

前処理 欠損値処理系

鈴木瑞人

東京大学大学院 新領域創成科学研究科

メディカル情報生命専攻

博士課程1年

目次

統計量算出

データの可視化

欠損値生成メカニズム

欠損値対処法

各列の要約統計量算出

iris

```
> iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa

各列の要約統計量算出

summary(iris)

```
> summary(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300
Median :5.800	Median :3.000	Median :4.350	Median :1.300
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500

Species
setosa :50
versicolor:50
virginica :50

Min: 最小値

1st Qu: 第一四位置点(小さい方から25%のところに位置する値)

Median: 中央値(小さい方から50%のところに位置する値)

Mean: 平均値

3rd Qu: 第3四分位点(小さい方から75%のところに位置する値)

Max: 最大値

欠損値の把握

airquality

```
> airquality
```

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10
11	7	NA	6.9	74	5	11
12	16	256	9.7	69	5	12

欠損値の把握

summary(airquality)

```
> summary(airquality)
```

Ozone		Solar.R		Wind		Temp	
Min.	: 1.00	Min.	: 7.0	Min.	: 1.700	Min.	: 56.00
1st Qu.:	18.00	1st Qu.:	115.8	1st Qu.:	7.400	1st Qu.:	72.00
Median :	31.50	Median :	205.0	Median :	9.700	Median :	79.00
Mean :	42.13	Mean :	185.9	Mean :	9.958	Mean :	77.88
3rd Qu.:	63.25	3rd Qu.:	258.8	3rd Qu.:	11.500	3rd Qu.:	85.00
Max.	:168.00	Max.	:334.0	Max.	:20.700	Max.	:97.00
NA's	:37	NA's	:7				

Month		Day	
Min.	:5.000	Min.	: 1.0
1st Qu.:	6.000	1st Qu.:	8.0
Median :	7.000	Median :	16.0
Mean :	6.993	Mean :	15.8
3rd Qu.:	8.000	3rd Qu.:	23.0
Max.	:9.000	Max.	:31.0

Ozone列で37個、
Solar.R列で7個の欠損値
(NA(NotAvailable))があること
がわかる。

complete.cases() 関数

- complete.cases() 関数は行ごとにNAがあるならFALSE、ないならTRUEをかどうかのTRUE/FALSEをベクトルで返します。

complete.cases(airquality)

```
> complete.cases(airquality)
 [1]  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
[13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[25] FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
[37] FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE
[49]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[61] FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
[73]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
[85]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
[97] FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
[109]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
[121]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[133]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[145]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
```

欠損値がある行の集計

```
table(complete.cases(airquality))
```

```
> table(complete.cases(airquality))
```

```
FALSE  TRUE  
    42   111
```


na.omit関数による、欠損値がある行の消去

summary(na.omit(airquality))

```
> summary(na.omit(airquality))
```

Ozone	Solar.R	Wind	Temp
Min. : 1.0	Min. : 7.0	Min. : 2.30	Min. : 57.00
1st Qu.: 18.0	1st Qu.: 113.5	1st Qu.: 7.40	1st Qu.: 71.00
Median : 31.0	Median : 207.0	Median : 9.70	Median : 79.00
Mean : 42.1	Mean : 184.8	Mean : 9.94	Mean : 77.79
3rd Qu.: 62.0	3rd Qu.: 255.5	3rd Qu.: 11.50	3rd Qu.: 84.50
Max. : 168.0	Max. : 334.0	Max. : 20.70	Max. : 97.00

Month	Day
Min. : 5.000	Min. : 1.00
1st Qu.: 6.000	1st Qu.: 9.00
Median : 7.000	Median : 16.00
Mean : 7.216	Mean : 15.95
3rd Qu.: 9.000	3rd Qu.: 22.50
Max. : 9.000	Max. : 31.00

行数と列数の把握

dim(airquality)

```
> dim(airquality)
[1] 153  6
```

列数が多い場合の要約統計量の算出

- 実際のデータは、列数が数百を超えているものも多々あり、普通にsummary関数を使用すると、コンソール画面が埋め尽くされてしまう。
- 一部だけにsummary関数を適用する必要がある。
- そこで、lapply関数でsummary関数を使用する必要がある。
- この結果はリストとして保持され、ひとつずつ取り出すことができる。

ISLRパッケージのCaravanデータセットを使用する

```
install.packages("ISLR",quiet=T, dependencies=T)
```

```
library(ISLR)
```

```
data(Caravan)
```

```
str(Caravan)
```

```

> library(ISLR)
> data(Caravan)
> str(Caravan)
'data.frame':  5822 obs. of  86 variables:
 $ MOSTYPE   : num  33 37 37 9 40 23 39 33 33 11 ...
 $ MAANTHUI  : num  1 1 1 1 1 1 2 1 1 2 ...
 $ MGEMOMV   : num  3 2 2 3 4 2 3 2 2 3 ...
 $ MGEMLEEF  : num  2 2 2 3 2 1 2 3 4 3 ...
 $ MOSHOOFD  : num  8 8 8 3 10 5 9 8 8 3 ...
 $ MGODRK    : num  0 1 0 2 1 0 2 0 0 3 ...
 $ MGODPR    : num  5 4 4 3 4 5 2 7 1 5 ...
 $ MGODOV    : num  1 1 2 2 1 0 0 0 3 0 ...
 $ MGODGE    : num  3 4 4 4 4 5 5 2 6 2 ...
 $ MRELGE    : num  7 6 3 5 7 0 7 7 6 7 ...
 $ MRELSA    : num  0 2 2 2 1 6 2 2 0 0 ...
 $ MRELOV    : num  2 2 4 2 2 3 0 0 3 2 ...
 $ MFALLEEN  : num  1 0 4 2 2 3 0 0 3 2 ...
 $ MFGEKIND  : num  2 4 4 3 4 5 3 5 3 2 ...
 $ MFWEKIND  : num  6 5 2 4 4 2 6 4 3 6 ...
 $ MOPLHOOG  : num  1 0 0 3 5 0 0 0 0 0 ...
 $ MOPLMIDD  : num  2 5 5 4 4 5 4 3 1 4 ...
 $ MOPLLAAG  : num  7 4 4 2 0 4 5 6 8 5 ...
 $ MBERHOOG  : num  1 0 0 4 0 2 0 2 1 2 ...
 $ MBERZELF  : num  0 0 0 0 5 0 0 0 1 0 ...
 $ MBERBOER  : num  1 0 0 0 4 0 0 0 0 0 ...

```

86列のデータがある。

データ全体に対して、lapplyとsummaryを使用

```
sm.Caravan <- lapply(Caravan, summary)
```

要約統計量を知りたい列を列名で指定。

```
sm.Caravan[["Purchase"]]
```

```
sm.Caravan[["MOSTYPE"]]
```

```
> sm.Caravan[["Purchase"]]
```

```
  No  Yes
```

```
5474 348
```

```
> sm.Caravan[["MOSTYPE"]]
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	10.00	30.00	24.25	35.00	41.00

cor関数での相関係数の算出

#subset関数でirisデータのSpecies列以外を選択

```
cor(subset(iris, select = -Species))
```

```
> cor(subset(iris, select = -Species))
```

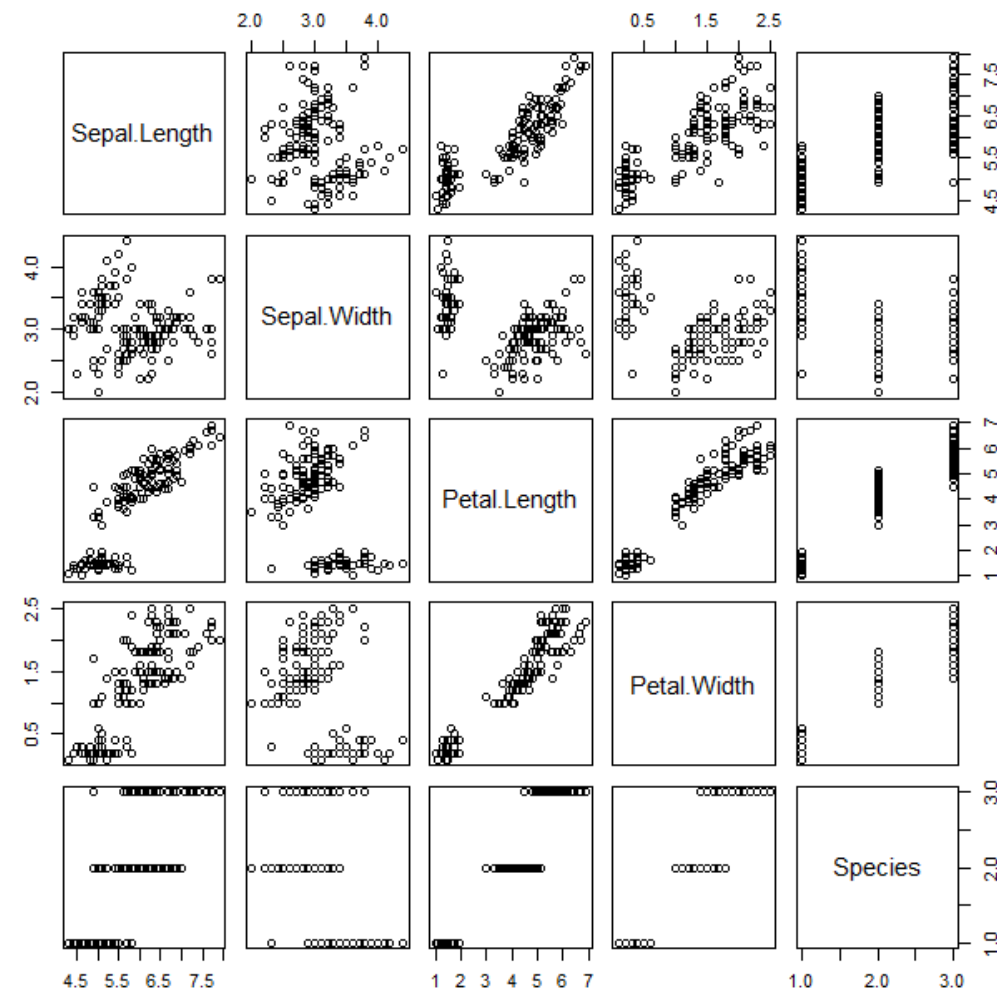
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

相関係数の意味の可視化

#plot関数かpairs関数を用いる

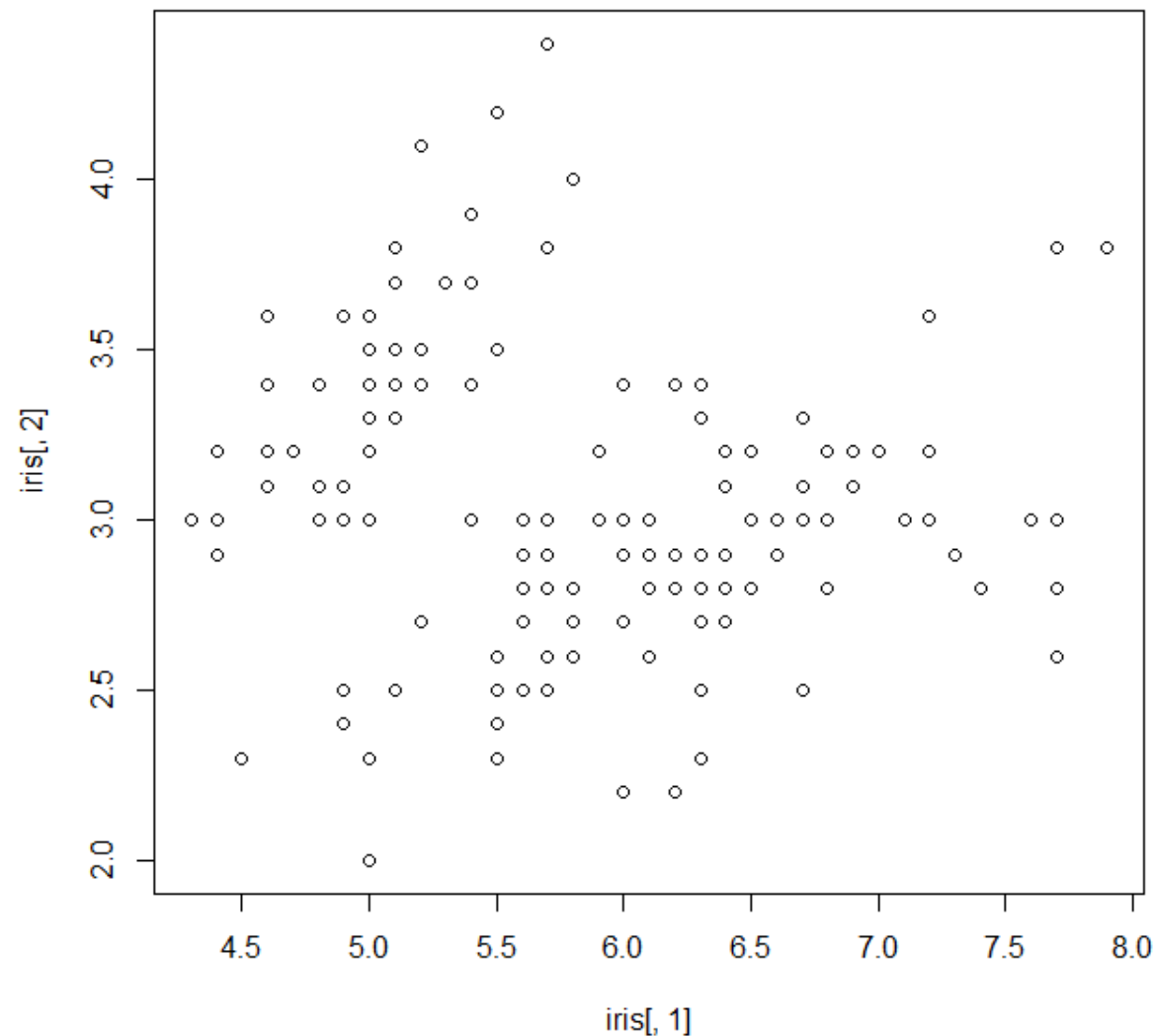
plot(iris)

#もしくはpairs(iris)



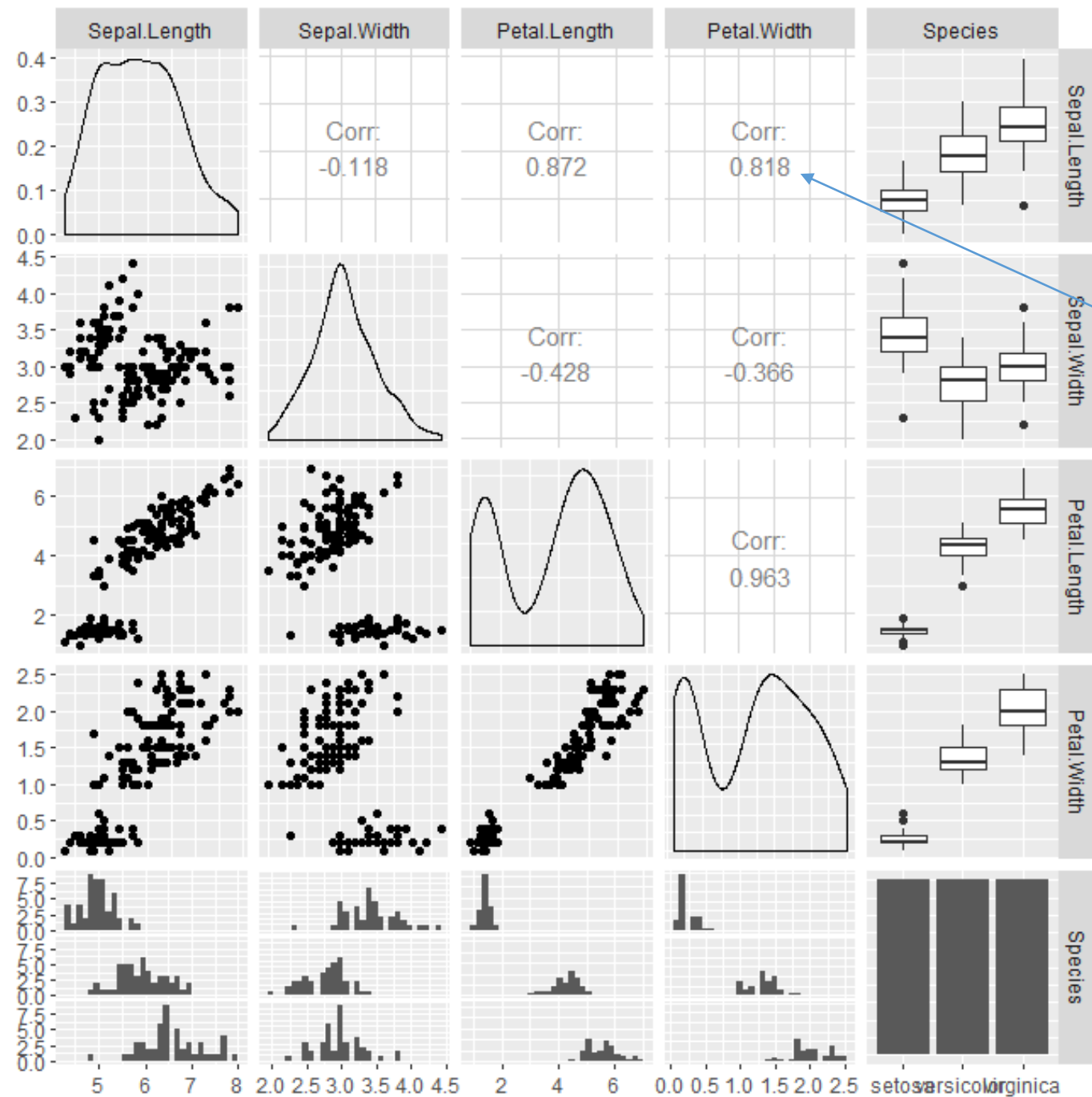
より詳細な2変数の関係の可視化

