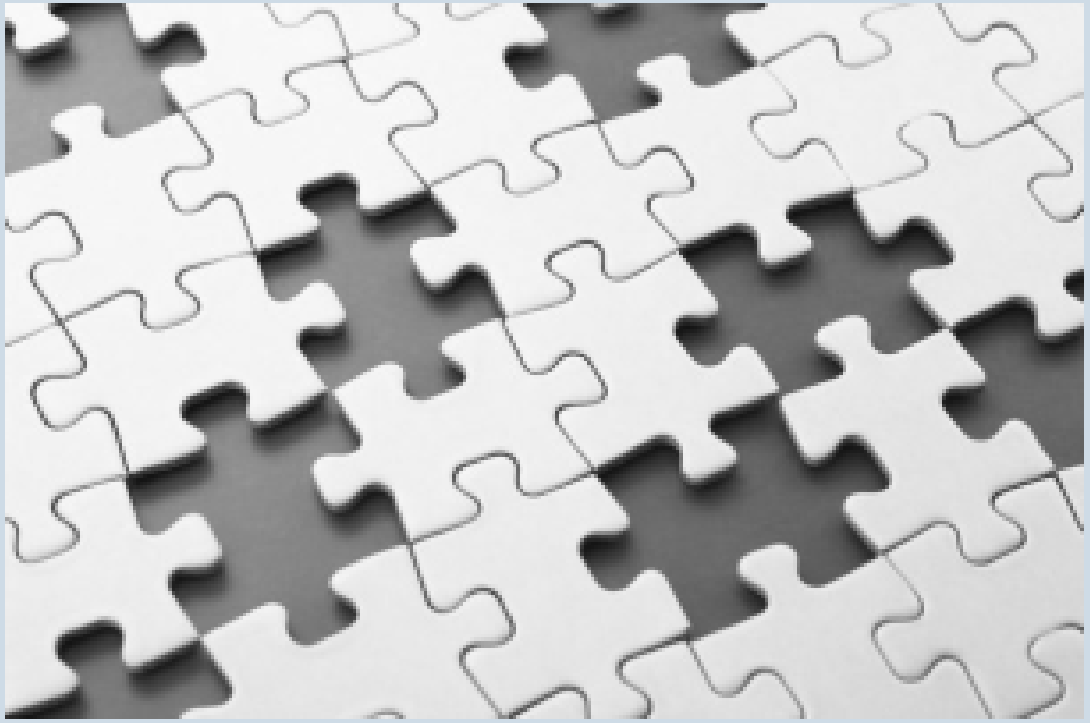
欠損値を補完する？

教育認知心理学講座 野村研究室 M1
高野 了太

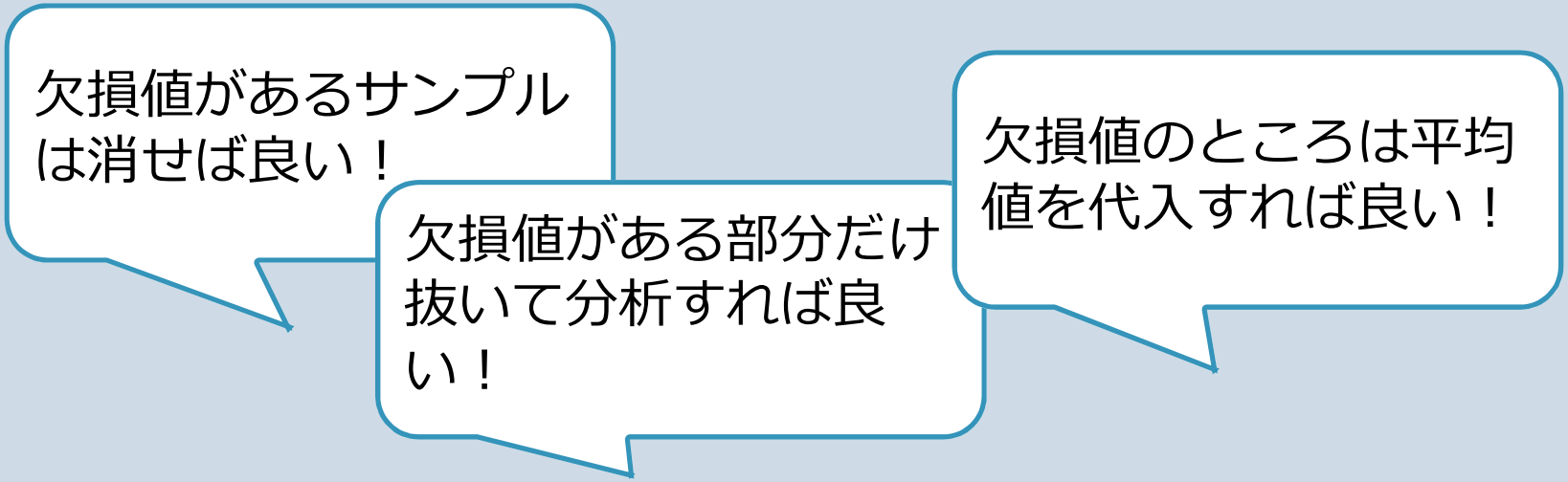
データ解析演習 2017/07/05

目次



1. はじめに

- 心理学の研究、とりわけ質問紙調査などでは、欠損値はつきもの。
- 欠損値があるデータをどのように扱うのかに関しては、様々な議論がなされてきた（Enders, 2010）。



- ……本当にそれで良いの？？という話

2. 欠損値の種類

- 欠損値は、「どのように欠損が生じたか」によって、3つに大きく分類
