# Stataによる多重代入

株式会社　ライトストーン

2020年8月

修正2020年10月

# はじめに

- 第2回では欠損値が存在するときの対応方法の一般論について解説した
- 今回はStataを使って欠損値がモノトーンパターンの場合と、モノトーンパターンに従わないときの、代表的な対応方法について説明する

# 第3回 多変量正規分布の利用

- 欠損メカニズムがモノトーンパターンの時
  - mi impute monotone:多変量正規分布とは無関係。モノトーンパターンに従う場合にだけ利用する代入コマンド
- データ拡大アルゴリズム(Data Augmentation)
  - mi impute mvn:欠損メカニズムがモノトーンパターンでない時に利用するベイズ統計による代入コマンド。変数の正規性を利用する
- キーワード:データ拡大アルゴリズム、Average RVI、Largest FMI、MCMCの収束、WLF

# 1.欠損メカニムズがモノトーンパターンの時

- 欠損値の発生メカニズムがモノトーンパターンの時は比較的、簡単な計算方法で多重代入が実行できる
- もう少し具体的に言えば、複数の変数に欠損値が存在しても、一変数ごとにモデルを利用した計算を繰り返していけばよい
- 今、欠損のある変数Xがp個あり、欠損値を含まない完全データの変数Zが存在するものとする

*モノトーンパターンに関する解説は第2回にもあり

# モノトーンパターンの代入

- 次に示す方法で理論値x*を求める(Rubin 1987)

$$
\begin{aligned}
x_1^* &\sim f_1(x_1|z) \\
x_2^* &\sim f_2(x_2|x_1^*, z) \\
&\cdots \\
x_p^* &\sim f_p\left(x_p|x_1^*, x_2^*, \ldots, x_{p-1}^*, z\right)
\end{aligned}
$$

- f()は欠損値の代入に利用する関数
- ポイントは欠損の多い変数の代入値の計算に、すでに計算した代入値を利用するところ

# 例題:モノ トーンパターン

操作:第1回と同じサンプルデータを利用して、モノトーンパターンの確認と、多重代入、そしてロジスティックモデルの推定を行う

```
. use https://www.stata-press.com/data/r16/mheart5s0, clear
. mi des
. misstable nested
```

1. age(12) -> bmi(28)

ageとbmiに欠損値が存在し、ageの12個の欠損値はbmiの28個の欠損値にネストしている

# 例題:モノ トーンパターン

bmiには回帰モデルを利用し、ageにはpmm(predictive mean matching)を利用した代入をおこなう。pmmの詳細は第4回で解説する

.mi impute monotone (reg) bmi (pmm,knn(3)) age = attack smokes hsgrad female ,add(10)

```
Conditional models:
                age: pmm age attack smokes hsgrad female , knn(3)
                bmi: regress bmi age attack smokes hsgrad female
```

ageの12個の欠損値に代入値を作成し、それを利用してbmiの代入値を作成している。代入回数は10回

# 例題:モノ トーンパターン

. mi est,dots:logit attack smokes age bmi hsgrad female

- オプションdotsは計算プロセスを示すもの。計算に長い時間を要する場合に利用する

| attack | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| smokes | 1.168286 | .357026 | 3.27 | 0.001 | .4685075 | 1.868065 |
| age | .0279243 | .0172083 | 1.62 | 0.106 | -.0060082 | .0618569 |
| bmi | .1039346 | .0513625 | 2.02 | 0.045 | .0023066 | .2055627 |
| hsgrad | .1415342 | .40294 | 0.35 | 0.725 | -.6482223 | .9312907 |
| female | -.0824037 | .4128752 | -0.20 | 0.842 | -.8916261 | .7268186 |
| _cons | -4.991369 | 1.743229 | -2.86 | 0.005 | -8.425476 | -1.557262 |

- オッズ比を表示する場合はeformオプションを次の位置に入力する

. mi estimate, eform: logit attack smokes age bmi hsgrad female

# Average RVI

- 欠損値の割合が高いほど、推定値の分散は大きくなる
- RVIはrelative variance increase。この値が小さいほど、欠損値が分散に与える影響が小さいことを示す
- したがって、もし、完全データで同じ推定を実行すれば、RVIはゼロになる

```
Imputations        =        10
Number of obs      =       154
Average RVI        =    0.1139
Largest FMI        =    0.2762
DF:      min       =    128.19
         avg       = 113,489.58
         max       = 530,022.84
F(   5, 2710.4)    =      3.09
Prob > F           =    0.0087
```