```
plot(iris[,1],iris[,2])
```



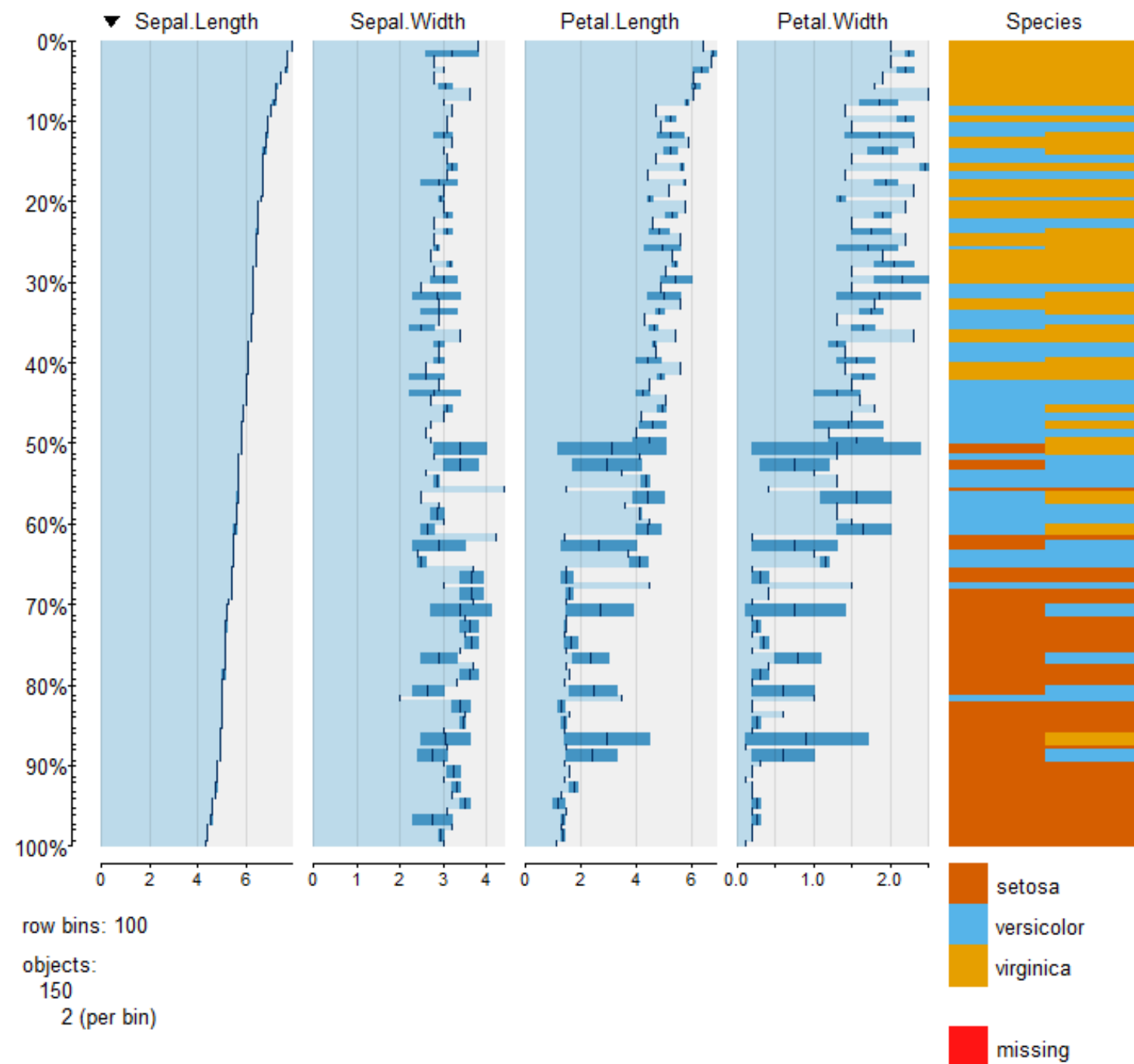
相関係数関係を一度に把握したい場合

```
install.packages("GGally", quiet = TRUE, dependencies=T)  
library(GGally)  
ggpairs(iris)
```

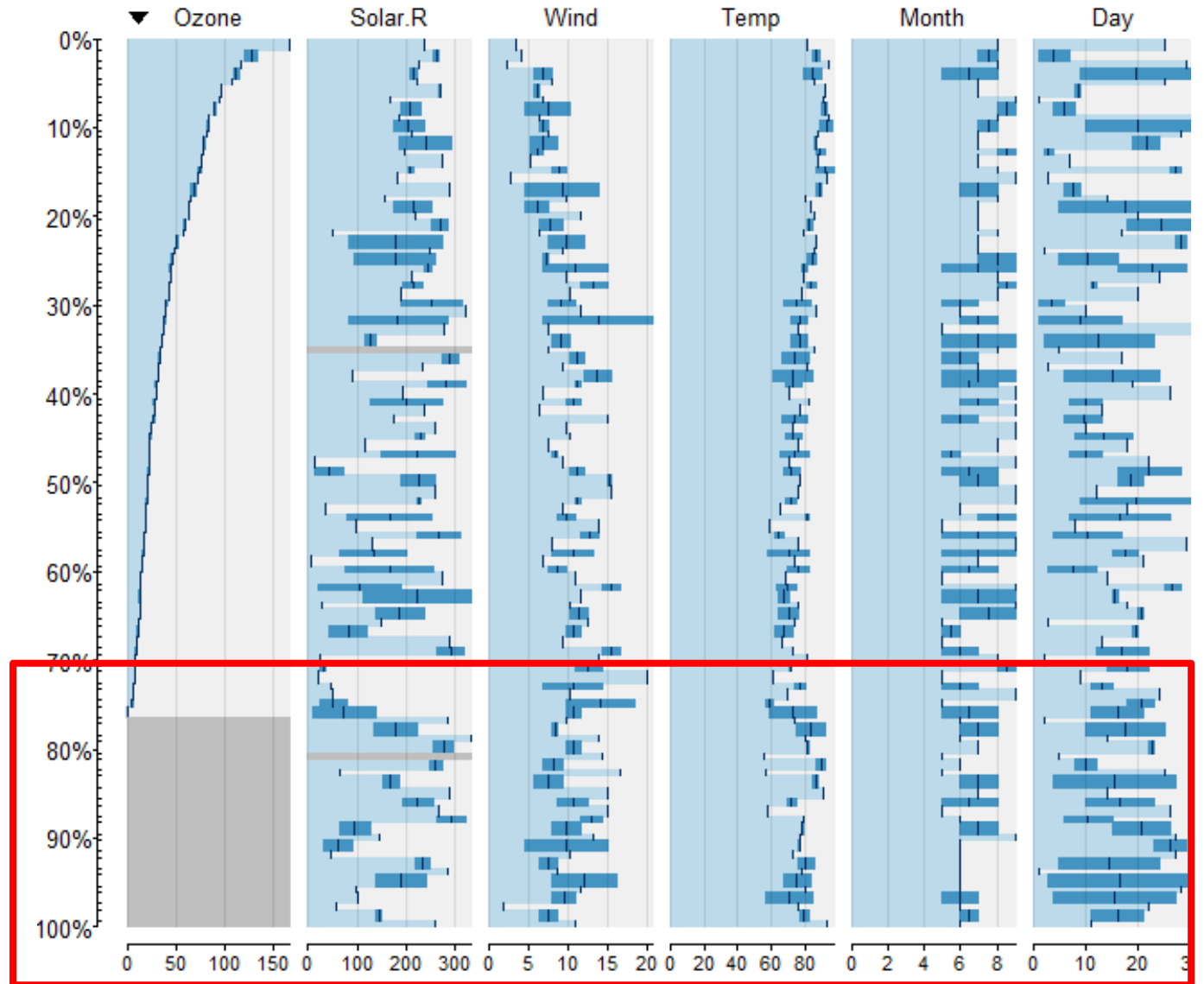


相関係数も算出

