研究者が開発した研究者のための
統計解析ソフトウェア

STATA®

期間限定!

アカデミック向けキャンペーン開催

2019年3月29日(金)ご注文分まで適用

大学生協をはじめ
全国の販売会社様から
ご購入いただけます

すべてのStataアカデミック版がキャンペーン価格となります

Stata IC アカデミック版

25% OFF

~~¥72,000(税別)~~

¥54,000(税別)~

取扱いのデータの数、処理速度の向上が可能となる
異なるエディションもキャンペーン対象です。
詳しくはお問合せください。

最新のStataを30日間
無料で体験できる!
評価版の申込み受付中

広範な統計機能群

マルチレベル分析、傾向スコア分析、ANOVA、ベイズ分析、生存分析、メタ分析、t検定など

分析操作の記録・再現が容易

分析結果を再現するためのDoファイル、解析記録を残すためのログファイルにより、統計解析作業のエビデンスの保存、適正な運用を支援します

無償テクニカルサポート

お電話・チャットサポート(平日9時～18時)、メールフォーム、Eメールにてインストールや操作方法に関する疑問にお答えします

StataPress書籍の邦訳本

Stataの使い方や医療統計の実施方法を解説した書籍の邦訳本「医療研究者のためのStata入門」を販売しています

国内初! Stata社公認日本語マニュアルをご提供

日本語GettingStartedマニュアルがダウンロードできます

各種セミナーの開催

Stataの使い方や分析手法を学べるセミナーを開催しております

- 初めてのStata
- メタ分析
- 傾向スコア分析
- マルチレベル分析
- 医療統計
- 回帰分析
- 他

製品、セミナーに関する詳細はWebで!
https://www.lightstone.co.jp/stata/lp/medical.html

お問い合わせ先
LightStone® 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

ベイズ推定による解決法 ～ logit モデルの分離問題～

Stata社の提供している下記Stataブログの内容を要約してベイズ推定のメリットとそれに関連するコマンドを紹介します。詳細は下記ブログをご参照ください。

Stata Blog
開発元のホームページ（Stata.com）より「Stata Blog」→「Bayesian logistic」と検索ください。
Bayesian logistic regression with Cauchy priors using bayes prefix
https://blog.stata.com/2017/09/08/bayesian-logistic-regression-with-cauchy-priors-using-the-bayes-prefix/
サンプルデータ「irisstd.dta」はこちらから入手可能です。

被説明変数はアヤメの種別を示す0と1の二値変数です，説明変数は花弁と萼片の大きさを示す連続変数です。
データ平均0、標準偏差0.5となるように標準化されています。

❶ 最初に記述統計量を確認します。
コマンド：summarize

Variable	Obs	Mean	Std. Dev.	Min	Max
virg	150	.3333333	.4729838	0	1
slen	150	4.09e-09	.5	-.8318901	1.241849
swid	150	-2.88e-09	.5	-1.215422	1.552142
plen	150	2.05e-10	.5	-.7817487	.8901885
pwid	150	4.27e-09	.5	-.7198134	.8525946

最初と最後の行のデータはベイズ推定によるモデル後にバリデーション用に利用しますので、推定の標本からは除外することにします。
コマンド：gen touse=_n>1 & _n<_N

Logit Modelの推定
コマンド：logit virg slen swid plen pwid if touse,nolog

```
Logistic regression                             Number of obs     =        148
                                                LR chi2(4)        =     174.09
                                                Prob > chi2       =     0.0000
Log likelihood = -5.9258976                     Pseudo R2         =     0.9369
```

virg	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
slen	-3.993255	3.953552	-1.01	0.313	-11.74207	3.755564
swid	-5.760794	3.849979	-1.50	0.135	-13.30661	1.785028
plen	32.73222	16.62764	1.97	0.049	.1426454	65.3218
pwid	27.60757	14.78357	1.87	0.062	-1.367681	56.58283
_cons	-19.83216	9.261786	-2.14	0.032	-37.98493	-1.679393

Note: 54 failures and 6 successes completely determined.

"Note : 54 Failures and 6 successes completely determined"と推定結果の下にデータにばらつきがないことによる、分離問題の発生を示すメッセージが表示されています。この状態になると推定値と標準誤差にバイアスが発生します。ここでの分離問題の発生原因は説明変数pwidにあります。そこで、対応策としてベイズ推定を実行します。

❷ ベイズ推定の為、乱数作成キーを設定します。
コマンド：set seed 15
Bayesの推定　コマンドの先頭にbayes:を付けるだけで実行できます。
コマンド：bayes:logit virg slen swid plen pwid if touse

