Stataによる多重代入

株式会社　ライトストーン

2020年9月

第4回 正規性が仮定できない時

- 欠損値がモノトーンでなく、当該変数に正規性が仮定できないときの代入方法について説明する
  - mi impute pmmは1変数(連続変数)に欠損がある場合の対応方法
  - 2つ以上の変数(連続変数)に代入を行いたい(mi impute chained)
  - 今回紹介する内容はMIのPDFマニュアルからの要約である。より詳しい情報が必要な場合はPDFマニュアル[MI]を参照のこと
  - キーワード:FCS, MICE, PMM, SRHI

1.mi impute pmm

- 第1回で多重代入の実行例として紹介したmi impute regressによく似た方法としてmi impute pmmというものがある
- mi impute regressはパラメータの不確実性を誤差分散経由で取り込んで代入値を用意した
- 欠損値のある変数は正規分布することを仮定していた
- mi impute pmmは正規分布に従わない変数に欠損値が存在する場合に利用する
- pmmはpredictive mean matchingの略

mi impute pmm

- 欠損値のある変数をxとする

$$x_i | z_i \sim N(z_i' \beta, \sigma^2) \qquad (1)$$

  - 完全データzによる上記のような回帰モデルを考える。xとzは観測できたグループに分ける(zは全部観測できている)

$$x = (x_0, x_m), \ z = (z_0, z_m)$$

	x	z_1	z_2
1	NA	2	1
2	NA	−3	0.5
3	3	−2	1
4	3.5	−3	0
5	1	1	5

Xmは1行目と2行目のx

mi impute pmm

Step4.Step2とStep3を繰り返してMセットの完全データセットを得る

Step2とStep3はposterior predictive distribution(事後予測分布)を利用したシミューレション(Gelman et al.[2014, 354-357])と呼ばれている

mi impute pmm

- knn()で利用する整数が小さいほど、代入値間の相関が強くなる。結果としてMIによる点推定値の分散が大きくなる
- 一方で整数値が大きいほど、MI推定値のバイアスが大きくなる
- Schenker and Taylor (1996), Morris, White, and Royston (2014)

利用例

- 変数bmiの分布は正規分布よりも、やや右裾が厚いので、mi impute regressの代わりにpmmを利用する

```
. use https://www.stata-press.com/data/r16/mheart0,clear
. mi set mlong
. mi register imputed bmi
. mi impute pmm bmi attack smokes age hsgrad ///
            female, add(20) knn(5)
. mi estimate: logit attack smokes age bmi hsgrad female
```

mi impute pmm

attack	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
smokes	1.208556	.3587367	3.37	0.001	.5054411	1.911671
age	.0358055	.0154283	2.32	0.020	.0055664	.0660447
bmi	.1072579	.0487065	2.20	0.028	.011685	.2028309
hsgrad	.1595527	.4034737	0.40	0.693	-.6312422	.9503477
female	-.1114382	.4168139	-0.27	0.789	-.9283816	.7055052
_cons	-5.548617	1.704012	-3.26	0.001	-8.890166	-2.207068

mi impute pmm

ここで第3回で利用したコマンドを振り返る

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

- 欠損値bmiとageの欠損値はモノトーンパターンに従っている
- したがって、1変数ごとに独立して代入を行なえばよい
- bmiは正規性を仮定して回帰モデルを利用して代入値を計算する
- ageの分布に偏りがあり、pmmを利用して差の小さい3つのフィット値からランダムに一つを選択し、ageに代入する

2.Chained Equations

- 欠損値がモノトーンパターンに従っていない
- 目的とする推定モデルの2つ以上の説明変数に欠損値が存在する
- 欠損値のある変数は正規分布に従っていない
- このような状況下で利用する多重代入の手法として MICE (Multivariate Imputation using chained equations) というものがある
- MICEはFCS(Fully Conditional Specification)や、SRMI(Sequential Regression Multivariate Imputation)とも呼ばれる

MICE

- MICEは一変数ごとに代入を行う方法で、ベイズ統計で利用するギブスサンプリングに似た計算手法である

$$
\begin{aligned}
X_1^{(t+1)} &\sim g_1\left(X_1 | X_2^{(t)}, \ldots, X_p^{(t)}, Z, \phi_1\right) \\
X_2^{(t+1)} &\sim g_2\left(X_2 | X_1^{(t+1)}, X_3^{(t)} \ldots, X_p^{(t)}, Z, \phi_2\right) \\
&\ldots \\
X_p^{(t+1)} &\sim g_p\left(X_p | X_1^{(t+1)}, X_2^{(t+1)} \ldots, X_{p-1}^{(t+1)}, Z, \phi_p\right)
\end{aligned}
$$

- 関数gは変数Xごとに変更できる(線形回帰やロジスティックなど)
- Xがすべて連続変数で、gが線形回帰モデルの場合はギブスサンプラーとなる

MICE

- 今回は代入と診断の方法だけに話を絞り、最後のモデル推定は実行しない
- モデル推定はこれまで同様、mi estimate:で実行すればよい

```
. use https://www.stata-press.com/data/r16/mheart8s0,clear
. mi misstable patterns, frequency
```