- **①MCAR (Missing Completely At Random)**
  ー完全ランダムに欠損が生じた場合
- **②MAR (Missing At Random)**
  ー分析に含まれる変数（X）とは関係するが、欠損データとそれを含む変数（Y）に対しては無関係な場合
- **③MNAR (Missing Not At Random)**
  ー欠損値の有無が欠損値を持つ変数自身と関係を持つ場合

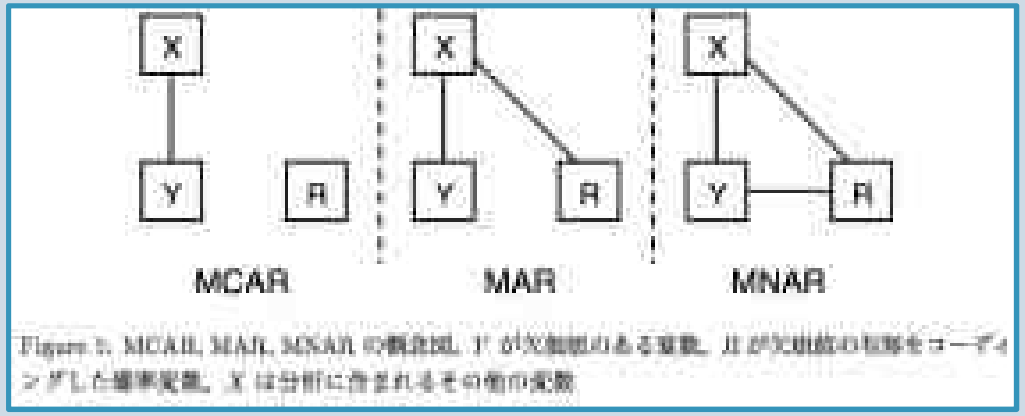
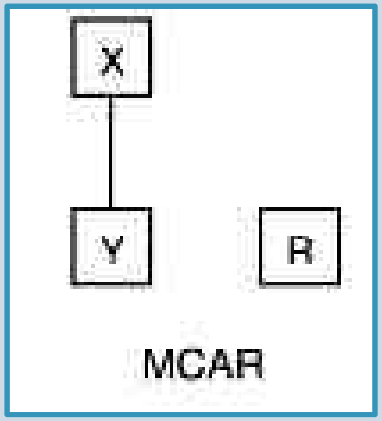