```
install.packages("tabplot", quiet = TRUE, dependencies=T)  
library(tabplot)  
tableplot(iris)
```



tableplot(airquality)



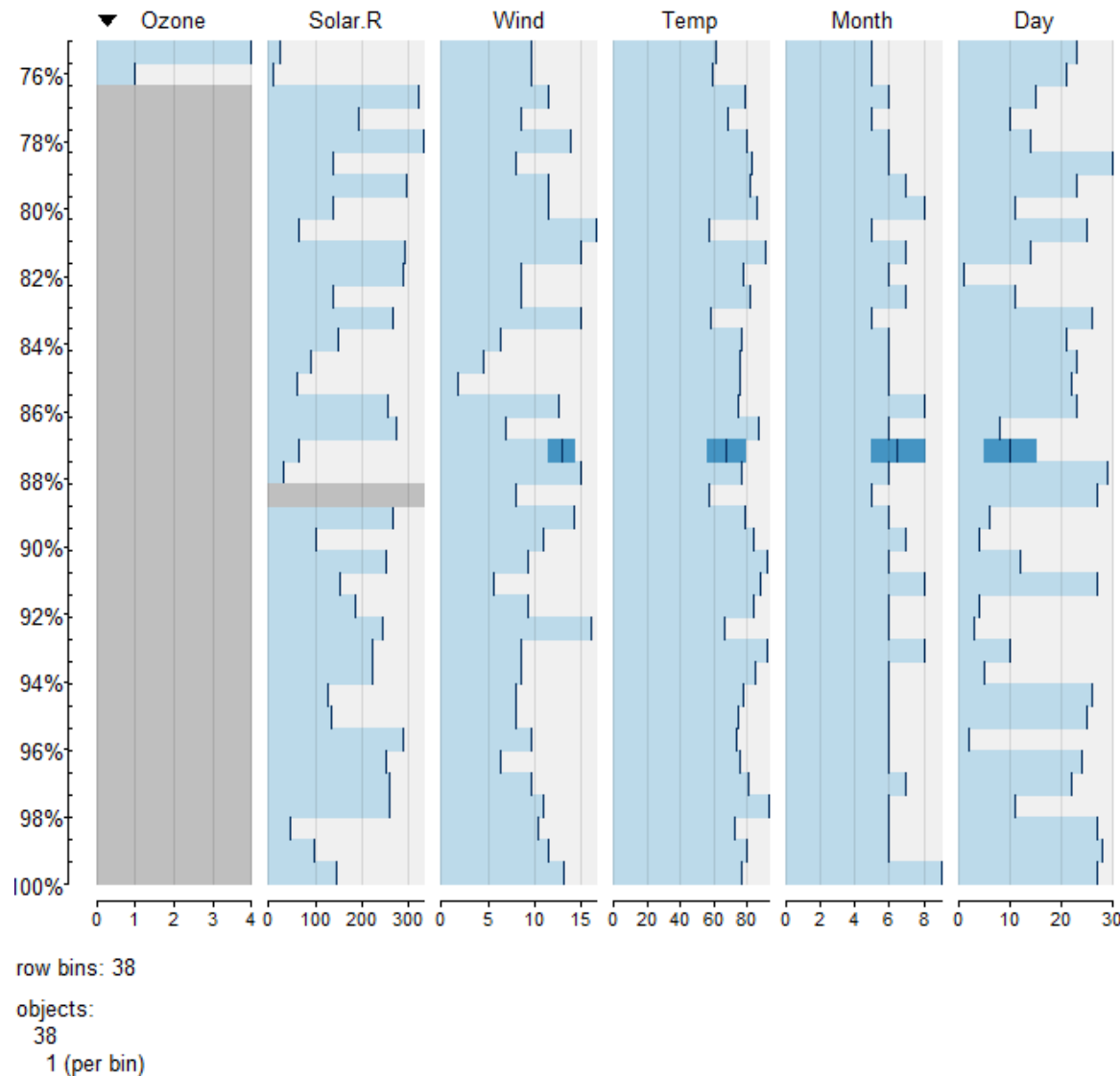
row bins: 100

objects:

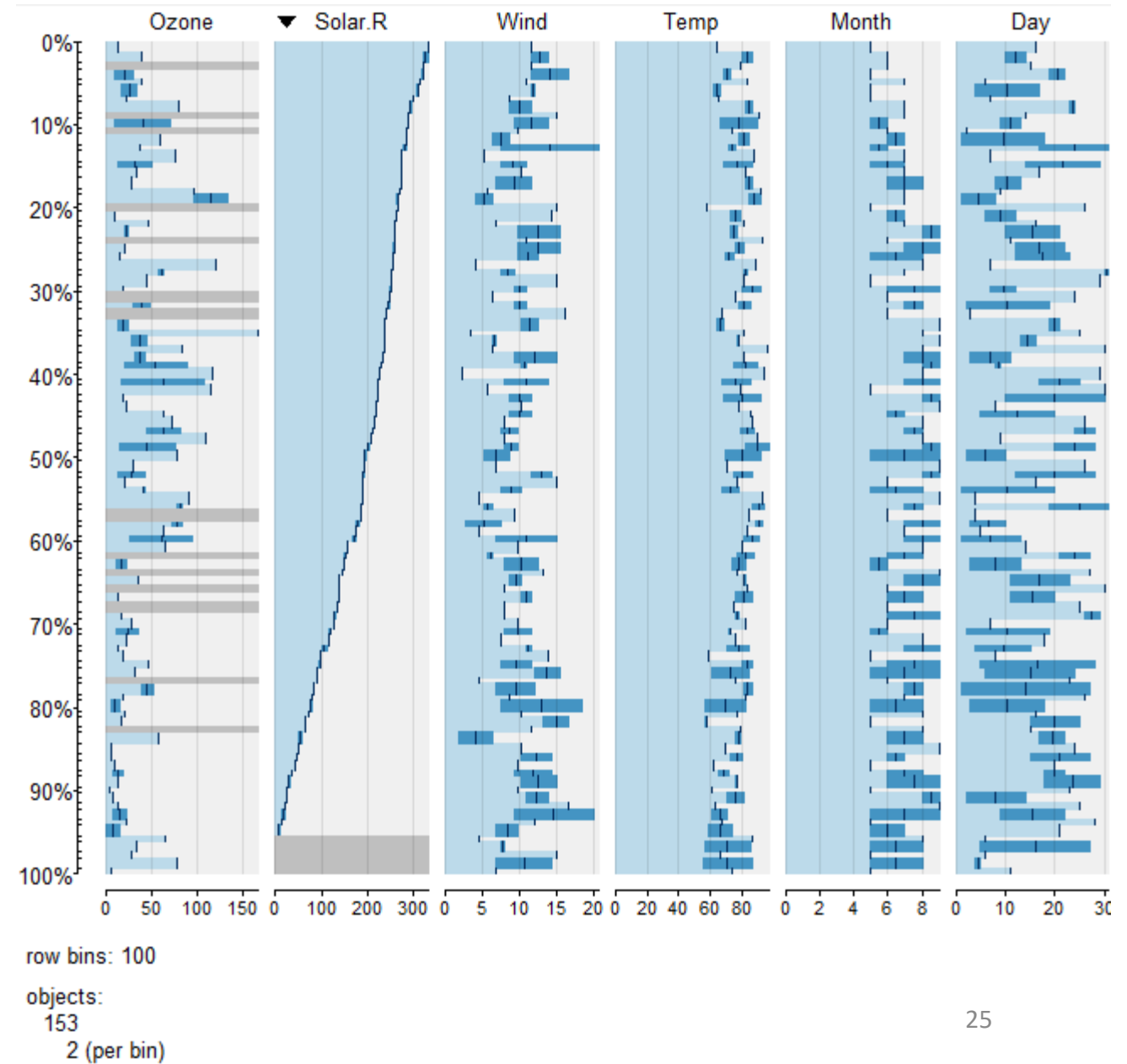
153

2 (per bin)

#表示するデータを構成割合で指定:from,toオプション
tableplot(airquality, from = 75, to = 100)




```
tableplot(airquality,sortCol="Solar.R")
```



分割を細分化する

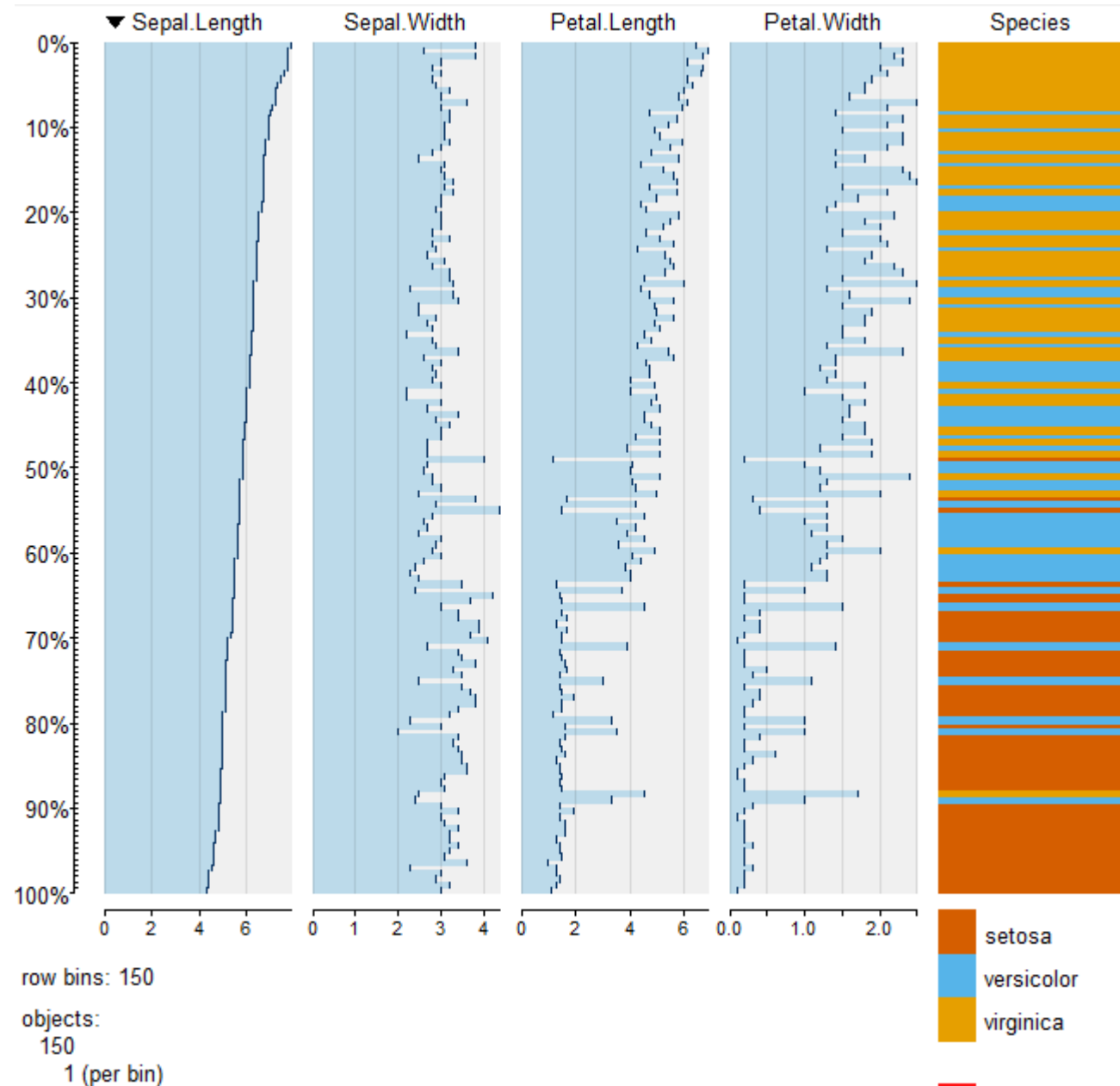
#デフォルトは100分割。

#もっと細かくすることもできる。

#分割して集計して、ソートされている。

#データ数(150)で分割

`tableplot(iris,nBins=150)`

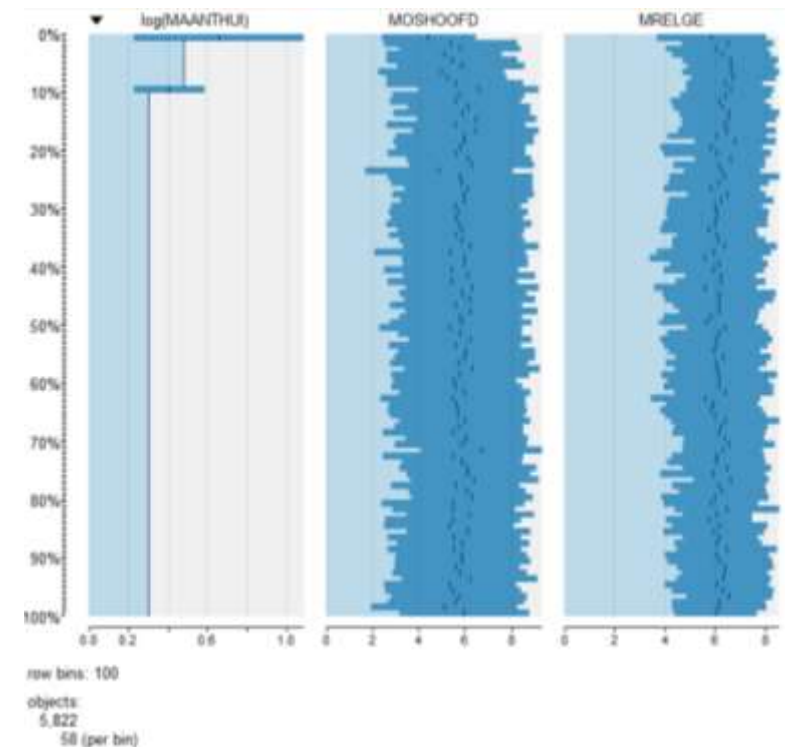


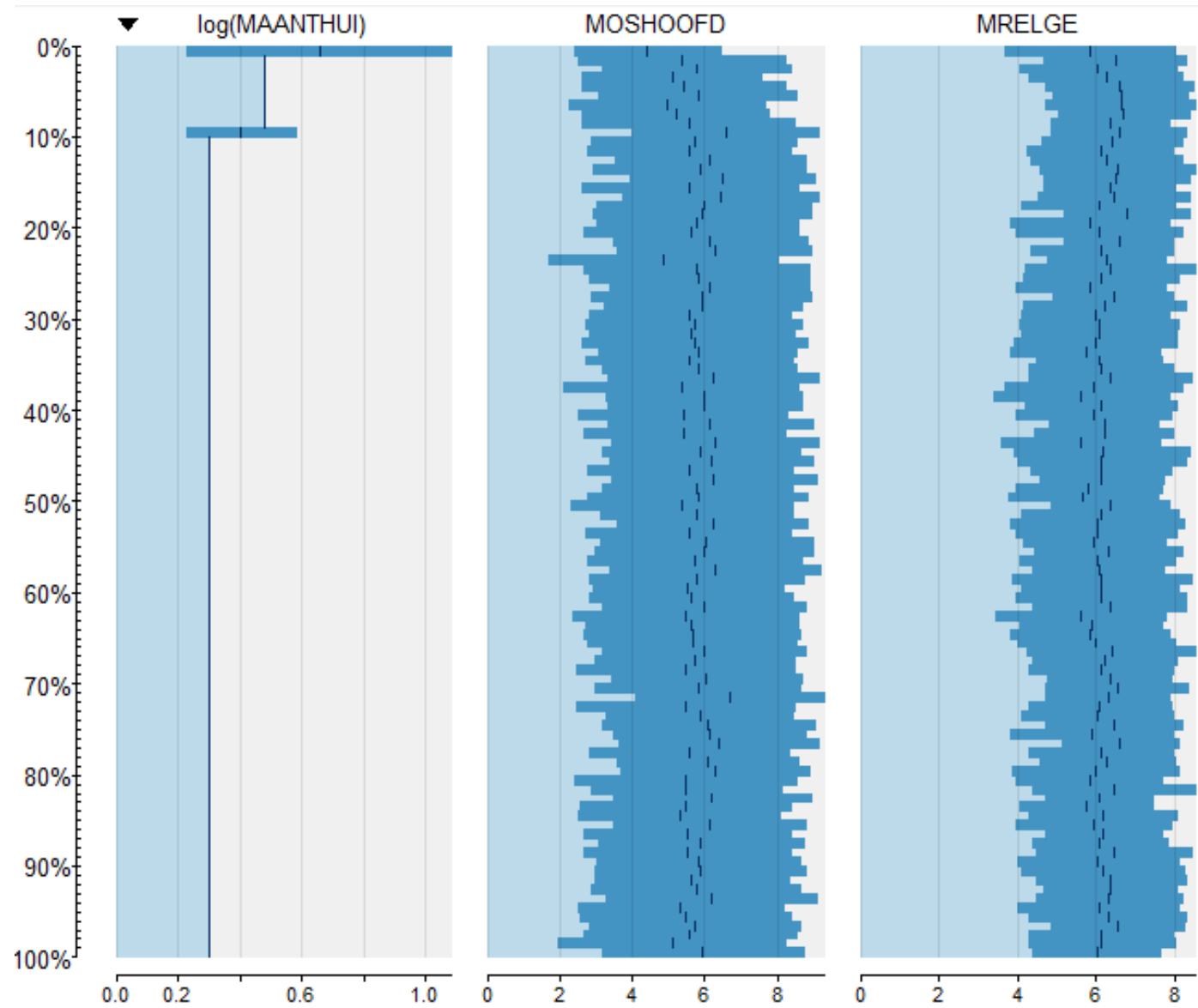
大きなデータを扱う際は列を選択して表示