MICE

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

                 |   Pattern
  Frequency      |  1   2
-------------------+-------------
        118      |  1   1
                 |
         24      |  1   0
          8      |  0   1
          4      |  0   0
-------------------+-------------
        154      |

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

*ageとbmiに欠損があり、モノトーンではない

MICE

```
. mi impute chained (regress) bmi age = attack smokes
                                        hsgrad female, add(10)
```

データセット数の累計

直近のコマンドで作成したデータセット数

```
Multivariate imputation                Imputations =       10
Chained equations                            added =       10
Imputed: m=1 through m=10                  updated =        0

Initialization: monotone                Iterations =      100
                                           burn-in =       10

                 bmi: linear regression
                 age: linear regression
```

全繰り返計算回数。add()とburn-inの積になる

MICEの計算でt=1より前に実行する稼働検査期間数

初期値の計算にはmonotoneの計算手法を利用する

MICE

```
. mi impute chained (regress) bmi age = attack smokes
                                        hsgrad female, add(10)
```

Variable	Observations per *m*			
	Complete	Incomplete	Imputed	Total
bmi	126	28	28	154
age	142	12	12	154

(complete + incomplete = total; imputed is the minimum across *m* of the number of filled-in observations.)

代入は欠損値の少ない変数ageから順番に実行される。コマンド入力の順番には関係しない

異なる代入方法を利用する

- 代入したデータセットがメモリー上に残るので、別の代入手法を用いる場合は、データの読み込みから開始する方が良い

```
. use https://www.stata-press.com/data/r16/mheart8s0,clear
. mi impute chained (pmm, knn(3)) bmi (reg) age = attack
                                    smokes hsgrad female, add(5)
```

```
bmi: predictive mean matching
age: linear regression
```

	Observations per *m*			
Variable	Complete	Incomplete	Imputed	Total
bmi	126	28	28	154
age	142	12	12	154

MICEの収束

- 代入した値がほぼ一定の範囲に収まっていることが大事
- 代入値を変数ごとに取り出して、その分布をトレースプロットで確認する
- ここではすべての代入値に興味はなく、その分布だけを調べたいので、chainonlyというオプションを利用する
- そして、その分布に関する統計量をsavetraceオプションでファイル保存する
- 個々の代入値は残っていないのでaddオプションは利用しない

MICEの収束

. use https://www.stata-press.com/data/r16/mheart8s0,clear
. mi impute chained (pmm, knn(3)) bmi (reg) age = attack smokes hsgrad female, chainonly burnin(100) replace savetrace(impstats,replace) rseed(1359)

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

Performing chained iterations ...

Note: No imputation performed.
```

MICEによる代入計算を実行したが、個々の代入値の代わりにその統計量だけをimpstatsに保存した

代入値の分布を確認する

```
. use impstats,clear
. describe
```

variable name	storage type	display format	value label	variable label
iter	byte	%12.0g		Iteration numbers
m	byte	%12.0g		Imputation numbers
age_mean	float	%9.0g		Mean of age
age_sd	float	%9.0g		Std. Dev. of age
bmi_mean	float	%9.0g		Mean of bmi
bmi_sd	float	%9.0g		Std. Dev. of bmi

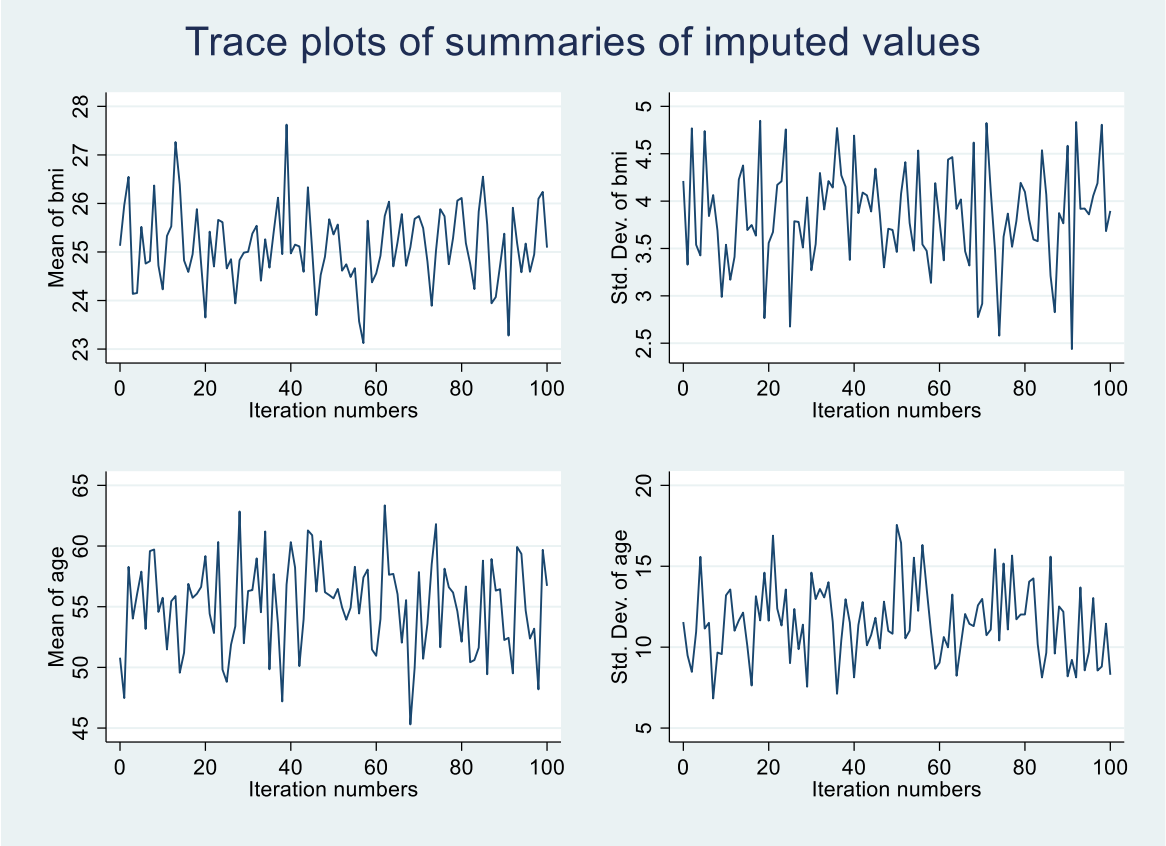
age(12個)とbmi(28個)の欠損値に代入した値の平均と標準偏差に関する情報が保存されている

代入値の分布を確認する

- 繰り返しの代入回数(0から100)が入っている変数iterを時間変数として4つのグラフを作成し、最後にタイル張りのグラフとする