Figure 1. MCAR, MAR, MNAR の概念図. Y が欠損値のある変数, R が欠損の有無をコーディングした確率変数, X は分析に含まれるその他の変数

http://koumurayama.com/koujapanese/missing_data.pdf

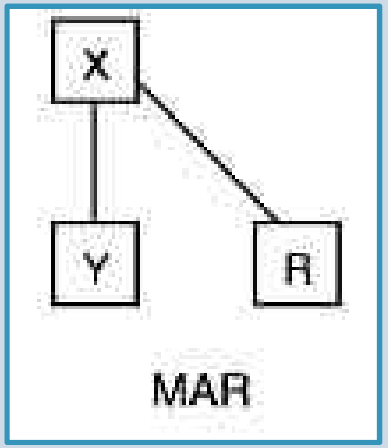
2-1. MCAR

- **MCAR (Missing Completely At Random)**
- 欠損が完全にランダムに生じている場合
- 欠損データを含む変数はもちろん、他の変数とも関連がない。
- 右図では...？
  ーYが欠損データのある変数
  ーRは欠損した時は0、観測した時は1をとる確率変数
  ーXは分析に含まれるその他の変数
- RがX、Yどちらとも関連していないことがわかる。



2-2. MAR

- **MAR (Missing At Random)**
- Yにおける欠損値の有無（R）が、他の変数Xと関連しているが、Xを統制するとその変数Y自体の値とは無関係である場合
- 他の変数との関連はOK。



IQが低い人に後の適性検査を実施しなかった時
・欠損の有無(R)がIQ(X)と関連
→MCARではない
・欠損なしの適性検査(Y)が欠損の有無(R)と関連してそうだが、それはIQによる偏相関
→統制すれば消えるMAR

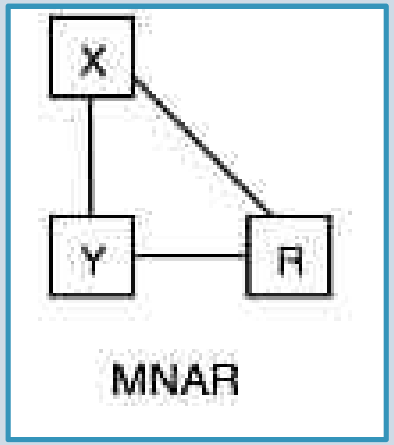
Table 1: MAR データの例

id	動機づけ	IQ	適性検査	適性検査（欠損なしの場合）
1	3	83	n/a	93
2	4	85	n/a	99
3	5	95	n/a	98
4	2	96	n/a	103
5	5	103	128	128
6	3	104	102	102
7	2	109	111	111
8	6	112	113	113
9	3	115	117	117
10	3	116	133	133
平均値	3.6	101.8	117.3	111.7

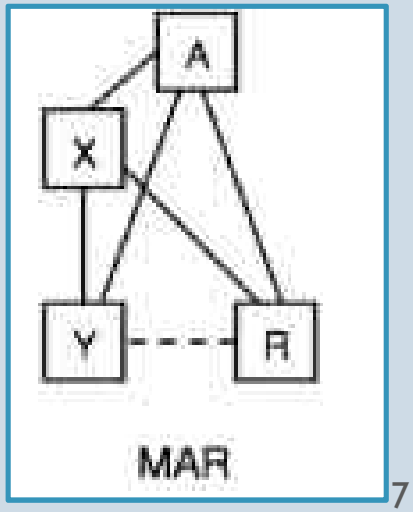
http://koumurayama.com/koujapanese/missing_data.pdf

2-3. MNAR

- **MNAR (Missing Not At Random)**
- 分析に含まれる他の変数を統制した後でも、欠損値の有無（R）が欠損値を持つ変数自体（Y）と関係を持つ場合



- ただし、他の変数を組み込むことで、それが分析には直接関係ない変数であっても、MARにすることが可能。
  - ーInclusive Analysis Strategy (Enders, 2010; Rubbin, 1996)
  - ー補助変数（Auxiliary Variable: 右図A）
- 補助変数はいくつあっても良いし、とりあえず全部投入すれば良い。
  - ーシミュレーション研究で実証（Enders, 2008）
- 具体的には、**多重代入法**や**完全情報最尤推定法**など



3. 欠損データの対処法

- 除去する手法
  ①**ペアワイズ削除法**
  　ー2変数の組み合わせで少なくとも1つが欠損していれば削除。
  ②**リストワイズ削除法**
  　ー1つでも欠損値があればオブジェクトを削除。
- 伝統的な処理法ではあるが、MCARを前提としている。
- 推定値にバイアスがかかり、仮にMCARであったとしても検定力が低下することが分かっているため、現在は他の手法を用いられることが多い。