```
library(ISLR)
```

```
data(Caravan)
```

```
tableplot(Caravan, select = c(MAANTHUI, MOSHOOFD, MRELGE))
```





row bins: 100

objects:

5,822

58 (per bin)

データ間の比較

#irisデータのSpecies列のvirginica列を抽出

```
Data1=iris[iris$Species=="virginica",]
```

#irisデータのSpecies列のversicolor列を抽出

```
Data2=iris[iris$Species=="versicolor",]
```

データ間の比較

#データをどのくらいの幅で集計するかをnBinsで指定(今回はデータ数)

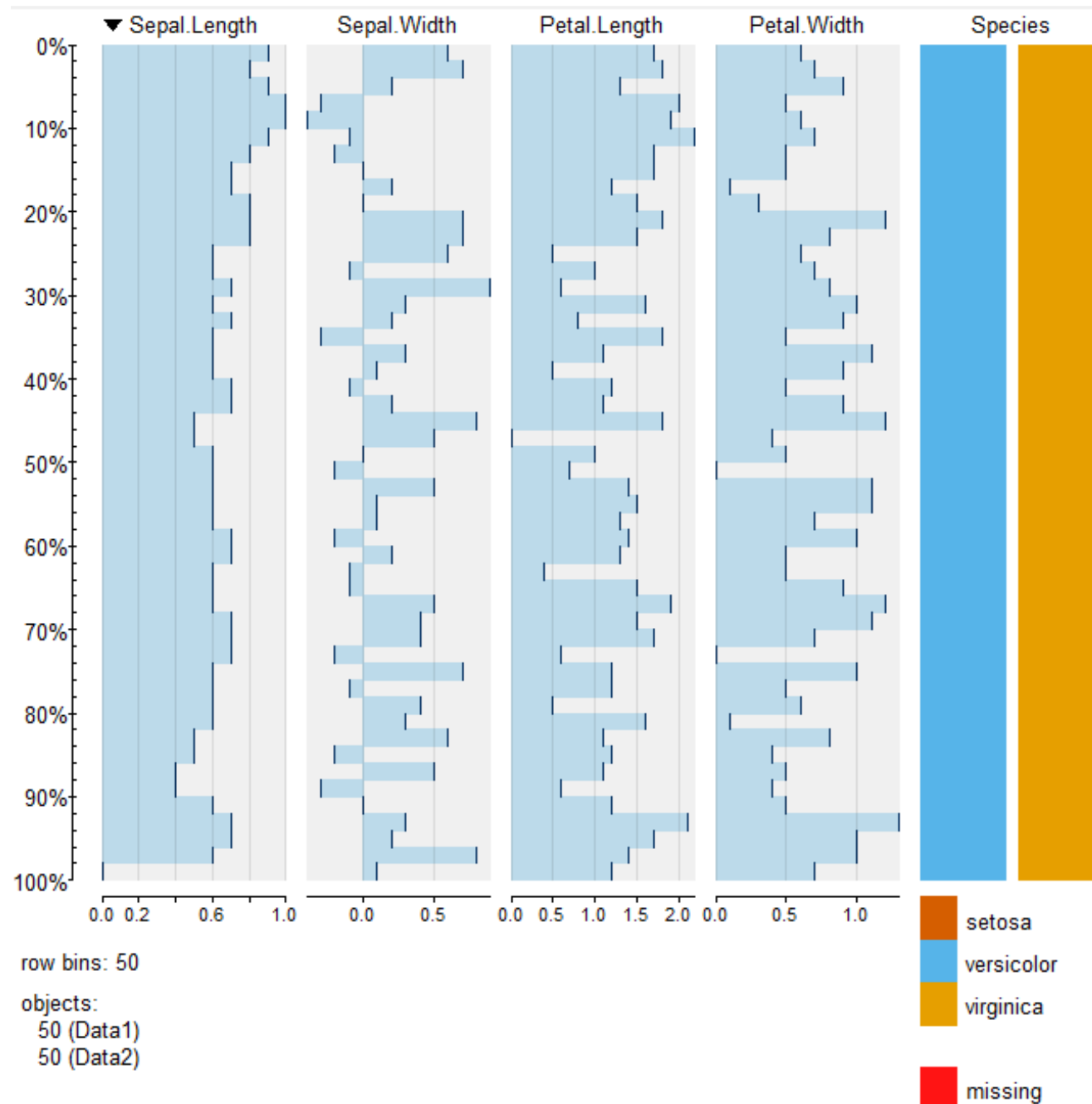
#図の出力はなしにして(plot=FALSE)、tableplot関数をつける