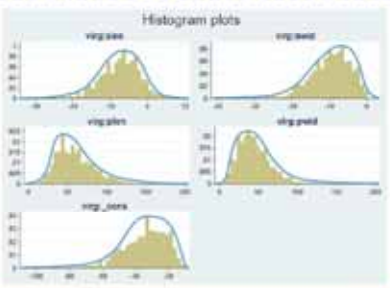
```
Likelihood:
  virg ~ logit(xb_virg)

Priors:
  {virg:slen swid plen pwid _cons} ~ normal(0,10000)                (1)

(1) Parameters are elements of the linear form xb_virg.

Bayesian logistic regression                    MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling        Burn-in          =      2,500
                                                MCMC sample size =     10,000
                                                Number of obs    =        148
                                                Acceptance rate  =      .1511
                                                Efficiency:  min =      .0119
                                                             avg =      .0204
Log marginal likelihood = -20.64697             max =      .03992
```

virg	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
slen	-7.391316	5.256959	.263113	-6.861438	-18.87585	2.036089
swid	-9.686068	5.47113	.492419	-9.062651	-21.32787	-1.430718
plen	59.90382	23.97788	1.53277	56.48103	23.00752	116.0312
pwid	45.65266	22.2054	2.03525	42.01611	14.29399	99.51405
_cons	-34.50204	12.77856	1.19136	-32.61649	-66.09916	-14.93324

Note: Default priors are used for model parameters.

Check! 最尤法の係数"coef"にくらべ、絶対値でみると推定値が大きく変化しています。

❸ パラメータの事前分布にはデフォルトの正規分布が利用されています。
乱数を用いたベイズ推定の結果として得られたパラメータの分布を確認。
コマンド：
bayesgraph histogram _all,combine(rows(3))



bad.. 事前分布として正規分布を適用すると、これらの事後分布に左右への歪が確認できます。

❹ そこで、Gelman et al. (2008)の提案を利用して事前分布としてコーシー分布を検討します。

```
set seed 15
bayes,prior({virg:_cons},cauchy(0,10)) prior({virg:slen
swid plen pwid },  ///
cauchy(0,2.5)):logit virg slen swid plen pwid if touse
```

```
Likelihood:
  virg ~ logit(xb_virg)

Priors:
                {virg:_cons} ~ cauchy(0,10)                         (1)
  {virg:slen swid plen pwid} ~ cauchy(0,2.5)                        (1)

(1) Parameters are elements of the linear form xb_virg.

Bayesian logistic regression                    MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling        Burn-in          =      2,500
                                                MCMC sample size =     10,000
                                                Number of obs    =        148
                                                Acceptance rate  =      .2072
                                                Efficiency:  min =      .01907
                                                             avg =      .02489
Log marginal likelihood = -18.651475            max =      .03497
```

virg	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
slen	-2.04014	2.332062	.143782	-1.735382	-7.484299	1.63583
swid	-2.59423	2.00863	.167406	-2.351623	-7.379954	.6207466
plen	21.27293	10.37093	.717317	19.99503	4.803869	45.14316
pwid	16.74599	7.506278	.54353	16.00158	4.382826	35.00463
_cons	-11.96009	4.157192	.272998	-11.20163	-22.45385	-6.549506

Log marginal likelihood = -20.64697
Log marginal likelihood = -18.651475

Good! 両者の対数尤度を比べると-20から-18となり、コーシー分布の方が適切であることが分かります。

❺ 最後に推定に利用しなかったデータを用いてバリデーションを行います。
バリデーション
コマンド：list if !touse

	virg	slen	swid	plen	pwid	touse
1.	0	-.448837	.5143057	-.668397	-.6542964	0
150.	1	.0342163	-.0622702	.3801059	.3939755	0

平均と95%信用区間
コマンド：

```
bayesstats summary (prob0:invlogit(-.448837*{virg:slen}
+.5143057*{virg:swid}  ///
-.668397*{virg:plen}-.6542964*{virg:pwid}+{virg:_cons}
)),nolegend
bayesstats summary (prob1:invlogit(.0342163*{virg:slen}
-.0622702*{virg:swid}  ///
+.3801059 *{virg:plen}+.3939755*{virg:pwid}+
{virg:_cons})),nolegend
```

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

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
prob0	7.26e-10	1.02e-08	3.1e-10	4.95e-16	3.18e-31	8.53e-10

```
. bayesstats summary (prob1:invlogit(.0342163*{virg:slen}-.0622702*{virg:swid}) ///
>         +.3801059 *{virg:plen}+.3939755*{virg:pwid}+{virg:_cons})),nolegend

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

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
prob1	.9135251	.0779741	.004257	.9361961	.7067089	.9959991

1行目の観測値0 ： 理論値 7.26e-10
150行目の観測値1 ： 理論値 0.9135251

Answer! 両者ともそれぞれ0と1に近い値を示しており、観測値の値と整合的であることが分かります。