- 推定する手法
  ①**完全情報最尤推定法（Full Information Maximum Likelihood method）**
  ②**多重代入法（Multiple Imputation Method）**

3. 欠損データの対処法

削除はバイアスが生じる

欠測処理		内容	推定したパラメータにバイアスが生じるどうか		
			MCAR	MAR	NMAR
削除	リストワイズ	欠測データを行単位で削除	○	×	×
	ペアワイズ	分析に用いた変数の範囲で欠測データを行単位で削除	○	×	×
最尤法		欠測を考慮した形で最尤法を適用する（EMアルゴリズム等適用）	○	○	△
代入	単一代入法	平均値や、他の変数による予測値を代入	○	×	×
	多重代入法		○	○	×

(Ichikawa, D., 2015)

欠損の割合：
10%未満→リストワイズでもOK？
10%以上→FIML or MI

https://www.slideshare.net/hajimesasaki1/wi2-55598897

3-1. FIML法

- **完全情報最尤推定法（Full Information Maximum Likelihood method）**
  ーケースごとに、欠損パタンに応じた個別の尤度関数を仮定した最尤推定法。
  ー普通の最尤推定法と変わらないが、FIMLと呼ばれることが多い。
- 個人ごとにデータのサイズが違っていても、個人ごと・全体の尤度を求めることができる。
- 欠損値があったとしても、同様の原理で最尤推定値を求めることができる。
- つまり、他の変数の情報を借りるような形で、欠損値のある変数のパラメタを推定することが可能。

- 詳しくは村山先生の資料 (http://koumurayama.com/koujapanese/missing_data.pdf) 及びEnders (2010) を参照。
- AMOSやMplus、SASでも実施可能。

3-2. 多重代入法

- 代入法（Imputation Method）
  ー平均値代入法
  　：欠損値以外の平均値を代入する。
  ー回帰代入法
  　：回帰モデルの予測値を代入する。
- これらは、測定に伴う不確実性を反映していないため、分散などが過小推定されてしまう。
- この問題の対処法として、**Stochastic Regression Imputation**
  ー回帰モデルの予測値にランダム誤差（誤差分散）を加える。
- この手法はある程度Goodだが、欠損値があることによる推定の不確定性が考慮されていないため、欠損値が多い場合に標準誤差を過小評価してしまう。

