0.01)$$

- 順序ロジットなどで利用するカットポイントのパラメータに対してはデフォルトで一様分布(flat)が利用される
- マルチレベル分析のランダム効果については逆ガンマ分布を利用されるが、ランダム効果の分散共分散行列については、その構造によって異なる事前分布が適用される

第5回のまとめ

- bayes:の基本的な用法
- MCMCのアルゴリズムであるギブスサンプリングの基本的な考え方を紹介した。欠損値問題に対応するための統計手法「多重代入」はこれと良く似た方法を利用している
- 多変量正規分布を事前分布に用いる場合はZellerのgを用いる
- ロジットモデルにおける完全予測もベイズ推定で対応可能
- パラメータを生み出すメカニズムとしてハイパーパラメータという考え方がある

参考文献

- STATA BAYESIAN ANALYSIS REFERENCE MANUAL, RELEASE 16

研究者向けの統計解析ソフトウェア Stata® 評価版のお申込み

機能制限なし！30日間無料で使える評価版でまずはお試しください！ https://www.lightstone.co.jp/stata/evaluate.html

※学生の方への評価版の提供はございません。大学の講義で使用するなど単位取得に関わる形でのご利用となる場合は学生版のご購入をご検討ください。

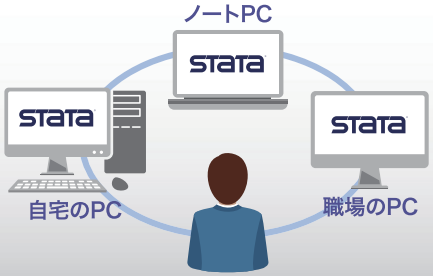
ライセンスはサブスクリプションがおすすめ！

サブスクリプションライセンスを選ぶメリット

- 常にStataの最新バージョンが利用できます
- 技術サポートの対象となります
- 毎年の経費としてStataを導入できます
- 利用者の増減に柔軟に対応可能
- 初期導入費を抑えたい場合にもおすすめ

シングルライセンス（サブスクリプション）

職場のPC・ノートPC・自宅のPC等、個人所有のPCにインストール可能です。
同時に使用できるのは1台のみとなります。



マルチユーザライセンス（サブスクリプション） 2ユーザ以上

ボリュームライセンス

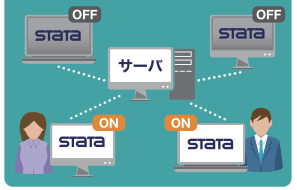
シングルライセンスを複数人でまとめてご購入いただく際のボリュームディスカウント商品です。

同時起動ライセンス

ご利用環境により下記3通りの運用方法からご選択いただけます。

PC毎にスタンドアロンインストールをして運用
例:2ライセンスの場合

クライアント＆サーバ環境によるネットワーク運用
例:2ライセンスの場合



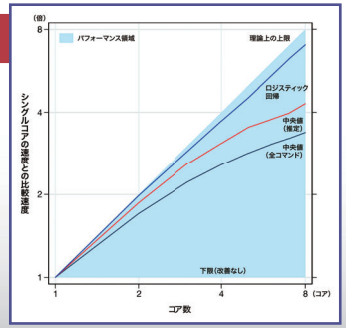
1台の仮想マシンにユーザがアクセスする運用
例:2ライセンスの場合

Stata／MPの演算能力

データ処理時間の削減ならMP

Stata MPはPCの持つマルチコアの特性を活かして、処理を分散・並列する機能を備えます。およそ85%以上のコマンドで処理速度が向上し、コア数に応じた計算時間の短縮が期待できます。
コマンドごとの処理速度の向上の度合いについては以下のページや資料をご覧ください。

コア数	全てのコマンド	推定コマンド	ロジスティック回帰
2	1.7倍	1.8倍	1.9倍
4	2.6倍	3.1倍	3.8倍
8	3.3倍	4.2倍	6.8倍



弊社Webページ内の『Stata／MP』
https://www.lightstone.co.jp/stata/statamp.html

株式会社 ライトストーン
〒101-0031
東京都千代田区東神田 2-5-12 龍角散ビル 7F
TEL：03-3864-5211
FAX：03-3865-0050
e-mail： sales@lightstone.co.jp
WEB： https://www.lightstone.co.jp/