欠損値があるデータを扱うMIプロシジャについて

小野裕亮
テクニカルサポートグループ
（株）SASインスティチュートジャパン

jpnyco@jpn.sas.com

内容：

multiple imputation（多重代入；以後MI）を、
SASで行なう方法を紹介
→プログラミングが中心。
×理論 ×実務 （？？？）

- 1）MIを行なうためのソフトウェア必要要件
- 2）Version6では？
- 3) MIの一般論
- 4）PROC MIについて
- 5）PROC MIANALYZEについて

ソフトウェア必要要件

＊MIを行なうプロシジャ＊

＊バージョン8.1～

＊SAS/STATプロダクト

＊評価版（experimental）

- バージョン6には存在しない。
- 日本：バージョン8.1（日本語版）を、
- Microsoft Windows版のみリクエストに応じて出荷中→ *http://www.sas.com/japan/service/v8/index.html*
- *11月の現状：マニュアルなどは英語,*
- *Base SASの一部分メッセージが日本語。*

The Power to Know.

V6における欠損値（. ）の扱い

- ほとんどのプロシジャはリストワイズで削除

x1	x2	x3
10.5	19.1	30.9
14.2	17.5	35.5
.	20.1	35.4
12.8	20.4	35.2
10.4	15.2	33.8

The Power to Know.

SAS V6における現状

- A)単一代入(single imputation)
  - 何らかの値で、事前にとりあえず穴を埋める努力
  - 1)STANDARDプロシジャ(Base SAS) … 平均。Easy!
  - 2)ユーザ自身のマクロプログラム、データステップなどによる涙ぐましい努力
    - 例)Hot-Deck Imputation … 米国SUGI 1998, USA Census
- B)CORRプロシジャ (Base SAS)

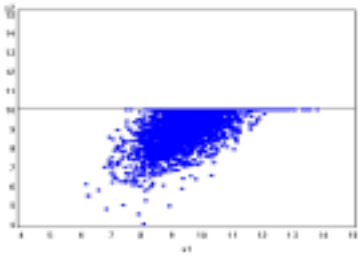
  ・デフォルトは、ペアワイズで削除

C)MIXEDプロシジャ(SAS/STAT)

…Observed-data likelihood

SAS V6における現状

D)打ち切りデータ(生存時間モデル・トービットモデル)



E)IMLによって自分自身でプログラミング。。。。。。

例)ペンシルバニア大学のPaul Allison…

余談：SASデータセットの欠損値（「欠損」といってもいろいろあるが）

SASデータセットの数値変数における欠損値

基本　.　　（ピリオド）

特殊欠損値　.A　.B　...（ピリオド＋アルファベット）

分析において、この違いを積極的に使うものは特にない（例外）

MEANSのCLASSステートメント,FREQの単純集計

UNIVARIATEの欠損値数出力

TRANSREG,PRINQUALプロシジャ

The Power to Know.

記号：（SASマニュアル）Shafer？

Yobs ...観測されたデータ

Ymis …観測されなかったデータ

R … 欠損値の位置を表すインデックス

Θ … 完全データに対するモデルパラメータ

ξ … 欠損メカニズムを決定するパラメータ

例） $X2 = \beta 0 + \beta 1\ X1 + \varepsilon$

Pr(X1=.)=p1　　Pr(X2=.) =p2

モデルパラメータ　$\theta = (\beta 0, \beta 1, \sigma)$

欠損メカニズムのパラメータ ξ =(p1 p2)

The Power to Know.

欠損値いろいろ

1）Missing Completely At Random

Yobs, Ymisに依存せず、欠損がランダムに生じている

2）Missing At Random（ランダムな欠損）

Yobsだけに依存して欠損がランダムに生じている。

Pr(R | Yobs, Ymis, ξ) = Pr(R | Yobs, ξ)

条件や方法によっては、欠損が無視できる

… …

3）nonignorable（欠損がinformativeな場合）

欠損のメカニズムをモデル化する必要

The Power to Know.

Multiple Imputation , Observed-data Likelihoodの前提: 欠損を無視できる？

L(θ | Yobs)が最大になるようなθ

| |

L(θ, ξ | Yobs , R) が最大になるようなθ, ξ

1） Missing At Random (or MCAR)

2） θとξがdistinct （ベイズ流: θとξが独立)。

・分析者は、ξには特に関心ない。

・分析者は、完全データ上のモデルパラメータ θに関心

The Power to Know.