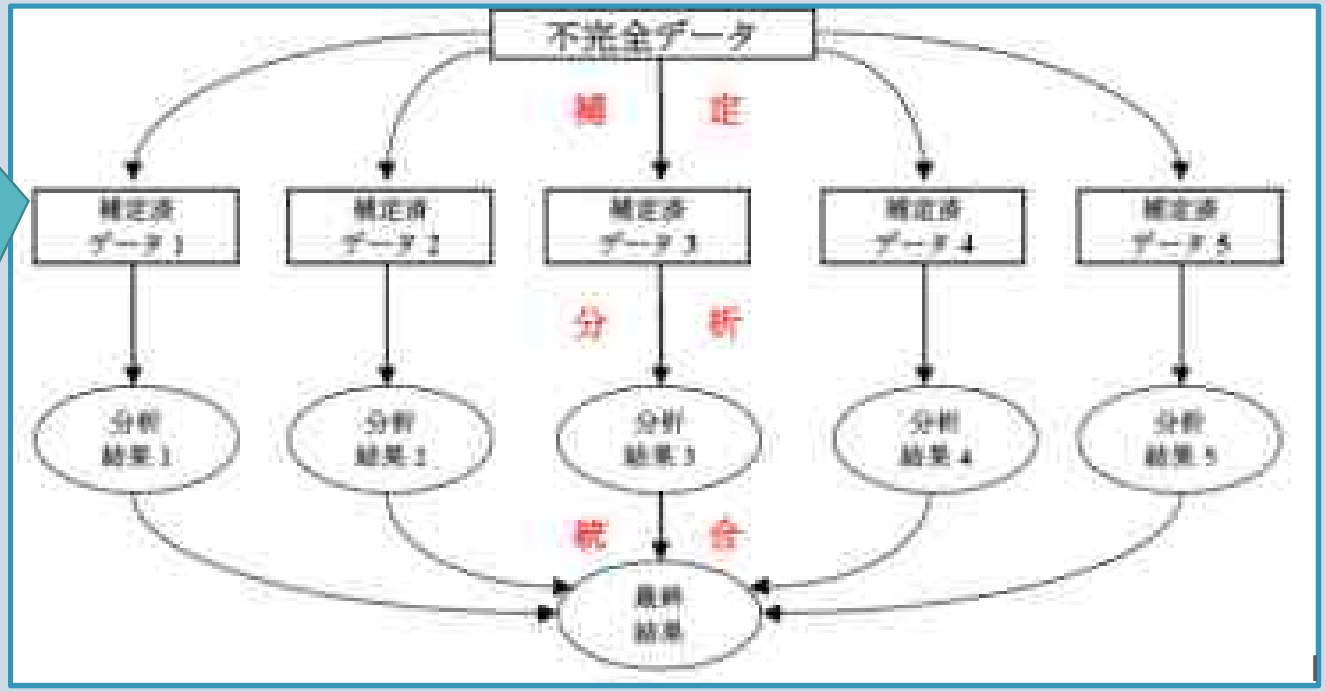
3-2. 多重代入法

- **多重代入法（Multiple Imputation Method）**では、欠損値を代入したデータセットを複数作成し、その結果を統合することで欠損値データの統計的推測を行う（Rubin, 1987）。
- データセットを複数作成することで、欠損値による推定の不安定性を結果に反映させている。

①代入ステップ
：擬似完全データセットを複数作成する。

②分析ステップ
：推定値とSEを得る

③統合ステップ
：複数の推定値とSEを統合して単一のそれらを算出。



3-2. 多重代入法

- ①**代入ステップ**（Imputation Step）
  ーデータ拡大法が主流、基本的にベイズ統計学の考えに大きく依拠。
  ー事後予測分布から乱数を発生させ、それを欠損値に代入したデータセットを複数作る。
  ー乱数の発生には**マルコフ連鎖モンテカルロ法**（Markov chain monte carlo; **MCMC**）
- マルコフ連鎖モンテカルロ法
  ーデータxが与えられた時、事後分布P(θ | x)からパラメータθをサンプリングする手法。

3-2. 多重代入法

- ③**統合ステップ**（Posterior / Integration Step）
  ー複数の擬似完全データセットが得られたら、それぞれのデータセットに関して、目的の分析（回帰分析、ANOVA、SEMなど）を実施する。
  ーパラメタの推定値とSEを統合する。