```
. tsset iter
. tsline bmi_mean, name(gr1) nodraw
. tsline bmi_sd, name(gr2) nodraw
. tsline age_mean, name(gr3) nodraw
. tsline age_sd, name(gr4) nodraw
. graph combine gr1 gr2 gr3 gr4, title(Trace plots of
                                  summaries of imputed values) rows(2)
```

トレースプロット



繰り返し計算の過程で、代入値にトレンドが発生していないことをチェックする

mi impute pmm

- pmmの処理

Step1.3-5行目のデータで回帰モデルを推定

Step2.回帰モデルの誤差分散でカイ二乗分布、係数推定値は正規分布でそれぞれ乱数を利用してシミュレーションを実行する(mi impute regressと同じ)

Step3.1-2行目のデータにおいてStep2で得た推定値を使ってNA部分をもとめる。例えば1行目のNAが3.1だとすると、一番近いのは3行目の3。これを代入値とする(k=1のとき)。k=2以上の場合、差の小さいものを2個(3と3.5)をリストアップし、ランダムに一つ選ぶ

	x	z_1	z_2
1	NA	2	1
2	NA	-3	0.5
3	3	-2	1
4	3.5	-3	0
5	1	1	5

被説明変数の欠損値

- ここまでの話はすべて解析モデルの説明変数に欠損値が存在する場合の話であった
- では、解析モデルの被説明変数に欠損値が存在する場合はどうすれば良いか?
- 問題の被説明変数と相関が0.5以上あるような補助変数が存在する場合のみ、代入を行う価値がある。相関の弱い補助変数を利用しても、不確実性が増すだけで意味がない

まとめ

- 代入の対象となる変数で正規性が仮定できない場合の代入方法としてpmmがある
- データセットの中に欠損値のある変数が2つ以上存在し、モノトーンパターンに従っていない場合はchained equationを利用する
- chained equationの場合、トレースプロットで代入値の分布状態を確認する。トレンドが存在しないことが大事

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。

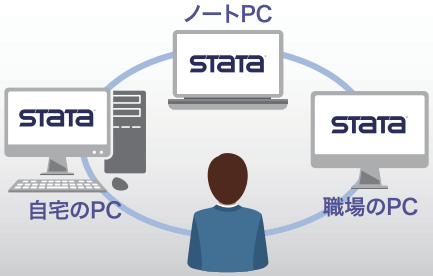
ライセンスはサブスクリプションがおすすめ！

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

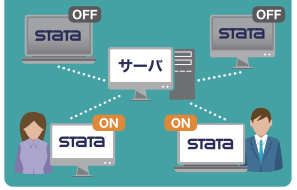
シングルライセンスを複数人でまとめてご購入いただく際のボリュームディスカウント商品です。

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



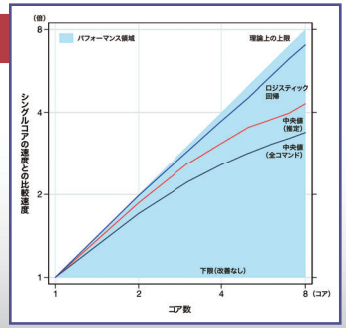
1台の仮想マシンにユーザがアクセスする運用
例:2ライセンスの場合

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/