喩え話:
2正規変量X1とX2において、X2に欠損値が存在

・A) MCAR

サイコロを振って、X1,X2の一部分を観測しない。

・B) MAR（Yobsのみに依存）。

X1＞1以上の場合には、サイコロを振ってX2の値を観測しない。

・C) Ymisに依存。

X2 ＞ 1の場合には、サイコロを振ってX2の値を観測しない。

The Power to Know.

平均と分散共分散行列の計算
研究者の興味は、μとΣ

A）

リストワイズの削除でもOK。

B)

リストワイズの削除

→ ×X2の平均および分散, 共分散, 相関×

C)

リストワイズの削除

→ × X2の平均および分散, 共分散, 相関×

The Power to Know.

補足:
REG:X2＝β0＋β1X1＋ε:
研究者の興味はβ0, β1, σ

- A)
リストワイズの削除でもOK（推定効率,検出力 ↓）
- B)
リストワイズの削除でもOK
推定効率,検出力 ↓。標準化偏回帰係数, R2乗値は×
- C)
リストワイズ削除:　β0,β1,σの推定値とも×。
打ち切りを考慮して、尤度関数を定義＝LIFEREG

The Power to Know.

補足:
REG:X1＝β0＋β1X2＋ε:
研究者の興味はβ0, β1, σ

-A)
リストワイズの削除でもOK（推定効率,検出力 ダウン）。

-B)
リストワイズ削除:　β0,β1,σの推定値とも×。

-C)
リストワイズ削除:　β0,β1,σの推定値ともOK（推定効率,検出力 ダウン、標準化偏回帰係数、R2乗値は×）

The Power to Know.

V6の機能：MIXEDプロシジャ
“Observed-data likelihood”を最大化

“Observed-data Likelihood ” を最大化　→
L(θ , ξ ｜Yobs , R) ではなく、L(θ | Yobs)を最大化。

? max(L(θ , ξ ｜Yobs , R))
→　max (L(θ | Yobs))　?

本当は、max(L(θ | Ymis, Yobs))がベスト。

The Power to Know.

V6の機能：MIXEDプロシジャ
多変量分析が目的ではないので…。

```
proc transpose data=data1 name=var out=out1;
  by id;
run;
proc mixed data=out1;
   class id var;
   model col1=var /ddfm=kenwardroger solution ;
   repeated var /subject=id type=un r rcorr;
run;
```

（注：小標本のときの振舞い）

The Power to Know.

MI推定を行なう手順
（2つに分類される）

MI実行者の任務

… “*モデルA*”に従う乱数で穴埋めを実行。この処理を複数回実行し、穴埋めされた複数個のデータセットを作成。

分析者の任務

… 各々の完全化されたデータセットを、“*モデルB*”にあてはめる。穴埋めによるバラツキも考慮して、複数の結果をまとめる。

The Power to Know.

MI推定のSASプログラム
… 3ステップに分けられる。

MI実行者

１）PROC MIを用いる。（重要任務：モデルAの選択）

分析者

２）既存プロシジャでモデルBをあてはめる。

3) 結果をまとめるためPROC　MIANALYZEを使用。

The Power to Know.

ステップ1：
MIプロシジャを実行する。

```
proc mi data=data1 out=out1 nimpu=70
                           seed=654321;
   multinormal method=MCMC
                ( prior=jeffreys
                  initial=EM(MLE)
                  biter=50
                  chain=multiple
                 );
   var x1 x2 x3;
run;
```

The Power to Know.

ステップ2：
分析者は既存プロシジャを実行する。

```
PROC  REG  DATA=WORK . OUT1
  OUTEST=WORK.OUT2  COVOUT ;
   BY  _IMPUTATION_;
   MODEL   X3=X1   X2 ;
RUN ;
```

The Power to Know.

ステップ3：分析者は、さらに MIANALYZEプロシジャを実行。

```
proc mianalyze data=work.out2 edf=97;
  var Intercept x1 x2;
run;
```

The Power to Know.

ステップ1：PROC MI 代入モデルを選ぶ。

PROC MIで用意されている方法は、全部で3つ。
MULTINORMAL　METHOD= ＊

（1）ロジスティックモデルに基づく方法 PROPENSITY
（propensity score method)
×（2）回帰モデル法 REGRESSION　Version8.1×
（3）多変量正規をMCMCで発生させる方法 MCMC

（制限）1,2は、欠損が単調(monotone)である時しか適用できない。

The Power to Know.

Monotoneな欠損構造

T1	T2	T3	T4
X	X	.	.
X	X	X	X
X	.	.	.
X	X	X	.

＊前が欠損だったら、後ろも欠損。

The Power to Know.