# Average RVI

- MI推定値の変動(分散)は代入した1セットのデータ内変動と、複数の代入セット間の変動により構成される
- つまり、サンプルサイズと代入回数に左右される
- サンプルサイズが変更できない場合は代入回数を増やして、Average RVIを小さくする

# Largest FMI

- Largest Fraction Missing Information (欠損している情報割合の最大値)
- 次の計算式で代入回数を設定する

$$M \geq 100 \times \mathrm{FMI}$$

- いま、0.2762×100=27.62なので、代入回数は28回以上に変更するべき

# Degree of Freedom

- Stataは推定コマンドによって、小/大標本の仮定を自動的に使い分ける
- DFはRVIの逆数に関連する統計量なので、RVIが小さい時は自動的に大きくなる
- 係数に関するF検定は、すべての係数に関する欠損情報の割合は一定であるという仮定を、便宜的においている

# Degree of Freedom

- regressコマンドの場合は小標本を前提とする
- regressコマンドを実行すると、他の推定コマンドの場合とは異なり、欠損値が存在しない場合の自由度Complete DFを追加表示する
- これは完全にデータが揃っている場合の、理想的な状態の自由度である
- データが完全に揃っていても自由度が小さい時に注意する

# 2.データ拡大アルゴリズム

- 欠損値の発生メカニズムがモノトーンパターンに従っていない場合はどう対応する?
- 第2回ではEMアルゴリズムを紹介したが、解析的に条件付き期待値を求める必要があった
- 一般的にはベイズ統計を利用したデータ拡大アルゴリムを利用することが多い
- ベイズ統計では事前分布と尤度、そしてMCMCを利用して事後分布の最頻値を求める
- Tanner and Wong (1987)の提案したベイズ統計によるデータ拡大アルゴリズムは、EMアルゴリズムと似たテクニックを利用して変数の分布を求める

# データ拡大アルゴリズム

- Tanner and Wong (1987)
- 完全データなxは観測データyと欠損したzにより構成されるものと考える

$$x^T = (y^T, z^T)$$

- 完全データxによるパラメータθの事後分布をπ(θ|x)とすると、それは観測データyを使って、

$$\pi(\theta|y) = \int \pi(\theta|x) f(z|y) dz = \boxed{\int \pi(\theta|y,z) f(z|y) dz}$$

ただし、

$$f(z|y) = \int f(z|y,\theta)\pi(\theta|y) d\theta$$

モンテカルロ積分による近似計算

# モンテカルロ積分近似

$$f(z|y) = \int f(z|y,\theta)\pi(\theta|y)d\theta$$

Step1.π(θ|y)に従う乱数$\theta_i$を作成

Step2.f(z|y,θi)に従う乱数$z_i$を作成

このようにして作成したziを利用して、π(θ|y)に対する次の近似式を計算する

$$\frac{1}{m}\sum_{i=1}^{m}\pi(\theta|y,z_i)$$

*データ拡大アルゴリズムは欠損値ziを何回も作成するところがポイント

# mi impute mvn

- Stataの多重代入コマンドはmi impute mvn
- データ拡大アルゴリズムはベイズ統計のシミュレーション手法あるMCMCを利用する
- 欠損値のパターンはモノトーンに限定しない。どんな欠損パターンでも利用可能
- ただし、欠損している変数は連続変数で、正規分布に従うことを仮定する
- MCMCを利用しているので、代入実行後に診断を行う必要がある

# 例題:ヘドニックモデルの推定

- 住宅価格の決定要因を探すヘドニックモデルは単純な重回帰モデルであるが、築年数(age)と課税(tax)にいくつかの欠損値がある
- 2つの欠損値の発生メカニズムはMARで、モノトーンパターンにはなっていない
- したがって、2つの変数の分布をきるだけ正規分布に近づけてMIを実行する
- サンプルデータはmhouses1993.dta

# データ

```
. use https://www.stata-press.com/data/r16/mhouses1993,clear
. des
. misstable nested
. misstable pattern
```

```
Missing-value patterns
  (1 means complete)

                |   Pattern
  Percent       |  1  2
----------------+-------------
     56%        |  1  1
                |
     35         |  1  0
      7         |  0  0
      2         |  0  1
----------------+-------------
    100%        |
```

ネストしていない

```
Variables are  (1) tax  (2) age
```

# 正規性

- 2つの変数の対数をとって、正規化する

```
. gen lnage = ln(age)
. gen lntax = ln(tax)
```

- 多重代入の設定

```
. mi set mlong
. mi register imputed lnage lntax
. mi register regular price sqft nfeatures ne custom corner
```