```
Tp1 <- tableplot(Data1, plot = FALSE, nBins=50)
```

```
Tp2 <- tableplot(Data2, plot = FALSE, nBins=50)
```

データ間の比較

```
#結果を引き算してplot  
plot(Tp1 - Tp2)
```



欠損値への対応

- 欠損値発生メカニズム

欠損値発生メカニズムの分類

- MCAR(Missing Completely At Random)
 - 列(変数)自体の値や、他の列との値とは完全に無関係にランダムに欠損。
- MAR(Missing At Rondon)
 - 列(変数)の値とは無関係だが他の列に関係して欠損値が発生している。
- MNAR(Missing Not At Rondon)
 - 列(変数)の値に関係して欠損。

D.B.Rubin. Inference and missing data. 63(3): 581-592, Biometrika, 1976

<http://www.stat.cmu.edu/~fienberg/Statistics36-756/Rubin-Biometrika-1976.pdf>

MCAR(Missing Completely At Random)

Id	A	B	C	D
1	3.3	6.2	1.1	7.0
2	1.2	5.1	3.4	5.9
3	4.2	2.8	3.2	8.2
4	1.1	6.1	2.6	5.3
5	4.5	3.1	5.7	2.1

欠損メカニズム
MCAR



Id	A	B	C	D
1	NA	6.2	1.1	7.0
2	1.2	5.1	3.4	5.9
3	NA	2.8	3.2	8.2
4	1.1	6.1	2.6	5.3
5	4.5	3.1	5.7	2.1

欠損値発生は、A列内の数とも、他の列の数とも関係ない。

MAR(Missing At Random)

Id	A	B	C	D
1	2.0	6.2	1.0	7.0
2	1.2	5.1	3.4	5.9
3	4.2	2.8	3.2	8.2
4	3.1	6.1	1.0	5.3
5	5.1	3.1	1.0	2.1

欠損メカニズム
MAR



Id	A	B	C	D
1	NA	6.2	1.0	7.0
2	1.2	5.1	3.4	5.9
3	4.2	2.8	3.2	8.2
4	NA	6.1	1.0	5.3
5	NA	3.1	1.0	2.1

A列内での欠損値発生は、C列の値に関係して起きている。この場合はC列の値が1.0ならA列で欠損値発生。

MNAR(Missing Not At Random)

Id	A	B	C	D
1	3.0	6.2	1.1	7.0
2	1.2	5.1	3.4	5.9
3	4.2	2.8	3.2	8.2
4	3.0	6.1	2.0	5.3
5	3.0	3.1	4.0	2.1

欠損メカニズム
MNAR



Id	A	B	C	D
1	NA	6.2	1.1	7.0
2	1.2	5.1	3.4	5.9
3	4.2	2.8	3.2	8.2
4	NA	6.1	2.0	5.3
5	NA	3.1	4.0	2.1

A列内でのみ関連がある値で、欠損値が発生。欠損値発生はほかの列の中身とは関係ない。

欠損値対応のフロー

- 3つのうちのどの欠損値発生メカニズムで欠損値が発生しているか突き止められるのが理想。
- 現実的には、MCARは観測データと欠損データの平均値と等分散の等質性を仮定しており、t検定などで、MCARの検定ができる(※1)
- MARについては、仮定の検証方法はほぼない。
- MNARについては、見つからず。

※1: t検定を多変量に拡張した、LittleによるMCAR検定。すべての変数を検査するが、type2 error(偽陰性)が起こりやすい欠点を持つ。

R.J.A.Little A test of missing completely at random for multivariate data with missing values
Journal of American Statistical Association 83(404):1198-1202, 1988.

https://www.jstor.org/stable/2290157?seq=1#page_scan_tab_contents

欠損値対応のフロー

- 以下では、miceパッケージによる、欠損値処理方法を解説する。

miceパッケージについて

- Miceパッケージの論文(※2)読みましょう。Rのサンプルコードとともに、丁寧な説明が書いてあります。67ページありますが、実質55ページ程度です。このスライドでも、論文のサンプルコードをいくつか使用しています。
- 余裕があれば、細かいことはReference manual(※3)読みましょう。

※2: S.v.Buuren and K.G-Oudshoorn. Mice: Multivariate imputation by chained equations in R.

Journal of Statistical Software, 45(3):1-67, 2011

<https://www.jstatsoft.org/article/view/v045i03>

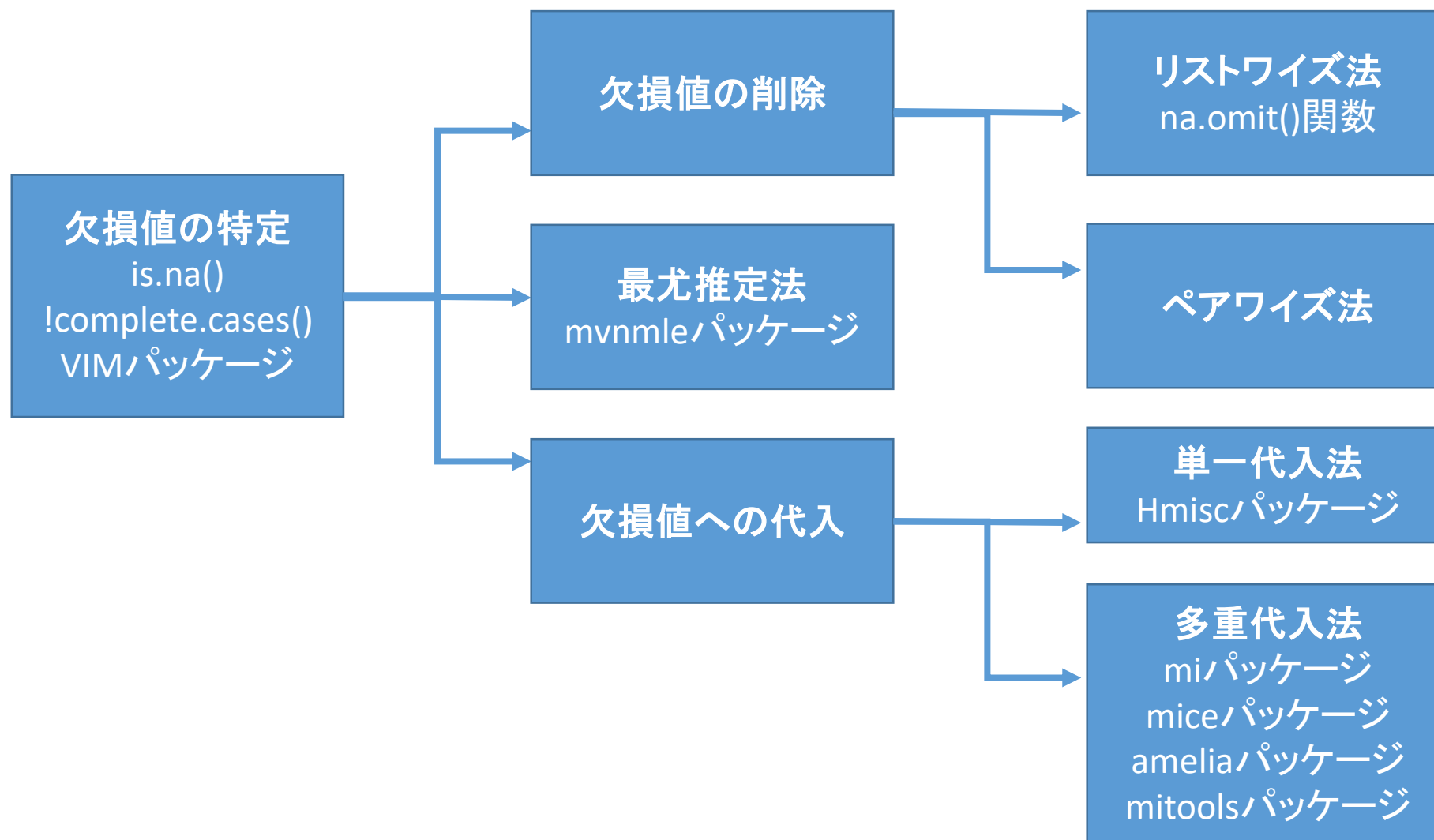
※3: <https://cran.r-project.org/web/packages/mice/mice.pdf>

欠損値対応のフロー

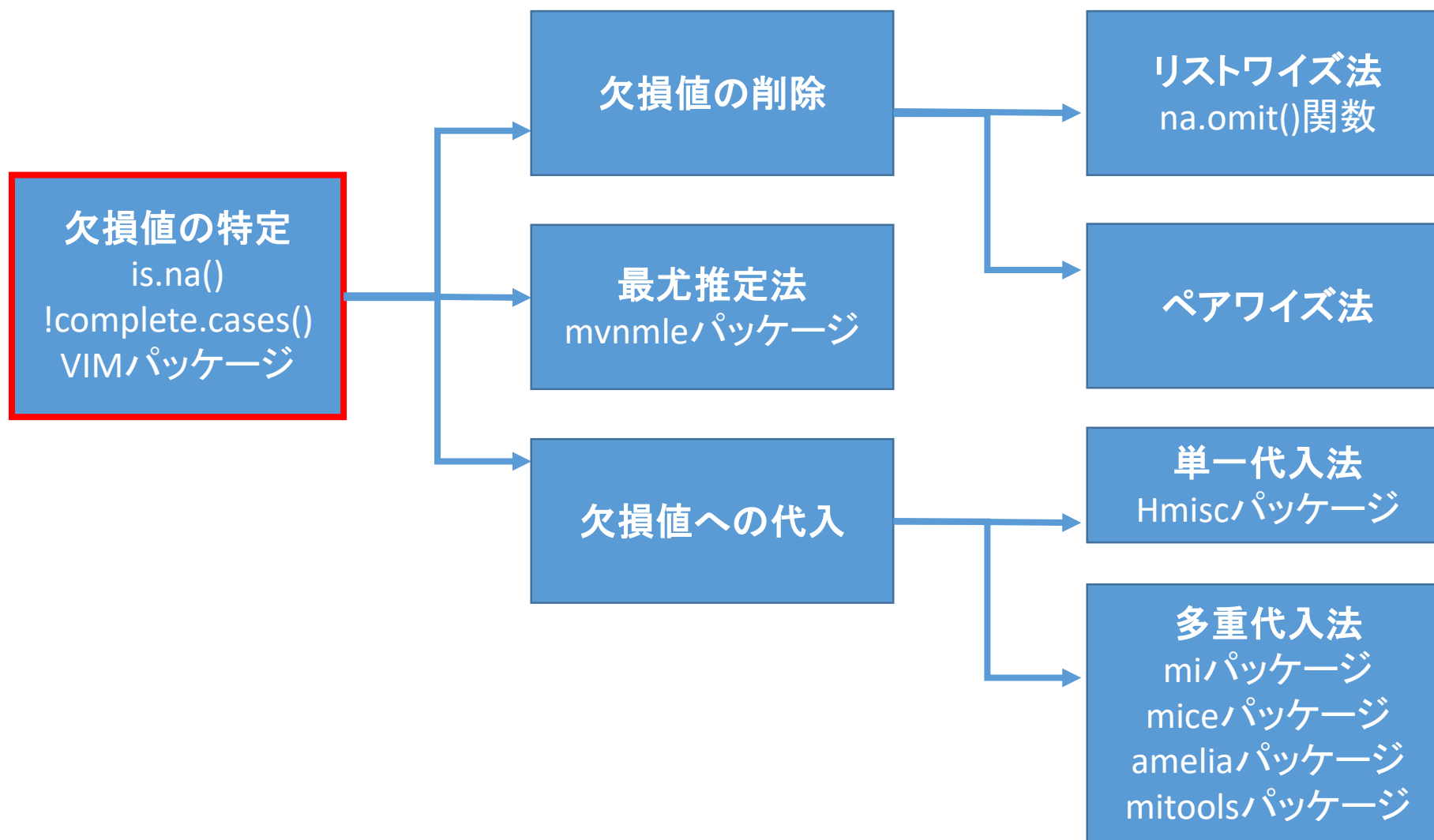
- 以下では、Kabacoff(※4)による、欠損値対応のワークフローを説明します。

※4: R.Kabacoff. R in Action: Data Analysis and Graphics with R. Manning Publications, 2011.
<http://kek.ksu.ru/EOS/DataMining/1379968983.pdf>

欠損値対応のフロー



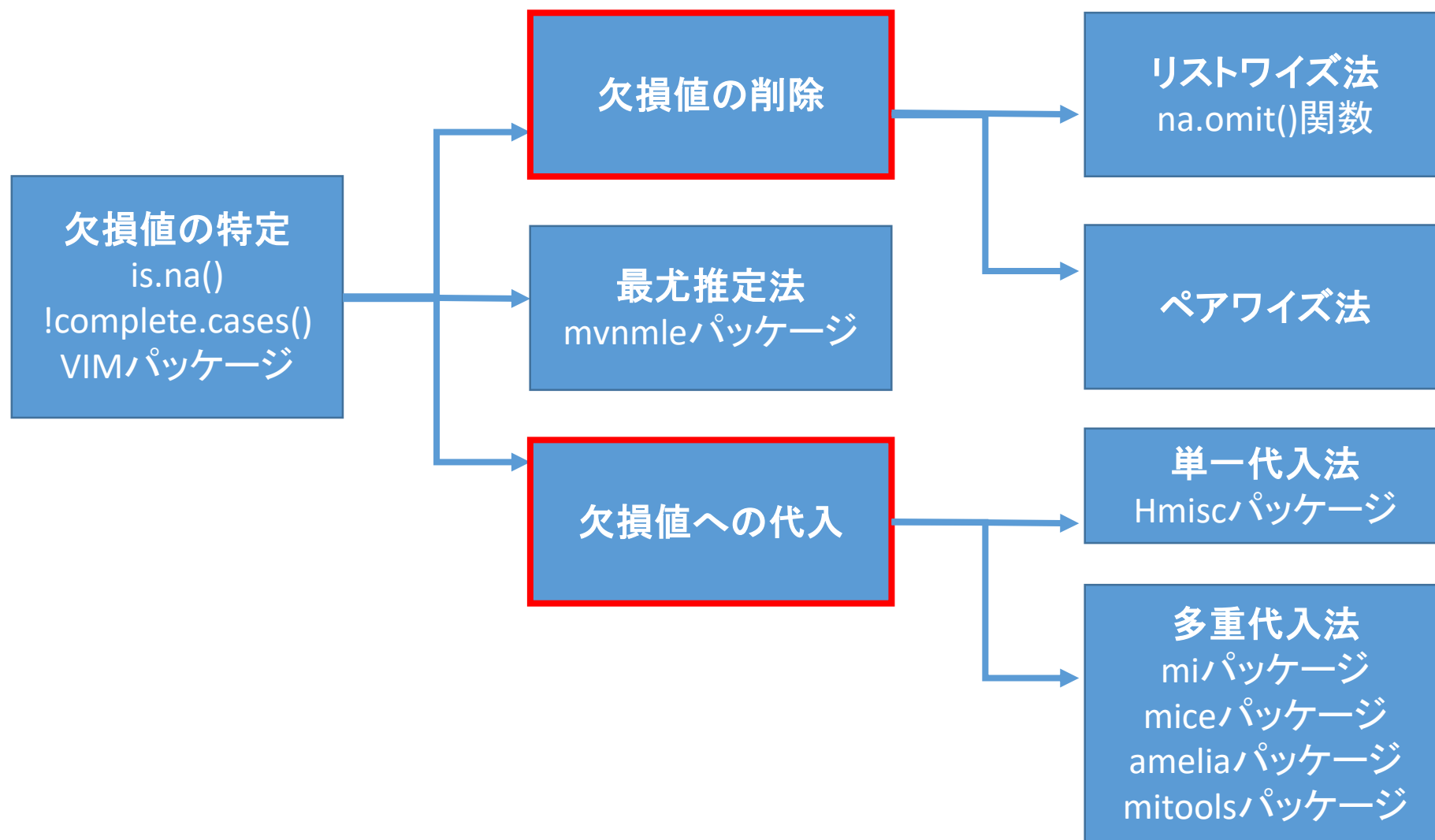
欠損値対応のフロー



欠損値対応フロー

- 欠損値を特定し、集計、またはtableplot関数などでのデータ可視化などにより、欠損値発生メカニズムを推定する。

欠損値対応のフロー




欠損値の削除と代入

- 欠損値が見つかった場合、対処方法として、主に削除と代入の2通りがある。
- それぞれさらに細分化され、
 - 削除は、リストワイズ法とペアワイズ法
 - 代入は、単一代入法、回帰代入法、確率的回帰代入法、完全情報最尤推定法、多重代入法がある。
- この中では、弱い過程の中で、バイアスのない推定値を与える、完全情報最尤推定法、多重代入法がよく用いられる方法。
- この二つは、欠損値発生メカニズムがMARであっても、バイアスのない推定値を算出する。

欠損値の削除と代入の参考文献

1, P.D. Allison. Missing data. SAGE Publications, Inc., 2001

Paul D. Allison氏がmissing dataについてすごくたくさん論文を出しています。



Paul D. Allison

Professor of Sociology, University of Pennsylvania
Quantitative Methods, Statistics, Sociology of Science, Missing Data,
Survival Analysis
Verified email at soc.upenn.edu - [Homepage](#)

[Follow](#)

Title	1-20	Cited by	Year
Missing data: Quantitative applications in the social sciences	PD Allison Sage Publications	5454 *	2002
Survival analysis using SAS: a practical guide	PD Allison SAS Institute	4766	2010
Event history analysis: Regression for longitudinal event data	PD Allison Sage	3992	1984

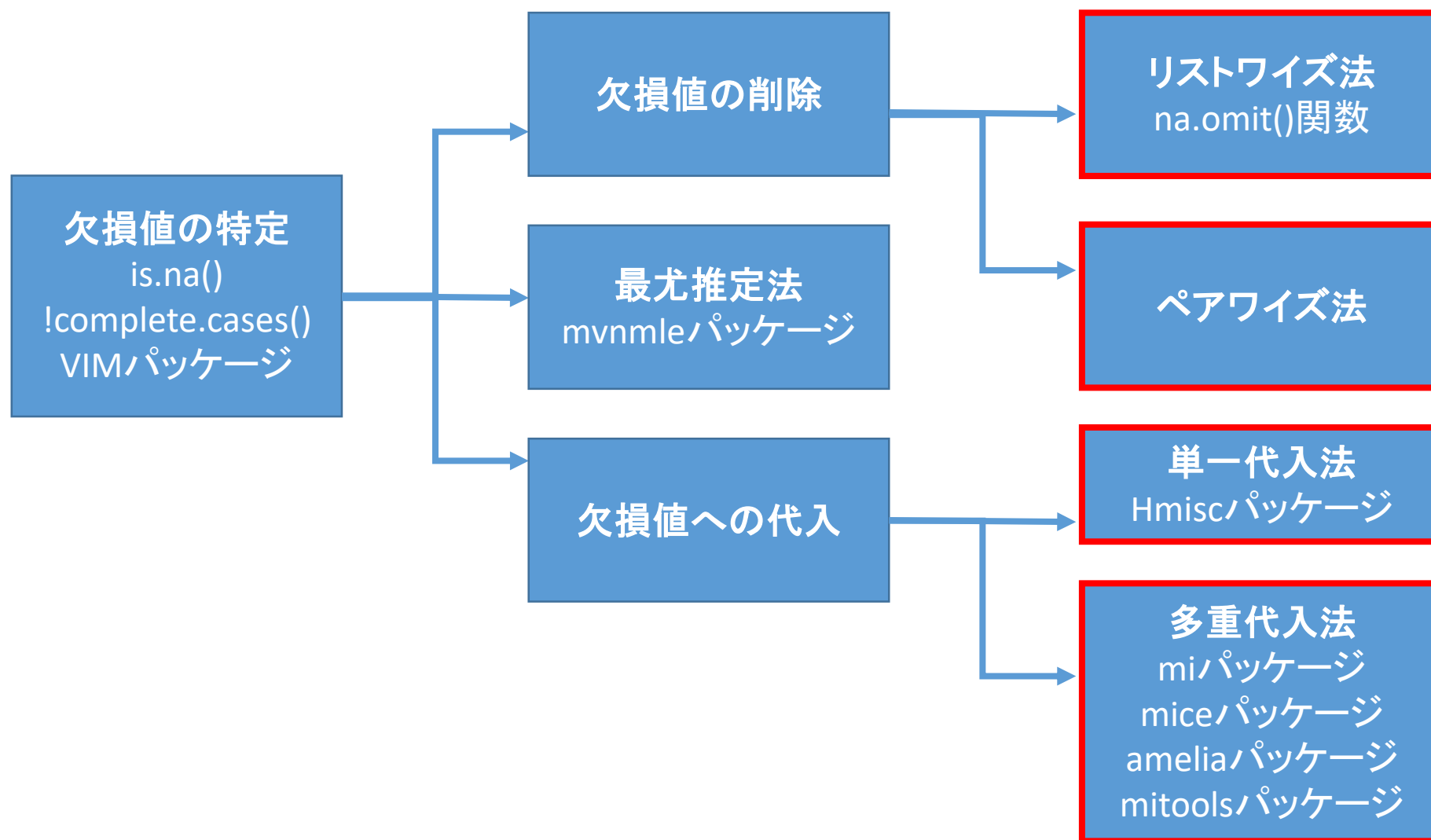
欠損値の削除と代入の参考文献

2,R.J.A.Little and D.B.Rubin.Statistical Analysis with missing data. Wiley-Interscience,2002.

3,C.K.Enders. Applied missing data Analysis.Guiford Press, 2010.

<http://hsta559s12.pbworks.com/w/file/fetch/52112520/enders.applied>

欠損値対応のフロー



削除または、代入方法

手法名	概要
リストワイズ法	欠損値を持つ行を削除
ペアワイズ法	相関係数や共分散等の算出の際2変数いずれかが欠損値をもつサンプルを削除
単一代入法	平均値や中央値など単一の値を欠損値へ代入。平均値を代入すれば平均代入法
回帰代入法	欠損値のないサンプルに回帰分析を行い、欠損値を含む項目の推定式を元に欠損値を補充。
確率的回帰代入法	回帰代入法により推定した値にランダムに誤差を与えて、欠損値を補充。
完全情報最尤推定法	サンプルごとに、欠損値パターンに応じた、尤度関数を仮定して、最尤推定を実施して、得られる多変量正規分布を用いて、平均値や分散共分散行列を推定。
多重代入法(※5)	欠損値に代入した、データセットを複数作成し、各データセットに対して、分析を実施し、その結果を統合することにより、欠損値を補充。

※5: S.v.Buuren. Flexible imputation of missing data. Chapman and Hall/CRC, 2012.

この本において、miceパッケージを中心として、Rの実装を交えながら多重代入法について詳しく書いてある。

[https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

[FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

[FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

miceパッケージのダウンロード・インストール

