多重代入法による回帰分析

原田悠介

Contents

モチベーション

実データにはほぼ必ず欠測がある!

欠測を無視した分析はバイアスを生む。欠測への適切な対処を

※欠測無視の例（リストワイズ除去）

ID	収入（年収）	世帯人数
A	500	3
B	1000	4
C	NA	1

⇨

ID	収入（年収）	世帯人数
A	500	3
B	1000	4

前提：用いる記号

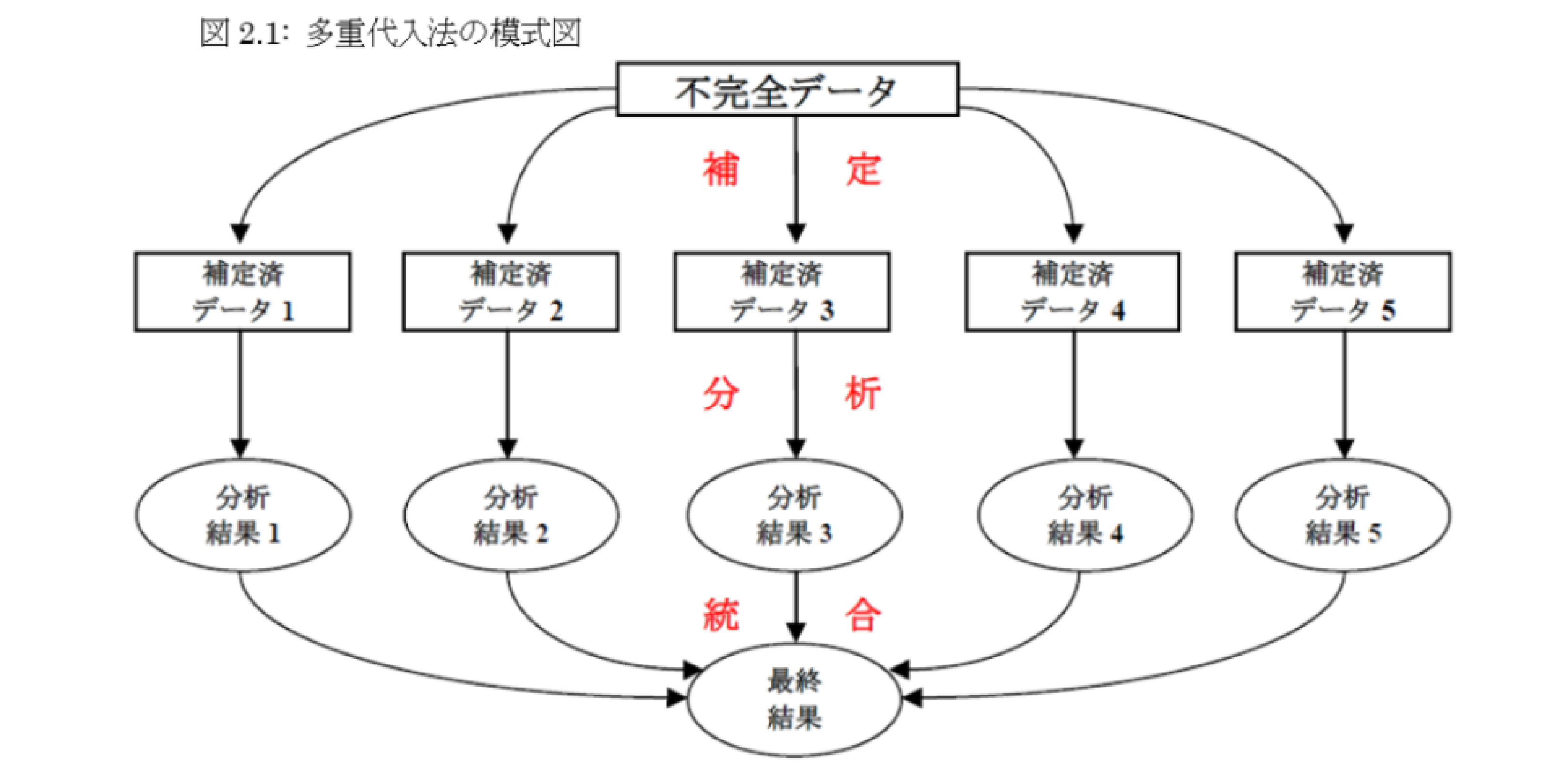
- D：データセット（n×p行列）$D \sim N_p(\mu_p \Sigma)$
  D $= \{Y_{obs}, Y_{mis}\}$
  i = 1,2,…,n （観測値のインデックス）
  j = 1,2,…,p（変数のインデックス）
- R：回答指示行列（n×p行列）