- パラメタ推定値の統合
  ー平均する。
- SEの統合

$$V_W = \frac{1}{m}\sum_{t=1}^{m} SE_t^2$$

$$V_B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}_t - \bar{\theta})$$

$$V_T = V_W + V_B + \frac{V_B}{m}$$

$$SE = \sqrt{V_T}$$

擬似データセット間のばらつきの指標

3-2. 多重代入法

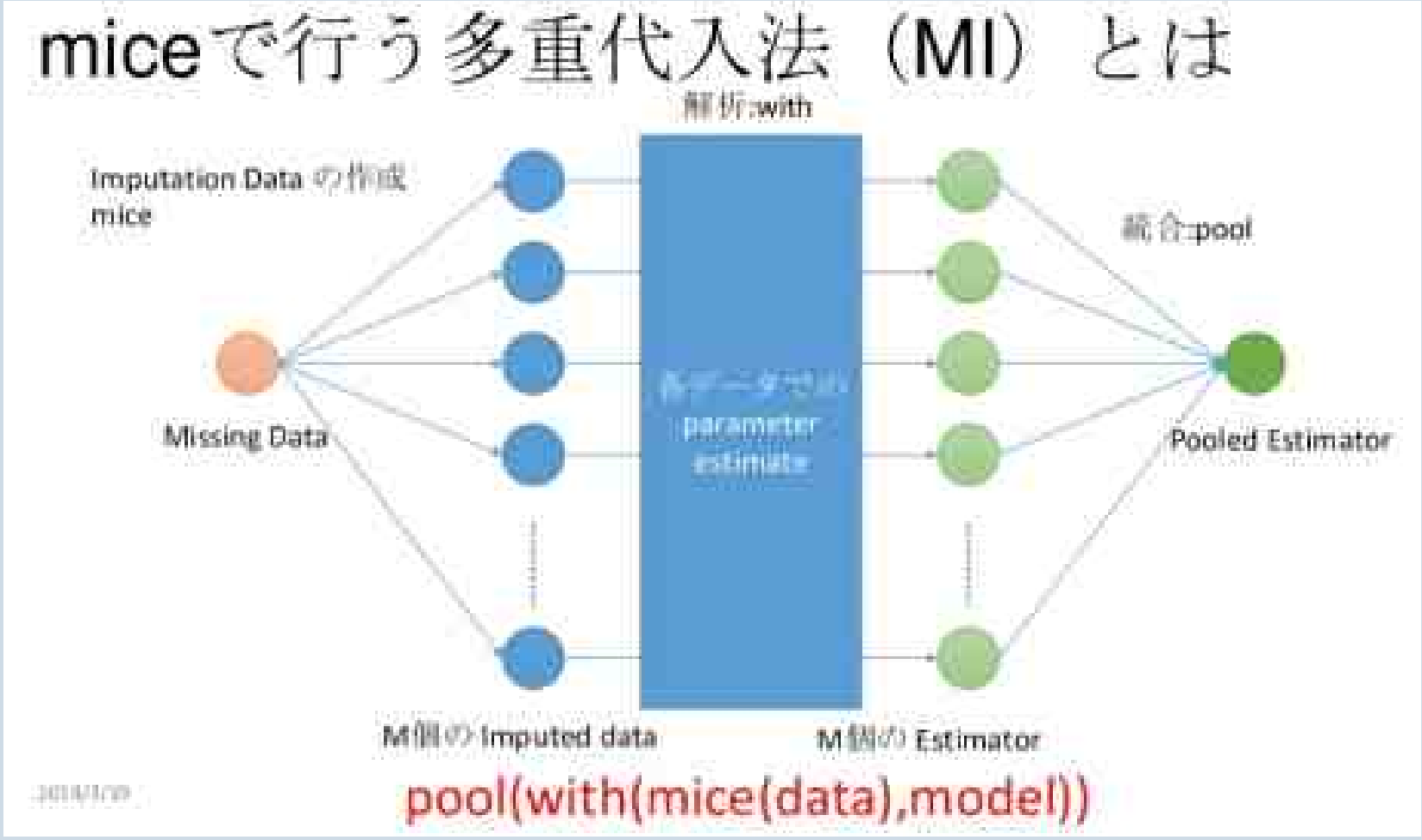
- **多重代入法の留意事項**
- 擬似完全データセットの数
  - Rubin (1987)は3~5、miceパッケージのデフォルトは5
  - Enders (2010)は20くらいを目安としている。
- 交互作用に興味がある時
  - 代入ステップで交互作用項もモデルに含めておく。
- 階層的なデータの分析を用いる場合
  - 階層性を代入時に仮定した方がBetterだが、それができるソフトウェアは少ない（NormとMplus ver.6では可能）。
- 尺度レベル？項目レベル？
  - 検出力の関係から、項目レベルでやった方が良いが、項目数が多いと結果が収束しなかったり、（回帰分析だと）そもそも代入できないことも。
  - Enders (2010)は両方使った代入も勧めている。

4. 実際に多重代入法をやろう！

- 使用するのはRのMICEパッケージ
- 他にも、SASやSPSSのパッケージで、代入ステップと統合ステップを自動的に行うことができる（SPSSはsequential regression modelを使用）。
- Rでは、MICEの他にもAmelia、Normなどのパッケージがある。

- **R パッケージ MICE**
  - ーオランダのユトレヒト大学のStef van Buuren (2012)を中心としてチームにより開発された多重代入法プログラム。
  - ーmice()関数で代入を行い、with()関数で分析、pool()関数で統合結果を見る。
  - ー生成した補定済データセットはcomplete()関数で作成できる。
  - ーデータ例はmiceパッケージにあるデータを使用。