```
install.packages("mice", quiet = TRUE, dependencies=T)
```

```
library(mice)
```

```
> md.pattern(nhanes)
      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
.
```

使用するデータ

nhanes

```
data(nhanes)
```

```
str(nhanes)
```

```
> str(nhanes)
```

```
'data.frame':   25 obs. of  4 variables:  
 $ age: num  1 2 1 3 1 3 1 1 2 2 ...  
 $ bmi: num  NA 22.7 NA NA 20.4 NA 22.5 30.1 22 NA ...  
 $ hyp: num  NA 1 1 NA 1 NA 1 1 1 NA ...  
 $ chl: num  NA 187 187 NA 113 184 118 187 238 NA ...
```

欠損値発生パターンの可視化

md.pattern(nhanes)

各列、age,hyp,bmi,chlに関して、それぞれの欠損の有無(欠損ありが0でなしが1)の組み合わせパターンとその頻度を表している。

```
> md.pattern(nhanes)
```

	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	

欠損項目の数

頻度

組み合わせパターン

一行目はすべての項目で欠損値がないサンプルが13個あることを示している。

各項の欠損数の合計

欠損値の個数合計

md.pairs(nhanes)

```
> md.pairs(nhanes)
$rr
      age bmi hyp chl
age   25  16  17  15
bmi   16  16  16  13
hyp   17  16  17  14
chl   15  13  14  15

$rm
      age bmi hyp chl
age     0   9   8  10
bmi     0   0   0   3
hyp     0   1   0   3
chl     0   2   1   0

$mr
      age bmi hyp chl
age     0   0   0   0
bmi     9   0   1   2
hyp     8   0   0   1
chl    10   3   3   0

$mm
      age bmi hyp chl
age     0   0   0   0
bmi     0   9   8   7
hyp     0   8   8   7
chl     0   7   7  10
```

md.pairs関数は、2項目の欠損値有無の組み合わせごとに件数を集計する。
結果は左のようなリストが返されるが、
rは非欠損値、
mは欠損値
rrは行方向も列方向も両方とも非欠損値
rmは行方向の項目が、非欠損値、列方向の項目が欠損値。

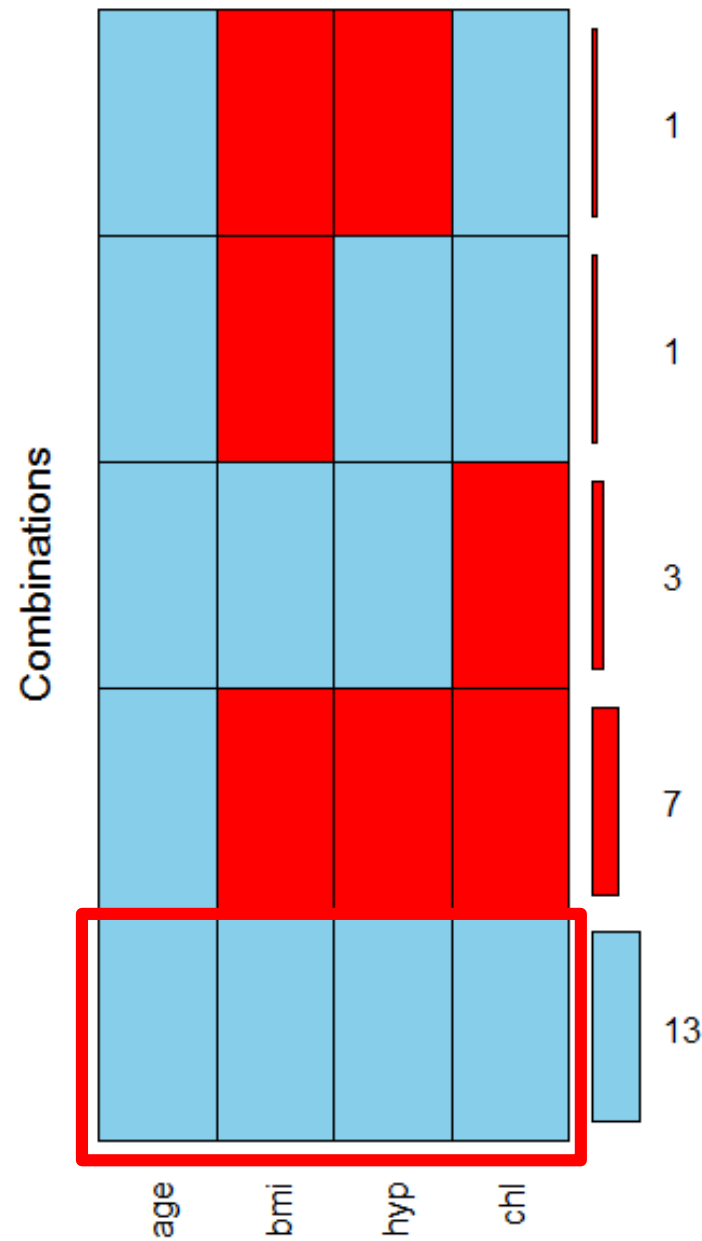
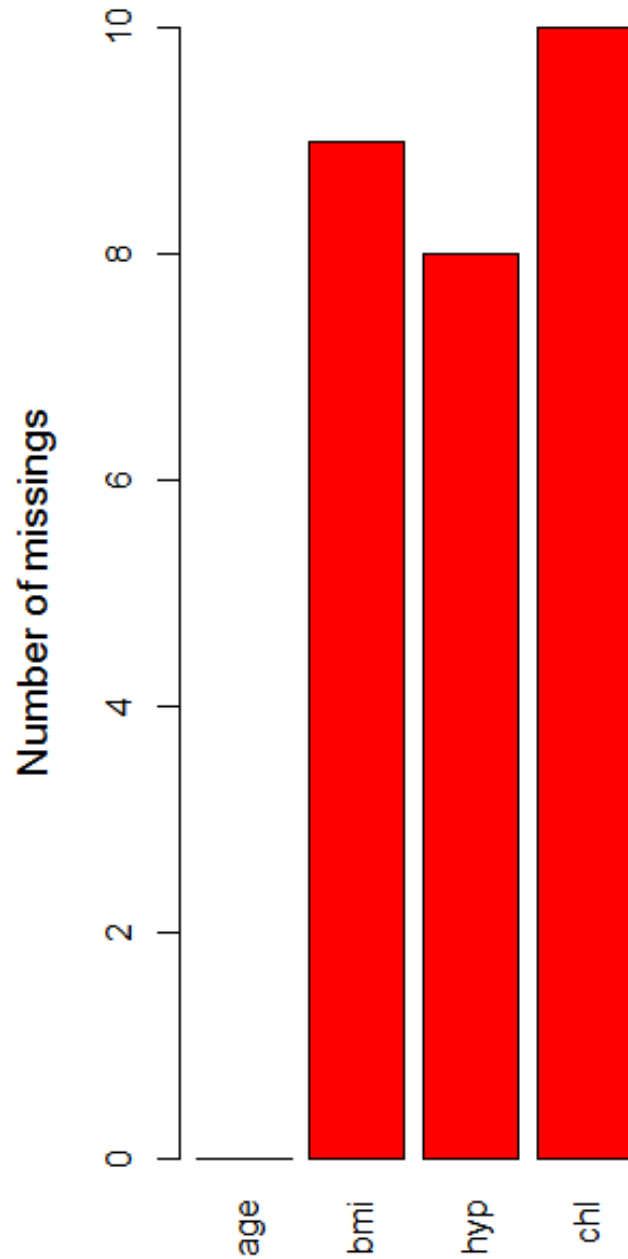
VIMパッケージのダウンロード・インストール

```
install.packages("VIM", quiet = TRUE, dependencies=T)  
library(VIM)
```


aggr関数による欠損値の可視化