前提：欠測のメカニズム

- MAR（Missing at Random)
  欠測は観測データを条件としてランダム
- MCAR （Missing Completely At Random)
  欠測は完全にランダム
- NMAR（Not Missing At Random）
  ランダムではない欠測、ある値が欠測する確率はその値による

多重代入ではMAR（MCAR）を仮定する

多重代入法の概要



高橋,伊藤(2014)より

多重代入法の概要

①抽出

- 欠測データの分布から独立かつ無作為に抽出されたM個のシミュレーション値によって欠測値を置き換える
- 欠測データの分布は観測できないため、観測データを条件として欠測データの事後予測分布を構築して抽出を行う

②分析

- 目的の分析を行いパラメータを推定する(検定もこの時に行う)

多重代入法の概要

③統合

- $\tilde{\theta}_m$をm番目のデータセットから得られたパラメータの推定値とする
- パラメータが正規分布に従わない場合は変換を行う
- パラメーターの統合: $\bar{\theta}_{\mathrm{M}} = \frac{1}{M}\sum_{m=1}^{M} \tilde{\theta}_m$
- $var(\bar{\theta}_M) = \frac{1}{M}\sum_{m=1}^{M} var(\tilde{\theta}_m) + (1 + \frac{1}{M})\frac{1}{M-1}\sum_{m=1}^{M}(\tilde{\theta}_m - \bar{\theta}_M)^2$

多重代入法のアルゴリズム①

◆DA（Data Augmentation）法

- MCMCに基づく伝統的なアルゴリズム
- θが収束するまで以下の2つのステップを繰り返す

I-step(imputation step): $Y_{mis}^{(t+1)}$を$P(Y_{mis}|Y_{obs},\theta^{(t)})$に基づいて生成する

P-step(posterior step): θ_{t+1}を$P(\theta|Y_{obs},Y_{mis}^{t+1})$に基づいて生成する

多重代入法のアルゴリズム②

◆完全条件付指定（FCS）アルゴリズム

- 条件付密度$P(Y_j|Y_{-j},R,\lambda_j)$によって多変量分布を指定し、ほかの変数を条件として欠損値の代入を行う

多重代入法のアルゴリズム②

1. $P\left(Y_{j,mis} \middle| Y_{j,obs}, Y_{-j}, R\right)$ ：欠測を含む各変数について、それ以外の変数と回答支持行列を条件として代入モデルを構築する
2. 初期値$\tilde{Y}_{j,0}$を設定する
3. 繰り返し回数 t = 1,2,...,T回まで繰り返す
4. $\tilde{Y}_{-j,t} = \left(\tilde{Y}_{1,t}, \ldots, \tilde{Y}_{j-1,t}, \tilde{Y}_{j+1,t-1}, \tilde{Y}_{p,t-1}\right)$：t番目の繰り返しにおける完全データ
5. $\tilde{\lambda}_{j,t} \sim P(\lambda_{j,t} | Y_{j,obs}, \tilde{Y}_{-j,t}, R)$： 代入モデルのパラメータλを抽出する
6. $\tilde{Y}_{j,t} \sim P\left(Y_{j,mis} \middle| Y_{j,obs}, \tilde{Y}_{-j,t}, R, \tilde{\lambda}_{j,t}\right)$ ：代入値の抽出