4. 実際に多重代入法をやろう！



http://sssslide.com/www.slideshare.net/Hiro_macchan/tokyo-r-37-hiromacchan

4. 実際に多重代入法をやろう！

- **データ概要**

  age: Age group (1=20-39, 2=40-59, 3=60+)
  bmi: Body mass index (kg/m**2)
  hyp: Hypertensive (1=no,2=yes)
  chl: Total serum cholesterol (mg/dL)

- 重回帰分析 chl<-age, bmi
- 年齢はすべて分かっているが、そのほかに欠損がいくつかある。
- **分析の流れ**

  ①欠損パターンを概観する。
  ②データを補完する。
  ③補完データを分析して統合する。
  ④補完後のデータを確認する。

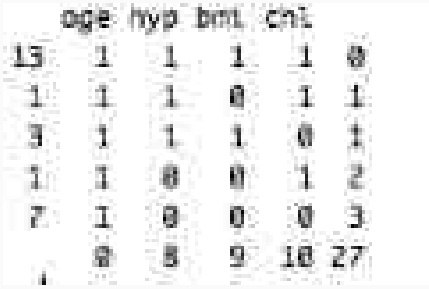
4. 実際に多重代入法をやろう！

- **①分析パターンを概観する**

```
data(nhanes)

# miceをinstall
install.packages("mice")
library(mice)
md.pattern(nhanes)

install.packages("VIM")
library(VIM)
aggr(nhanes, prop = FALSE, number = TRUE)
```



	age	hyp	bmi	chl	
13	1	1	1	1	0
1	1	1	0	1	1
3	1	1	1	0	1
1	1	0	0	1	2
7	1	0	0	0	3
	0	8	9	10	27

4. 実際に多重代入法をやろう！

- **②データを補完する**

```
tempData <- mice(nhanes,
                 m=10,
                 # refers to the number of imputed datasets. Five is the default value.
                 maxit=50,
                 meth='pmm',
                 # refers to the imputation method, pmm: predictive mean matching
                 seed=500)
summary(tempData)
```

4. 実際に多重代入法をやろう！

- **③補完データを分析して統合する**

```
fit <- with(data=tempData, lm (chl ~age +bmi) )
summary (pool (fit))
```

```
                   est         se          t         df
(Intercept) -11.306053 72.442023 -0.1560704 10.543807
age          35.583698 15.539323  2.2899130  4.978333
bmi           5.340508  2.149577  2.4844463 13.670432
                Pr(>|t|)       lo 95      hi 95 nmis
(Intercept) 0.87892545 -171.5955931 148.983486   NA
age         0.07088574   -4.4137514  75.581148    0
bmi         0.02659780    0.7196751   9.961341    9
                  fmi    lambda
(Intercept) 0.4679667 0.3757903
age         0.7543002 0.6721035
bmi         0.3419975 0.2522929
```

4. 実際に多重代入法をやろう！

- **④補完後のデータを確認する**

```
library(lattice)
xyplot(tempData, chl ~ age+bmi ,pch=18,cex=1)

densityplot(tempData)

stripplot(tempData, pch = 20, cex = 1.2)
```

参考資料/文献・URL

- 村山先生のpdf(スライド4~14)
  http://koumurayama.com/koujapanese/missing_data.pdf
- 清水先生のサイト(スライド4~14)
  http://norimune.net/1811
- miceパッケージの使い方のサイト(スライド15~21)
  https://datascienceplus.com/imputing-missing-data-with-r-mice-package/
- 広大・徳岡さんのスライド(スライド4~14)
  https://www.slideshare.net/masarutokuoka/darm3-missing-data-analysis
- 外科医の方のサイト(スライド15~21)
  http://pediatricsurgery.hatenadiary.jp/entry/2016/11/20/233217
- 高橋さん・伊藤さんの資料(スライド15~21)
  http://www.stat.go.jp/training/2kenkyu/ihou/71/pdf/2-2-713.pdf
- Matsui Hirokiさんのスライド(スライド15~21)
  http://sssslide.com/www.slideshare.net/Hiro_macchan/tokyo-r-37-hiromacchan
- Hajime Sasakiさんのスライド(スライド9)
  https://www.slideshare.net/hajimesasaki1/wi2-55598897
- Enders, C.K. (2010). Applied missing data analysis. New York: Guilford.
- Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.