- 最終的に推定するヘドニックモデルではlnageとlntaxは元の変数に戻すことに注意

# 代入

- データ拡大アルゴリズムによる多重代入の実行

. mi impute mvn lnage lntax = price sqft nfeatures ne
custom corner, add(20)

```
Performing EM optimization:
note: 8 observations omitted from EM estimation because of all imputation variables
      missing
  observed log likelihood =   112.1464 at iteration 48
```

代入値の初期値を求めるEMアルゴリズムの繰り返し計算回数

# 稼働検査期間

- EMアルゴリズムの計算回数48がデフォルトの稼働検査期間(burn-in)100よりも小さいことが必要

```
Prior: uniform                                  Iterations =     2000
                                                   burn-in =      100
                                                   between =      100
```

| | Observations per *m* | | | |
|---|---|---|---|---|
| Variable | Complete | Incomplete | Imputed | Total |
| lnage | 68 | 49 | 49 | 117 |
| lntax | 107 | 10 | 10 | 117 |

# MCMCの収束

- 稼働検査期間が条件を満たしていれば、MCMCによるパラメータの分布はうまく収束できている可能性が高い
- Schafer (1997)は複数のモデルパラメータの定常状態への収束状況を判定するためのworst linear functionを提案した

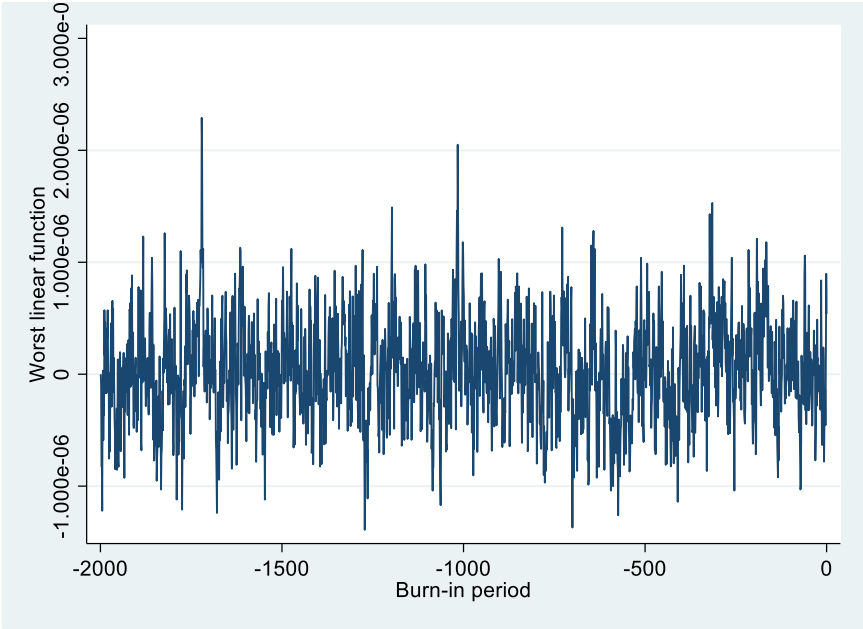
```
. mi impute mvn lnage lntax = price sqft nfeatures ne ///
        custom corner, mcmconly burnin(2000) rseed(23) ///
        savewlf(wlf, replace)

. save mhouses1993out,replace
```