多重代入法のアルゴリズム③

◆EMBアルゴリズム

- 期待値最大化法(EMアルゴリズム)とノンパラメトリック・ブートストラップを組み合わせた

アルゴリズム

- ノンパラメトリック・ブートストラップ：サイズNの母集団から無作為抽出によってサイズnの標本Sを得る。標本Sを疑似的に母集団とし、さらに再標本S_{boot}を再抽出する。これをM回繰り返す

多重代入法のアルゴリズム③

- EMアルゴリズム：S_{boot}に欠測が含まれる場合、 次の期待値ステップと最大化ステップをパラメータの推定値が収束するまで繰り返す
- 期待値ステップ：$Q(\theta|\theta_t) = \int l(\theta|Y)P(Y_{mis}|Y_{obs};\theta_t)dY_{mis}$
- 最大化ステップ：$\theta_{t+1} = argmax_\theta Q(\theta|\theta_t)$ をθについて最大化する

多重代入を用いた回帰分析

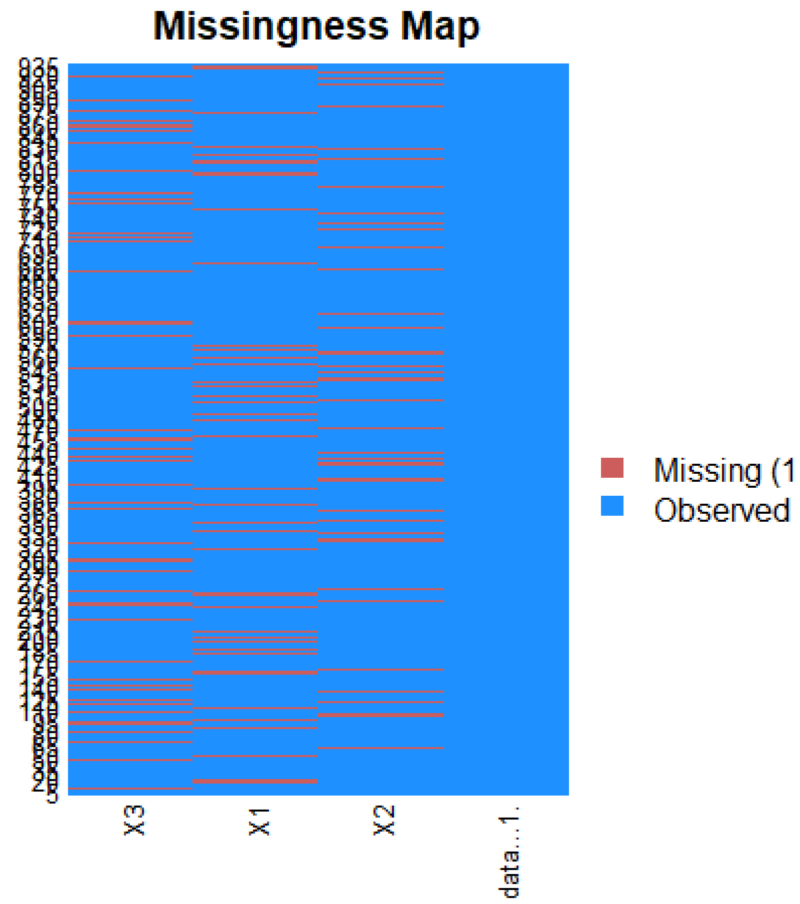
- RのパッケージAmelia(EMBアルゴリズム)を採用

- 使用するデータ：Rのwooldridgeパッケージに含まれるwage2データから、MARにもとづいて欠測を発生させたもの

- 推定式　$\log(\text{wage}) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 teure$
- Wageが中央値より高いならそれぞれ10%の確率で、中央値より低いならそれぞれ20%の確率で説明変数が欠測

