mwp-102

多重代入機能の概要　【 評価版 】

多重代入 (MI: multiple imputation) 機能は欠損値を補完し、その上で推定を行う機能体系を提供します。本 whitepaper では多重代入の機能概要を解説すると共に、簡単な用例について紹介します。

1. 多重代入とは

ここでは多重代入機能がなぜ必要になるかを具体的なデータを用いて説明します。使用する Example データセットは心臓発作に関する症例対照研究データを記録した `mheart0.dta` です。

```
. use https://www.stata-press.com/data/r18/mheart0.dta [*1]
(Fictional heart attack data; BMI missing)
```

データセット中には 154 人の被験者に関するデータが記録されているわけですが、そのデータの一部をリスト出力すると次のようになります。

[*1] メニュー操作：File ▷ Example Datasets ▷ Stata 18 manual datasets と操作、Multiple-Imputation Reference Manual [MI] の mi intro substantive の項よりダウンロードする。

```
. list attack smokes age bmi hsgrad female in 14/21, separator(0)
```
[*2]

	attack	smokes	age	bmi	hsgrad	female
14.	0	1	83.78423	29.83765	1	1
15.	0	0	73.5787	25.32784	1	0
16.	1	1	68.71399	22.45127	1	1
17.	1	0	49.79274	32.00198	1	0
18.	0	1	29.96287	.	1	0
19.	0	1	59.51569	.	1	1
20.	1	1	36.57973	31.22007	1	1
21.	0	1	59.25924	23.93727	0	1

変数 attack は症例 (attack＝ 1) か対照 (attack＝ 0) かを識別するための指標変数です。このような 2 値のアウトカムの推定には通常ロジスティック回帰が用いられますが、その際の説明変数としては次の 5 つを対象とすることにします。

変数	意味
smokes	現時点での喫煙慣行
age	年齢
bmi	肥満指数 (body mass index)
hsgrad	高校卒業生
female	性別 (0 = 男性、1 = 女性)

ここで注意を要するのは変数中 bmi に欠損値が含まれているという点です。

```
. logit attack smokes age bmi hsgrad female
```
[*3]

```
. logit attack smokes age bmi hsgrad female

Iteration 0:  Log likelihood = -91.359017
Iteration 1:  Log likelihood = -79.374749
Iteration 2:  Log likelihood = -79.342218
Iteration 3:  Log likelihood =  -79.34221
```

[*2] メニュー操作： Data ▷ Describe data ▷ List data
[*3] メニュー操作： db logit

```
Logistic regression                                     Number of obs =    132
                                                        LR chi2(5)    =  24.03
                                                        Prob > chi2   = 0.0002
Log likelihood = -79.34221                              Pseudo R2     = 0.1315

------------------------------------------------------------------------------
      attack | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      smokes |   1.544053   .3998329     3.86   0.000     .7603945    2.327711
         age |    .026112    .017042     1.53   0.125    -.0072898    .0595137
         bmi |   .1129938   .0500061     2.26   0.024     .0149837     .211004
      hsgrad |   .4048251   .4446019     0.91   0.363    -.4665786    1.276229
      female |   .2255301   .4527558     0.50   0.618    -.6618549    1.112915
       _cons |  -5.408398   1.810603    -2.99   0.003    -8.957115    -1.85968
------------------------------------------------------------------------------
```

推定に用いられた観測データ (observations) の数は 154 ではなく 132 とレポートされています。これは bmi が欠損値であるデータが 22 件存在することによるものです。上記のリスト出力を例に取るなら、bmi が欠損値であるが故に、被験者 18 と 19 に関する情報はすべて無視されてしまったことになります (listwise/casewise deletion)。計測値を含んだ共変量についてはその情報を最大限活用しようという趣旨で開発された技術が多重代入法です。

多重代入 (multiple imputation) 法は欠損値をシミュレーションによって補完しようとする統計手法であり、次に示す 3 つのステップから構成されます。

(i) 代入ステップ – ある選択された代入モデルに基づき、M 組のデータ（欠損値を補完したデータセット）を生成する。
(ii) 推定ステップ – 補完された M 組のデータ各々について推定を実行する (completed-data analysis)。
(iii) プーリングステップ – M 個の推定結果を結合し、単一の多重代入推定結果を抽出する。

通常、ステップ (ii) と (iii) は分離されない形で分析ステップを構成することになります。

代入の段階で $M = 1$ とする手法は単一代入 (single imputation) 法と呼ばれます。しかし単一代入の場合、補完された値は既知のデータとして扱われることになるため、推定値の分散は過少評価されてしまう傾向があります。これに対し多重代入の場合には複数組のデータを維持した状態で推定を行うため、欠損値に起因するサンプリング変動 (between-imputation variability) を考慮に入れた形での推定が行われます。

2. 理論的裏付け

評価版では割愛しています。

3. M の適正値

評価版では割愛しています。

4. 欠損値に関する仮定

評価版では割愛しています。

5. 欠損データのパターン

評価版では割愛しています。

6. Proper な代入手法

評価版では割愛しています。

7. 多重代入データの分析

評価版では割愛しています。

8. MI の用例

Stata では代入ステップからプーリングステップに至るまで、MI 分析全般をカバーする機能を提供しています。

代入ステップについて言えば、それは単一変数を対象にした形でも、あるいは複数変数を対象にした形でも実行できます。代入手法としては様々な変数種別に対応できるものや、他変量正規分布に基づく反復的マルコフ連鎖モンテカルロ法など、種々の手法がサポートされています。詳細については [MI] **mi impute** (*mwp-378*) を参照ください。