（1）Propensity Score Method METHOD=PROPENSITY

（a）「欠損値 or 非欠損値」の2値を従属変数としてロジスティック
（b）ロジスティックモデルの予測値が近い観測値を抽出して、欠損値を穴埋め。

－欠損値 or 非欠損値の情報しか用いていない。

付随するオプション：
METHOD=PROPENSITY（ GROUP= n ）
－何グループにするか？

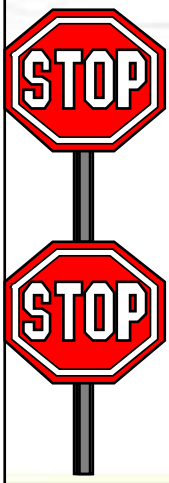
The Power to Know.

(2)回帰モデルに基づく方法 METHOD=REGRESSION

回帰モデルに基づき、穴埋めを行なっていく。
欠損構造がmonotoneの時のみ。
でも、MCMC法よりも高速。
基本的に、多変量正規分布。

・非欠損値の部分から、β および σ の推定値
→ β^* , σ^*を乱数で生成
→ $Y = X\beta^* + \sigma^* z$ で欠損値部分を埋める。

(2)回帰モデルに基づく方法 METHOD=REGRESSION;



重要：Version8.1では、METHOD=REGRESSIONは常に間違った結果になっています。Version8.2で、このバグは修正されています。

(3)MCMC法　METHOD=MCMC
多変量正規分布

“Impute”-step と“Posterior”-stepを交互に行なう。

（Ymis｜μ, Σ,Yobs) →(μ, Σ ｜Y)　→ （Ymis｜μ, Σ,Yobs)
→(μ, Σ|Y)→ ….

＜オプション＞

INITIAL＝EM（ BOOTSTRAP = p
MLE (デフォルトは事後分布）

CHAIN＝SIGLE｜MULTIPLE 単鎖 or 複鎖

PRIOR= 事前分布（デフォルトは、Jeffreys)

BITER＝ 各連鎖のまえに、何回、回しておくか？

ITER＝ 単鎖のときの間隔

The Power to Know.

PROC MIステートメントのオプション

幾つのデータ(完全化されたデータ)を作成するか？
NIMPU= k ;

乱数系列のシード値 SEED=

代入値の最大, 最小 MAX= MIN=

用いるデータ名 DATA=

作成する穴埋めデータ名 OUT=

結果を表示しない NOPRINT

The Power to Know.

PROC MIANALYZE

推定値および推定値間の分散共分散行列を与えなければいけない。
入力データの形式は2通り。

・DATA＝
TYPE=“EST” ＆ _TYPE_=“COV”

・PARMS= 推定値
COVB= 分散共分散行列

The Power to Know.

出力される情報：（別紙）

☆MI推定値＝ 各推定値の平均
☆欠損値を埋めることに伴う変動
・Within-impute Variance (W)
・Between-impute Variance (B) →欠損なしの時0
・Total Variance ＝W ＋ B /m ＋ B

☆欠損がMI推定値の変動にどれだけ影響しているか？
- a)fraction of missing information about Q
- b)relative increase in variance due to nonresponse

The Power to Know.

今後の予定　　（11/28 現在）

バージョン8.2で、変更・拡張を行なった。マニュアルも、V8.1のPROC MIは46ページだが、現在、V8.2は72ページ。

バージョン8.2でも、MIおよびMIANALYZEは、「評価版」。バージョン8.2の次バージョンでは、プロダクト版となる予定。

なお、カテゴリーデータに対する処理は、今後の取り組むべき課題。

The Power to Know.

Version8.2に追加される機能（MIプロシジャ）

1. EMステートメント

　MIを行なわずに、EMアルゴリズムで(μ , Σ)の推定だけを行ないたい時。

2. TRANSFORMステートメント　データを変換。

3. Monotone MCMC method （monotone missing patternのデータを作成する）

4. MCMC法において、定常になったかどうかをチェックするための自己相関プロット。

The Power to Know.

バージョン8.1のドキュメント （英語）

- MIプロシジャ （Version8.1)

http://www.sas.com/service/techsup/faq/stat_proc/miproc.html

- MIANALYZEプロシジャ (Version8.1)

http://www.sas.com/service/techsup/faq/stat_proc/mianalyzeproc.html

（PDF形式。Adobe社が無料提供しているAcrobat Readerが必要です）。

The Power to Know.

アウトプット例：

別紙）

The Power to Know.


sas
sas
The Power to Know.