多重代入を用いた回帰分析

欠測地図

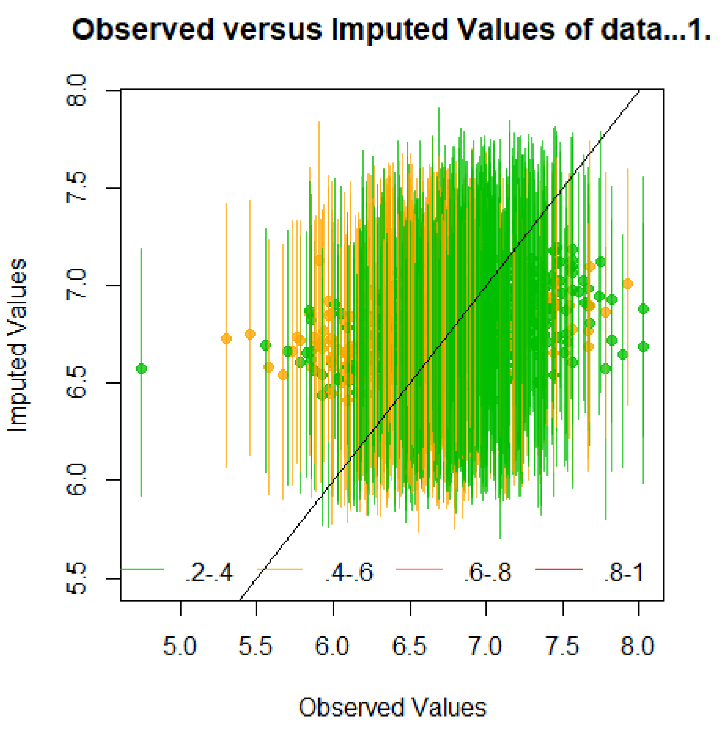


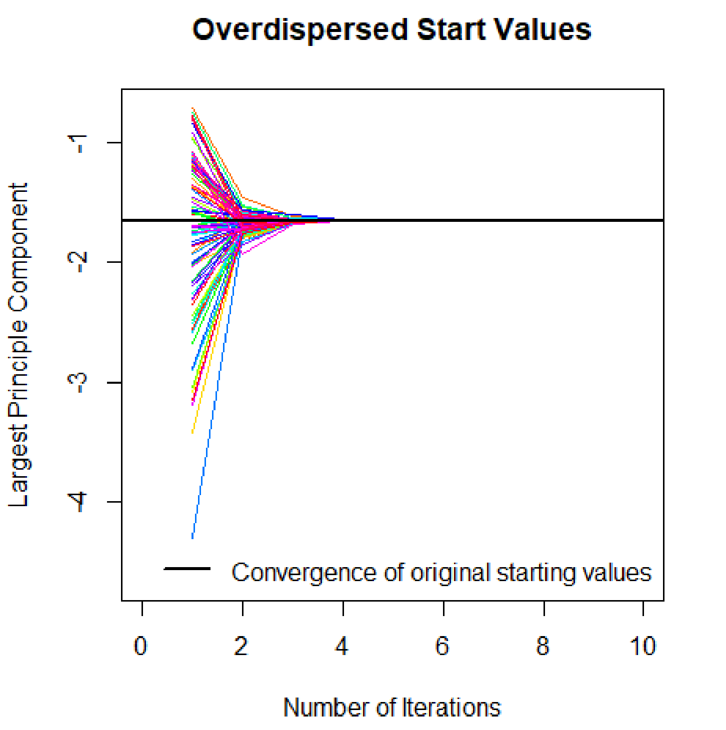
多重代入を用いた回帰分析

```
library(Amelia);library(lattice);library(miceadds)
#代入済データの個数
M<-20
set.seed(1)
#代入
a.out<-amelia(missdata,m=M)
```