```
aggr(nhanes, prop = FALSE, number = TRUE)
```

列数(変数)の数が少ない時には、aggr関数はおすすめ。
本質的には、miceパッケージのmd.pattern関数と同じ。



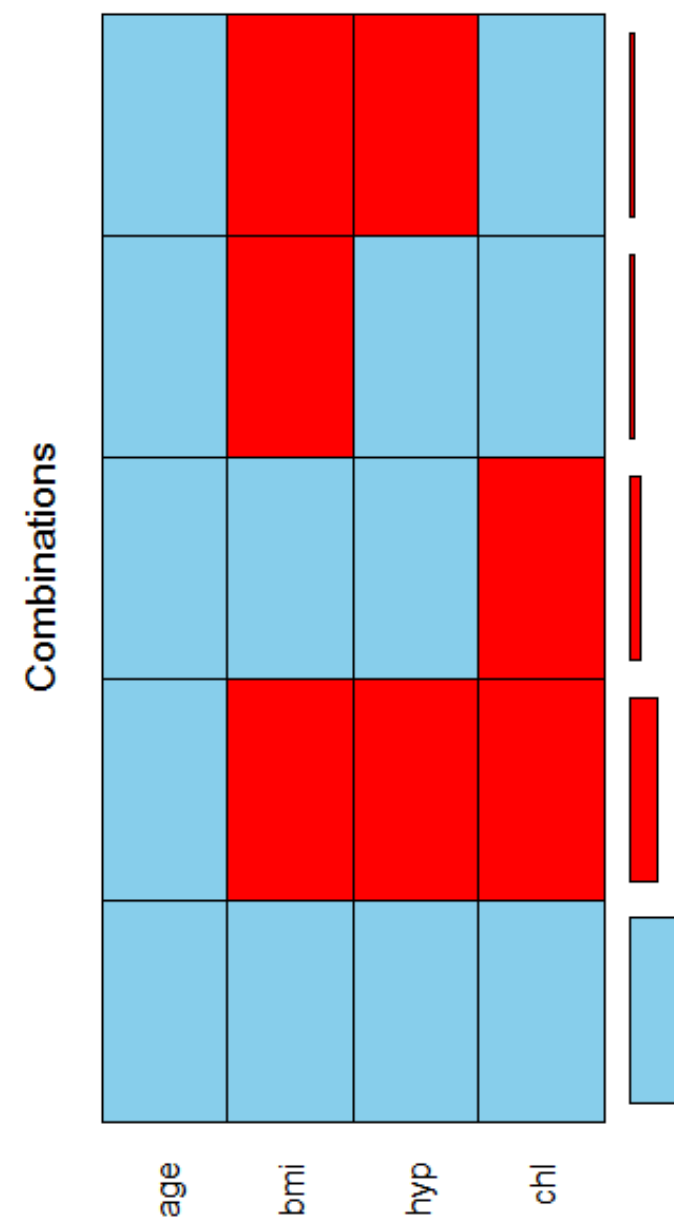
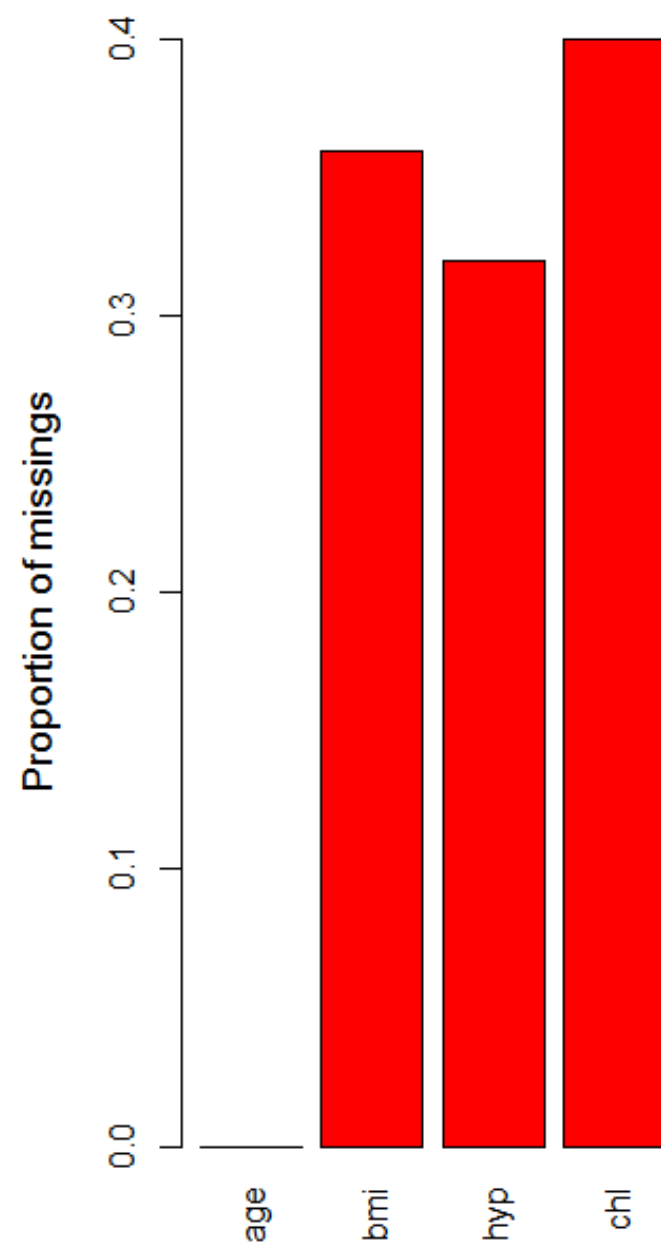
aggr関数は単一のデータ項目に関して、欠損値の棒グラフを出力し、複数のデータ項目に関して、欠損値がどのようなパターンで存在するかを可視化している。

例えばこれがすべての項目で欠損値がないパターンが13サンプルあることを示している。

aggr関数による欠損値の可視化

#prop = TRUEにすると、棒グラフの縦軸が割合になる(先ほどは件数)

```
aggr(nhanes, prop = TRUE, number = FALSE)
```

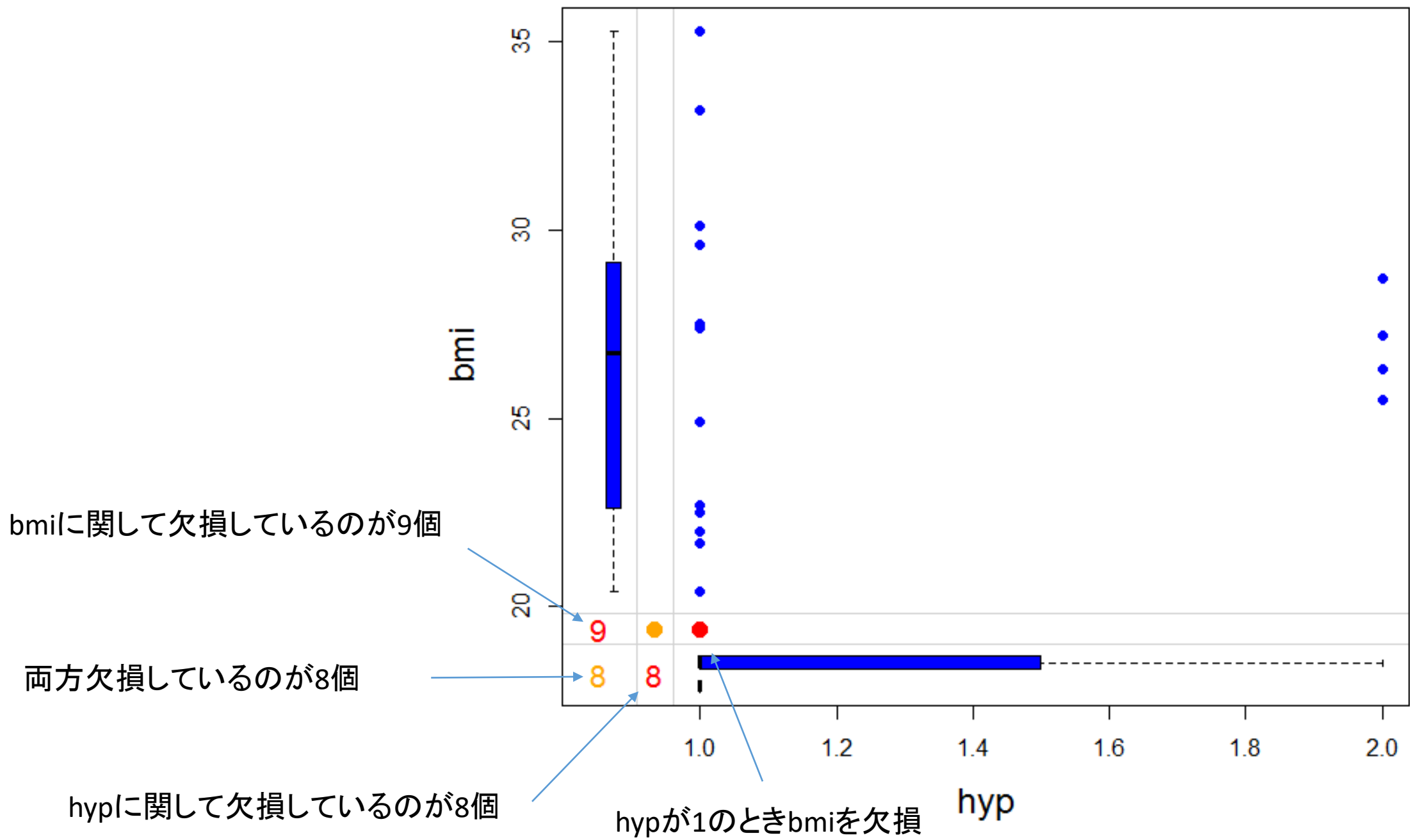


2項目の欠損状況のプロット

#marginplot関数により、2項目の欠損状況のプロットできる。

#ここでは、nhanesデータセットのhyp列とbmi列についての、それぞれの値の範囲と欠損値をプロットする。

```
marginplot(nhanes[, c("hyp", "bmi")], col = c("blue", "red", "orange"),  
           cex = 1.5, cex.lab = 1.5, cex.numbers = 1.3, pch = 20, ps = 1)
```

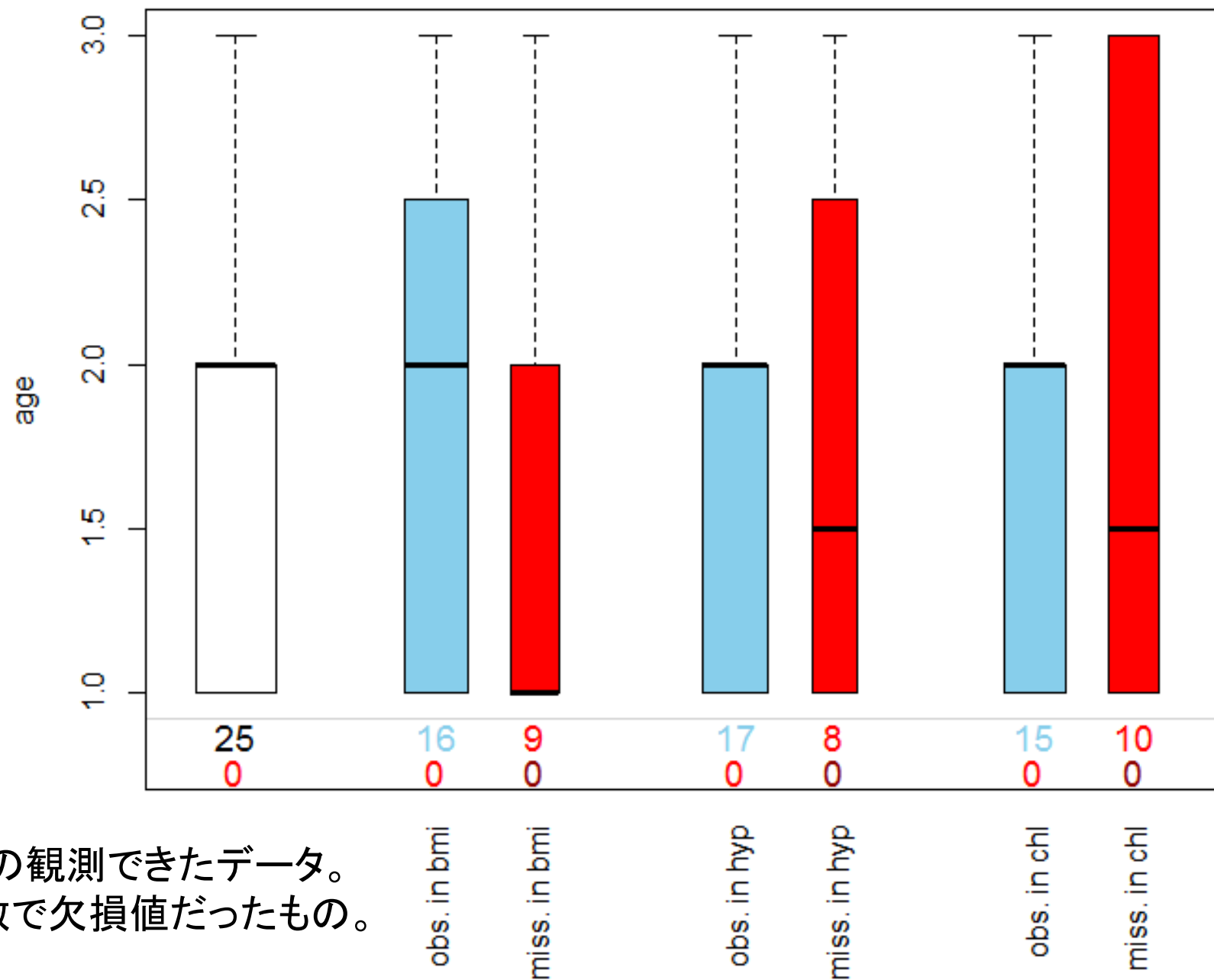


VIMパッケージのpbox関数

#VIMパッケージのpbox関数はデータの特定の項目と他の項目の間
の関係について欠損値の有無ごとに箱ひげ図をplotする。

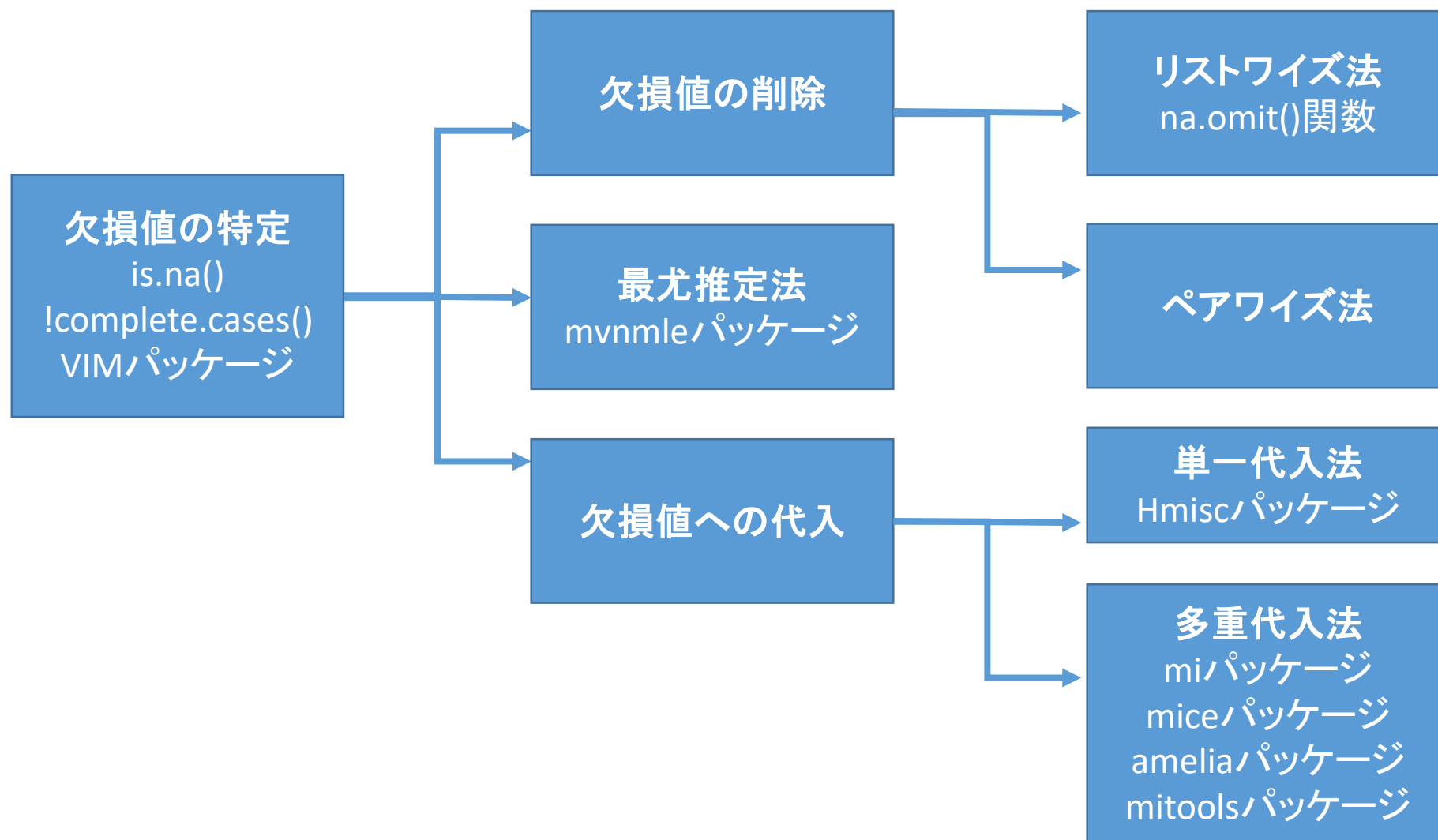
#次の例はnhanesデータセットに対して、項目ageと他の項目の欠損値
の有無に応じた項目ageの箱ひげ図のplot。