# WLF

- wlfファイルに保存した関数値が定常状態になっていることを確認する

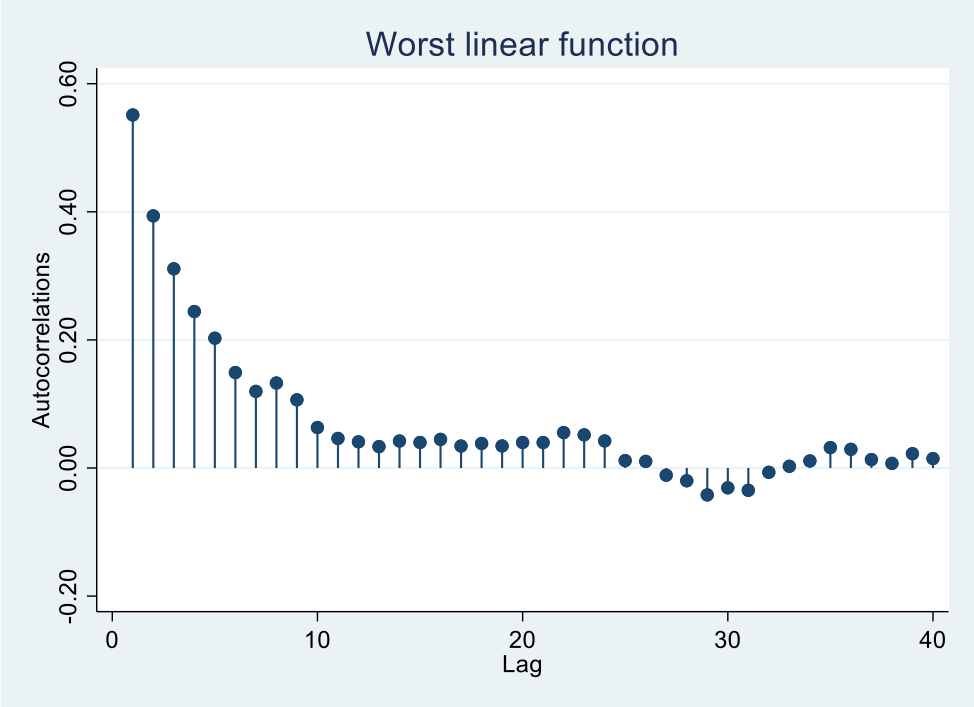
```
. use wlf, clear
. tsset iter
. tsline wlf, ytitle(Worst linear function) xtitle(Burn-in period)
```



# WLFの自己相関

- WLF値の自己相関を確認する。自己相関の変化にトレンドがなく、短いラグでゼロに収束することが望ましい

```
. ac wlf, title(Worst linear function) ytitle(Autocorrelations)
                ciopts(astyle(none)) note("")
```



# MI推定

- データセットをもとに戻し、MI推定を実行する。対数変換して登録した2つの変数を登録しなおす

```
. use mhouses1993out,clear
      *MIによるモデル推定
. mi passive: gen newage = exp(lnage)
. mi passive: gen newtax = exp(lntax)
. mi estimate:reg price newage newtax sqft nfeatures ne
```

# MI推定

- 任意の欠損パターンのデータで体重代入を実行したところ、ほとんどの変数の有意性は失われた

| price | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| newage | -.4416874 | .7900943 | -0.56 | 0.582 | -2.084179 | 1.200804 |
| newtax | .6095571 | .1479177 | 4.12 | 0.000 | .3118694 | .9072448 |
| sqft | .3096159 | .0916235 | 3.38 | 0.002 | .1237405 | .4954912 |
| nfeatures | 8.272231 | 14.22707 | 0.58 | 0.562 | -20.01092 | 36.55538 |
| ne | 2.771427 | 35.54388 | 0.08 | 0.938 | -67.71154 | 73.25439 |
| _cons | 51.71564 | 67.63651 | 0.76 | 0.447 | -82.86647 | 186.2978 |

# まとめ

- 欠損値の発生メカニズムがMARで、モノトーンパターンである時と、無い時の多重代入の実行方法を紹介した
- モノトーンパターンでない時、変数には正規性を仮定した。必要に応じて正規化して、データ拡大アルゴリズム(DA)を利用する
- DAを利用したら、MCMCの収束判定を行う。問題があるときは、モデルを再考する
- WLFは複数のパラメータの収束状況を単一の尺度で判断するための統計量である

# 参考文献


- ALLISON, P.D. (2002). Missing Data. Sage Publications
- STATA MULTIPLE IMPUTATION 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。

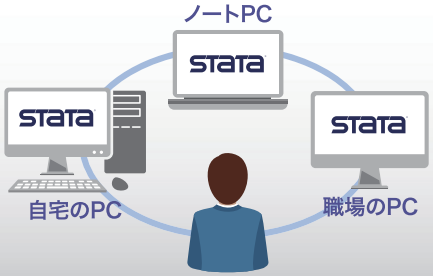
## ライセンスはサブスクリプションがおすすめ！

### 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

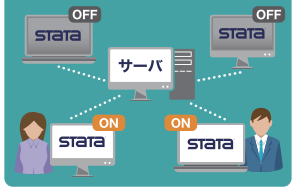
シングルライセンスを複数人でまとめてご購入いただく際のボリュームディスカウント商品です。


#### 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



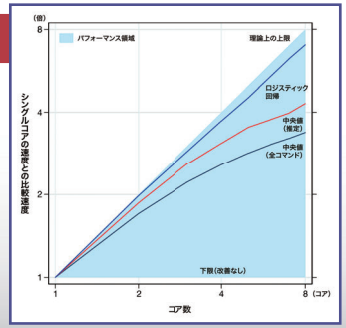

**1台の仮想マシンにユーザがアクセスする運用**
例:2ライセンスの場合


## 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/