多重代入を用いた回帰分析

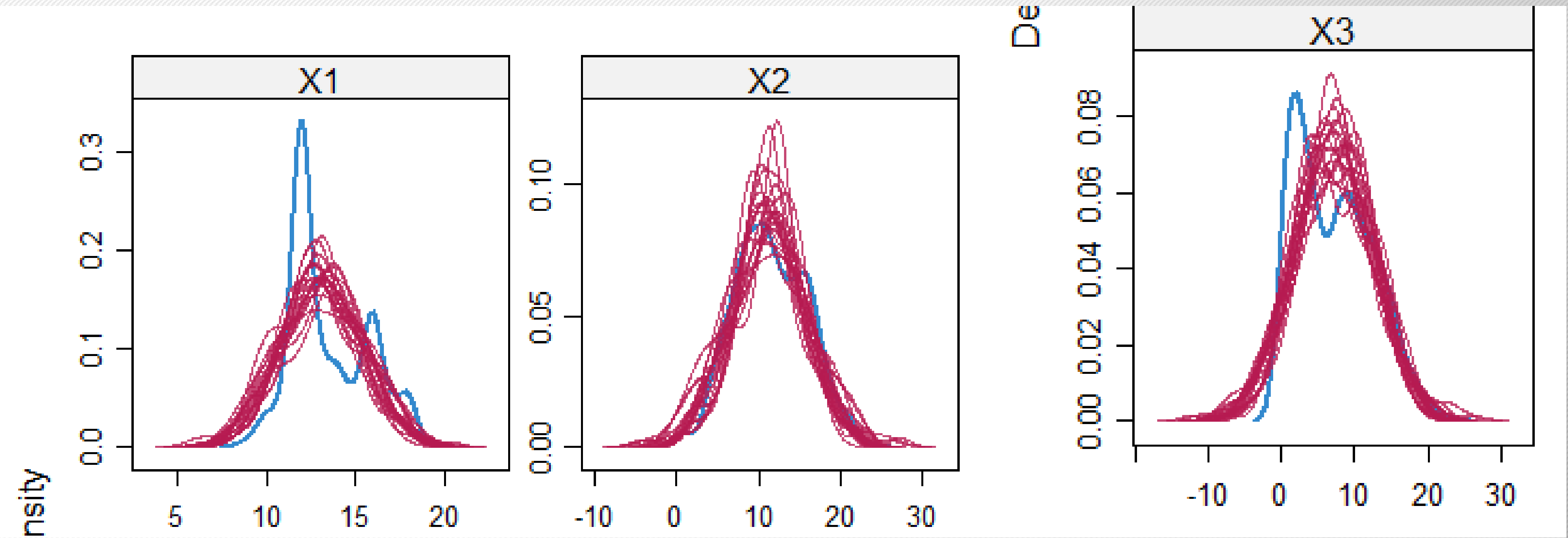
- overimpute(a.out,var=1)　#代入モデルの当てはまりの良さの確認
- disperse(a.out,dims=1,m=100)　#EMアルゴリズムの収束チェック





多重代入を用いた回帰分析

```
a.mids<-datlist2mids(a.out$imputations)
densityplot(a.mids)  #欠測値の密度と観測値の密度の比較
```



多重代入を用いた回帰分析

modelA<-lm.mids(data...1.~X1+X2+X3,data=a.mids)

summary(pool(modelA))

pool.r.squared(modelA)　#決定係数の統合

```
> summary(pool(modelA))
               estimate   std.error statistic       df      p.value
(Intercept) 5.49678013 0.117310404 46.856715 436.5758 0.000000e+00
X1          0.07415588 0.006880244 10.778089 471.5083 0.000000e+00
X2          0.01541846 0.003598165  4.285089 444.1019 2.215514e-05
X3          0.01458151 0.002929793  4.976976 217.7199 9.069905e-07
> pool.r.squared(modelA)
          est     lo 95     hi 95 fmi
R^2 0.1606915 0.115273 0.2105089 NaN
```

参考文献

- 高橋・渡辺(2017)「欠測データ処理-Rによる単一代入法と多重代入法」共立出版
- 高橋・伊藤(2014)「様々な多重代入法アルゴリズムの比較～大規模経済系データを用いた分析～」,統計研究彙報71号