

# 多重代入法を用いた 欠測データ処理

経済学部4年 原田悠介



# 内容



# 目次

1. 欠測データの問題点
2. 欠測データの発生メカニズム
3. 単一代入法
4. 多重代入法の概要と分析手順
5. シミュレーション

# 欠測データの問題点

- ・ 計算処理が不可能になる

(ほとんどの統計ソフトはリストワイズ除去で対応)

リストワイズ除去で対応すると

- ・ データ資源が無駄になる
  - ・ 分析結果に偏りがでる
- などの問題

欠測データの例

	エンジンの消費量 (本/月)	家庭の構成人数(人)
佐藤	10	3
鈴木	無回答(本当は6)	2
田中	2	1
高橋	20	無回答(本当は5)



リストワイズ除去

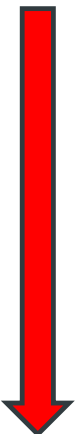

	エンジンの消費量 (本/月)	家族の人数(人)
佐藤	10	3
田中	2	1

# 欠測データの発生メカニズム

欠測はどのようなメカニズムで発生しているかによって無視可能と無視不可能に分かれる

$D$  ... データセット ( $n \times p$  行列)、 $D_{obs}$  ...  $D$  の観測部分、 $D_{mis}$  ...  $D$  の欠測部分

$K$  ...  $D$  と同次元の回答指示行列 (回答ありなら1, なしなら0になる要素で構成される)

現実的	扱いやすさ	発生メカニズム	無視可能 (代入法による 対処が可能)
		MCAR(Missing Completely At Random) $\Pr(K D) = \Pr(K)$	○
		MAR(Missing At Random) $\Pr(K D) = \Pr(K D_{obs})$	○
		MNAR(Missing Not At Random) $\Pr(K D) \neq \Pr(K D_{obs})$	×

# 単一代入法

## 単一代入法とは

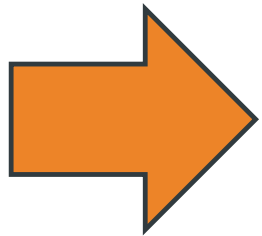
代入モデルから得られる唯一の代入値で欠測値を置き換える手法  
平均値代入法、ホットデック法、回帰代入法などが用いられる。

- 平均値代入法...観測部分の平均値を欠測値に代入
- ホットデック法...観測部分の値が似ているサンプルの回答を欠測値に代入  
観測部分の”似ている”を判断するために距離関数を定義
- 回帰代入法...観測部分で興味のある変数を説明する回帰分析を行い、  
回帰モデルから算出した値を欠測値に代入

# 単一代入法

## 単一代入法の問題点

- 使用する代入モデルによって代入値が変化する
- 代入された値を唯一絶対のものとして分析をしてしまう  
(1つの代入値の背後にある様々な可能性を無視している)



不確実性を反映させるために、複数の値を算出する必要がある。  
多重代入法へ

# 多重代入法の概要と分析手順

## 多重代入法

欠測データの分布から独立かつ無作為に抽出した  
 $M$ 個( $M > 1$ )のシミュレーション値によって欠測を置き換える方法

欠測データの分布は観測不可能なので...

MAR(MCAR)を仮定し、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た後抽出を行う



# 多重代入法の概要と分析手順

## ①代入

欠測データの事後予測分布を構築することで、  
M個の代入済みデータセットを生成する。

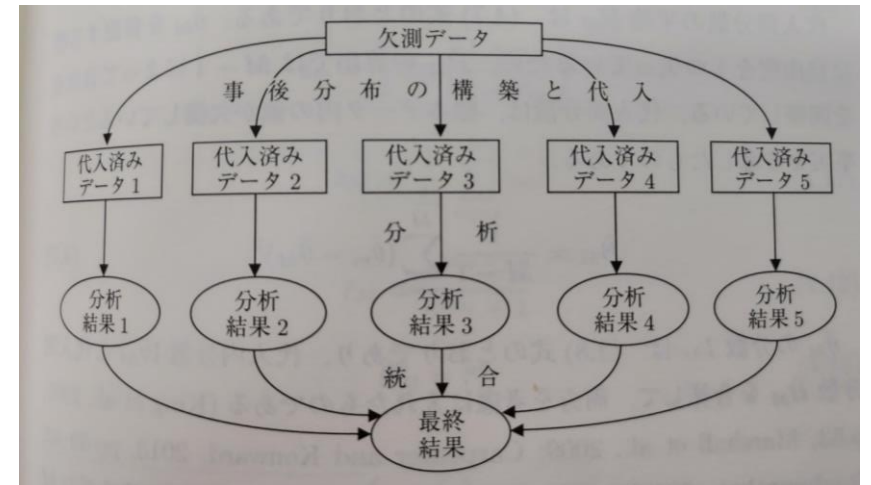
多重代入法の分析手順の模式図(M=5)

## ②分析

それぞれのデータセットで分析を行う。

## ③結果の統合

結果の分布が正規なら平均値をとる。  
正規でないならZ変換を行って平均値を取る



# シミュレーション

## シミュレーションデータ

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

ただし

- $X_k$  は平均値ベクトル0、下の相関行列の多変量正規分布に従う

$$\begin{bmatrix} 1 & 0.18 & -0.32 \\ 0.18 & 1 & 0.01 \\ -0.32 & 0.01 & 1 \end{bmatrix}$$

- $\beta_k \sim U(-2, 2)$  ( $k = 0, \dots, 3$ )
- $\varepsilon \sim N(0, U(0.5, 2))$
- $Y$  が中央値以上なら  $X_k$  は0.5の確率で欠測
- $Y$  が中央値未満なら  $X_k$  は0.1の確率で欠測

# シミュレーション

## ①EMBアルゴリズムによる多重代入法

### EMBアルゴリズム

EM(期待値最大化;Expectation-Maximization)法とノンパラメトリック・ブートストラップから構成される多重代入法のアルゴリズム

# シミュレーション

## ②FCSアルゴリズムによる多重代入法

### 完全条件付指定(fully conditional specification)

多変量分布を一連の条件付き分布によって指定し、他の変数を条件として欠測値の代入を行う