分析ステップとプーリングステップは結合された形で `mi estimate` コマンド（[MI] **mi estimate** (*mwp-379*) 参照）として実装されています。種々のタイプのモデルがフィットでき、結合された形の係数推定値を得ることができます。またユーザ独自の推定コマンドを用意し、それを `mi estimate` prefix を介して実行させることもできます。

これらに加えてデータ操作や診断を支援する種々のコマンドも一式用意されています。コマンドの一覧については [MI] **Intro** を参照ください。

セクション 1 では心臓発作に関するデータに対して通常の `logit` コマンドによるモデルフィットを行ったわけですが、ここでは MI の枠組みを使ったときの分析の流れを紹介します。より一般的な操作ガイドについては [MI] **Workflow** を参照ください。

以下に示す操作のポイントを記しておくと次のようになります。

1) `bmi` の欠損値に対して多重代入法を使って値の補完を行う（具体的には線形回帰モデルによる代入を行う）。
2) 多重に代入されたデータをロジスティック回帰を使って分析する。

最初に Example データセット `mheart0.dta` をロードします。

```
. use https://www.stata-press.com/data/r18/mheart0.dta
(Fictional heart attack data; BMI missing)
```

ここではダイアログインタフェースを使って MI の操作を紹介しますが、その場合、

- Statistics ▷ Multiple imputation

と操作することによって表示される MI 制御パネルが操作の中心となります。

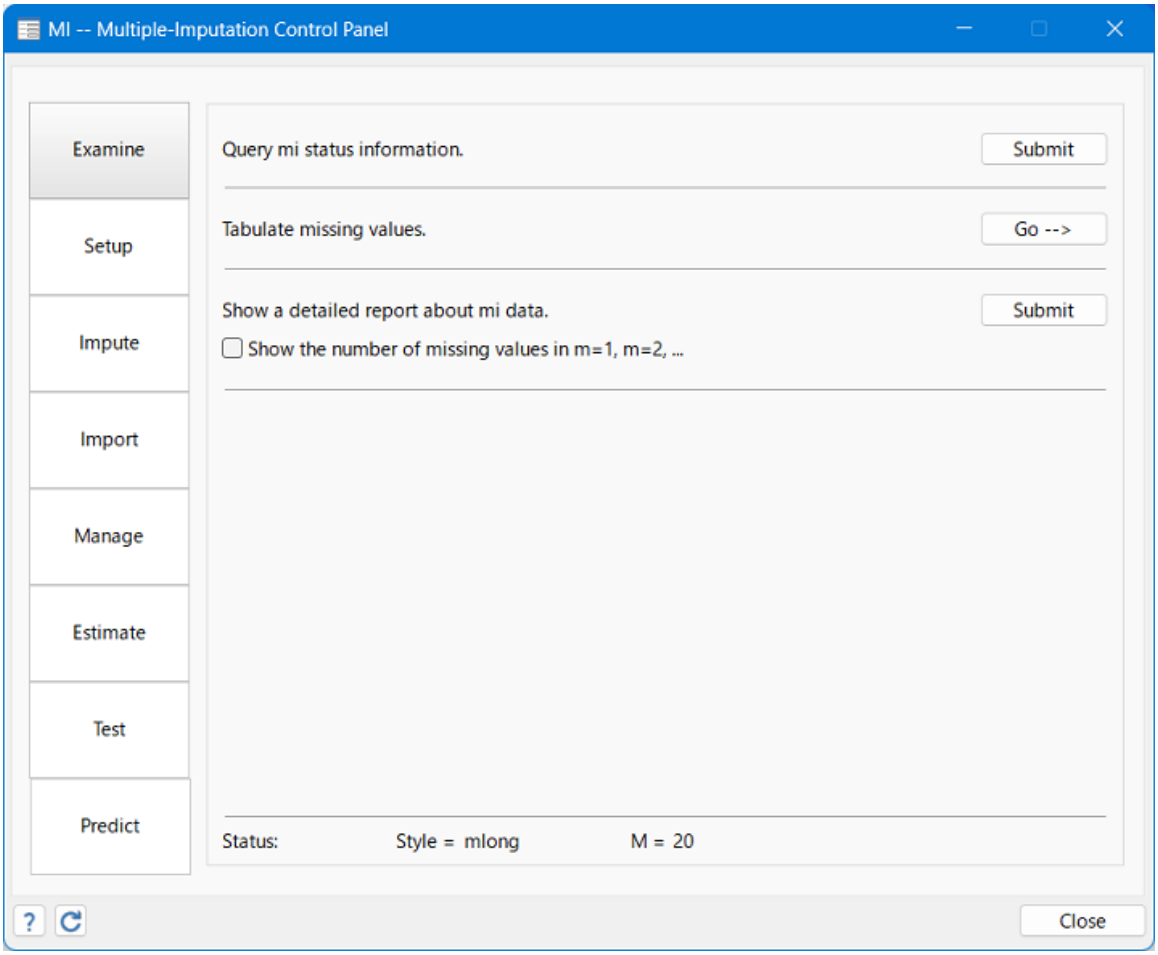

図 1　MI 制御パネル

(1) MI **データセットの初期設定**

評価版では割愛しています。

(2) **変数の登録**

評価版では割愛しています。

(3) **多重代入の実行**

評価版では割愛しています。

(4) **推定の実行**

評価版では割愛しています。

9. まとめ

評価版では割愛しています。

補足１ – 注意事項

評価版では割愛しています。

■