```
pbox(nhanes, pos = 1, int = FALSE, cex = 1.2)
```



青いほうが各変数の観測できたデータ。
赤いほうが、各変数で欠損値だったもの。

欠損値対応のフロー



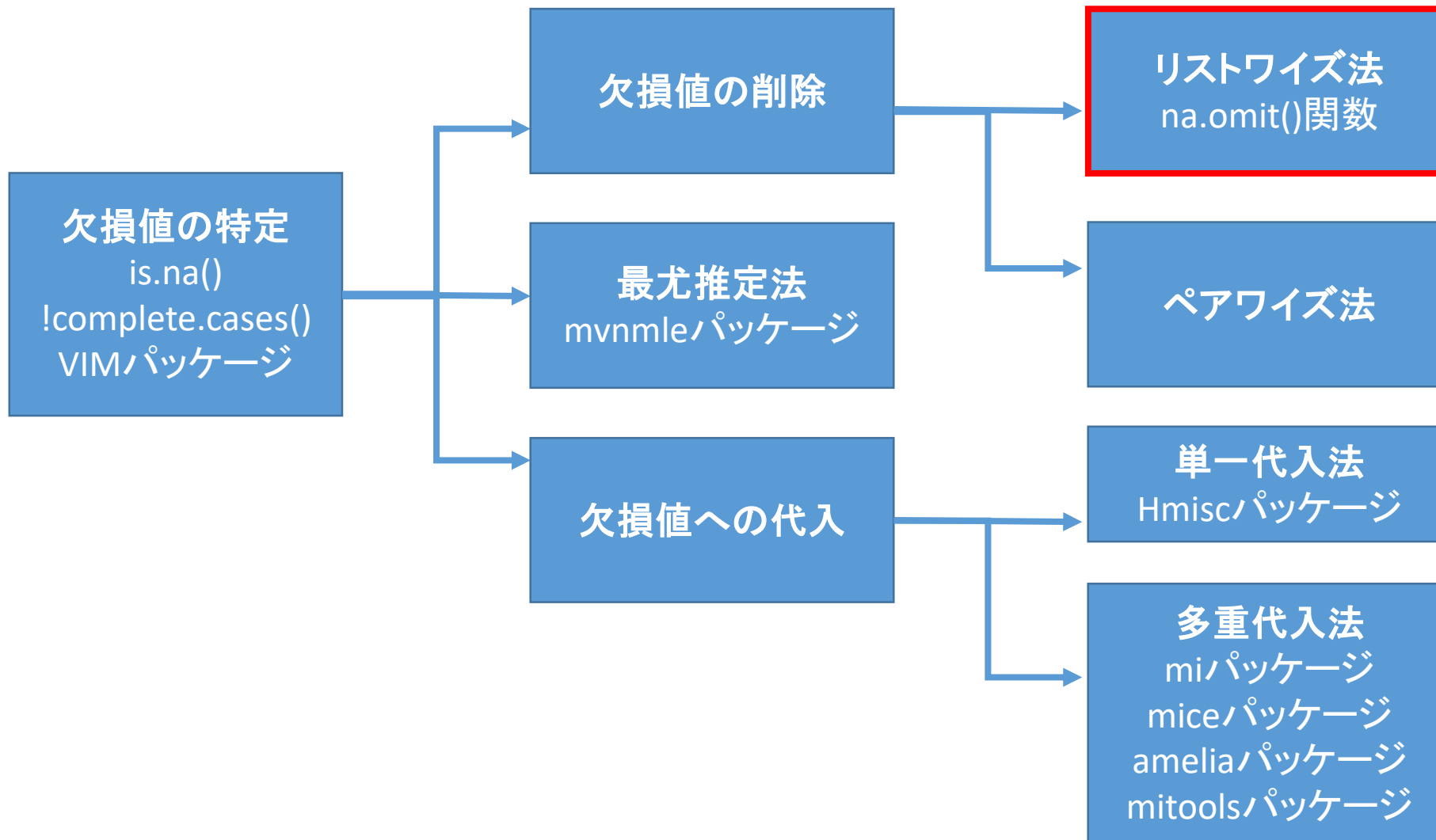
```
data(nhanes)
```

```
str(nhanes)
```

```
> str(nhanes)
```

```
'data.frame':   25 obs. of  4 variables:  
 $ age: num  1 2 1 3 1 3 1 1 2 2 ...  
 $ bmi: num  NA 22.7 NA NA 20.4 NA 22.5 30.1 22 NA ...  
 $ hyp: num  NA 1 1 NA 1 NA 1 1 1 NA ...  
 $ chl: num  NA 187 187 NA 113 184 118 187 238 NA ...
```

リストワイズ法



```
dim(nhanes)
```

```
head(nhanes, 10)
```

25行4列のデータであり、
欠損値(NA)が多数あるのが確認できる。
NAがある行を消去するのがリストワイズ法。
Rでは、na.omit()関数で実行する。

```
> dim(nhanes)
```

```
[1] 25  4
```

```
> head(nhanes, 10)
```

	age	bmi	hyp	chl
1	1	NA	NA	NA
2	2	22.7	1	187
3	1	NA	1	187
4	3	NA	NA	NA
5	1	20.4	1	113
6	3	NA	NA	184
7	1	22.5	1	118
8	1	30.1	1	187
9	2	22.0	1	238
10	2	NA	NA	NA

リストワイズ法

#リストワイズ法を使用

```
nhanes.lw <- na.omit(nhanes)
```

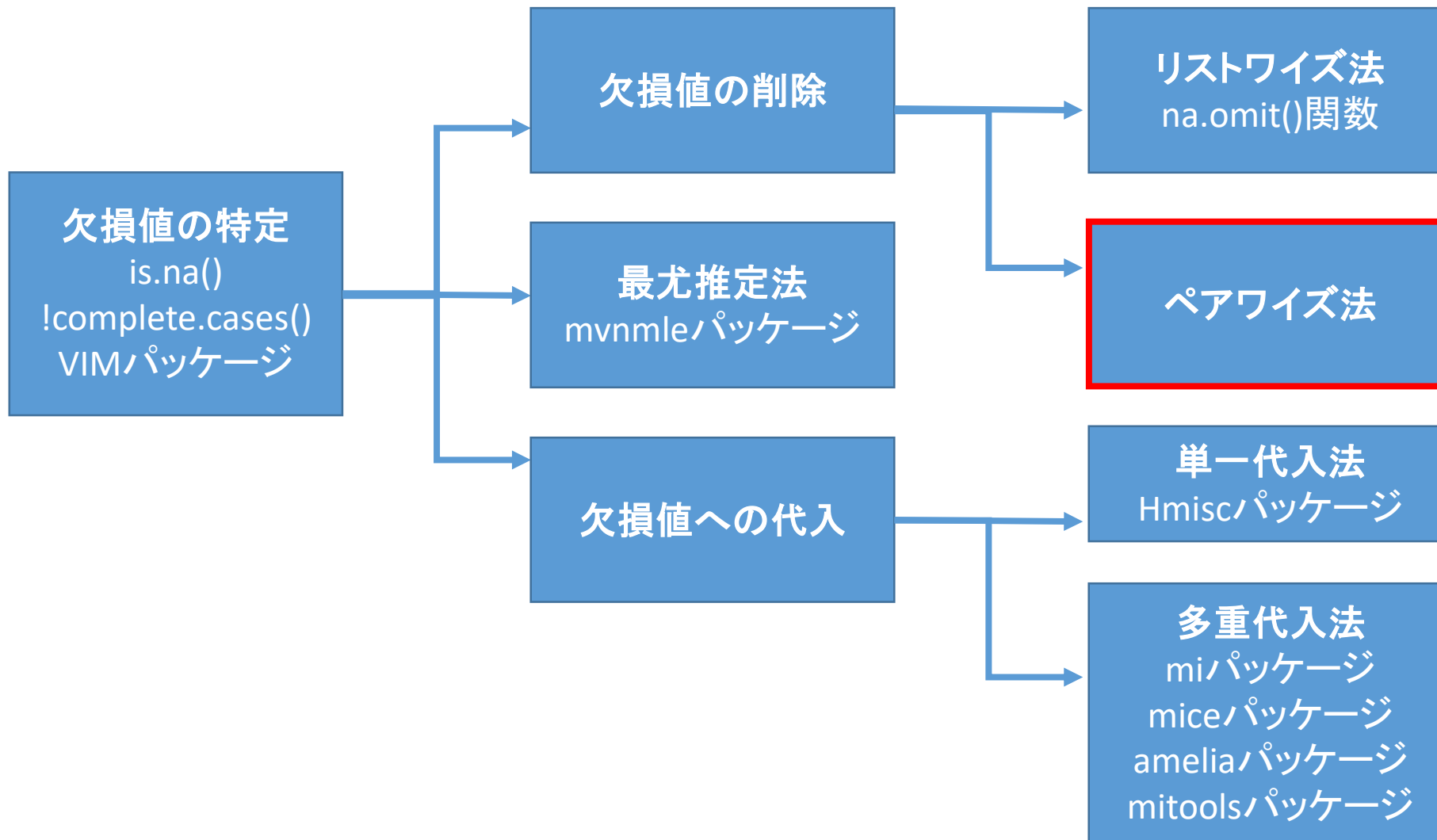
#リストワイズ法使用後の結果

```
dim(nhanes.lw)
```

```
head(nhanes.lw, 10)
```

```
> nhanes.lw <- na.omit(nhanes)
> dim(nhanes.lw)
[1] 13  4
> head(nhanes.lw, 10)
      age  bmi hyp chl
2       2 22.7  1 187
5       1 20.4  1 113
7       1 22.5  1 118
8       1 30.1  1 187
9       2 22.0  1 238
13      3 21.7  1 206
14      2 28.7  2 204
17      3 27.2  2 284
18      2 26.3  2 199
19      1 35.3  1 218
```

ペアワイズ法



ペアワイズ法

- ペアワイズ法は、リストワイズ法において、欠損値を含むサンプルをすべて除去した結果、データの情報量の損失が起きる問題を緩和するために提案された手法。
- ペアワイズ法では、相関係数や共分散などを求めるときに、2変数のいずれかが欠損値を持つ値を使用せずに、これらを計算する手法である。
- ペアワイズ法は、欠損値生成メカニズムとしてMCARを想定しており、データがその仮定を満たさない場合は、算出する相関係数や共分散にバイアスが生じるので注意。

ペアワイズ法

#相関係数を算出するcor関数のuse引数に、pairwise.complete.obsを指定することで相関係数を算出できる。
`cor(airquality, use = "pairwise.complete.obs")`

```
> cor(airquality, use = "pairwise.complete.obs")
```

	Ozone	Solar.R	Wind	Temp	Month	Day
Ozone	1.00000000	0.34834169	-0.60154653	0.6983603	0.164519314	-0.013225647
Solar.R	0.34834169	1.00000000	-0.05679167	0.2758403	-0.075300764	-0.150274979
Wind	-0.60154653	-0.05679167	1.00000000	-0.4579879	-0.178292579	0.027180903
Temp	0.69836034	0.27584027	-0.45798788	1.00000000	0.420947252	-0.130593175
Month	0.16451931	-0.07530076	-0.17829258	0.4209473	1.000000000	-0.007961763
Day	-0.01322565	-0.15027498	0.02718090	-0.1305932	-0.007961763	1.000000000

pairwise.complete.obsはpairwiseと省略も可能。

ペアワイズ法

#cov関数での共分散の算出(use引数でpairwise.complete.obsを指定)

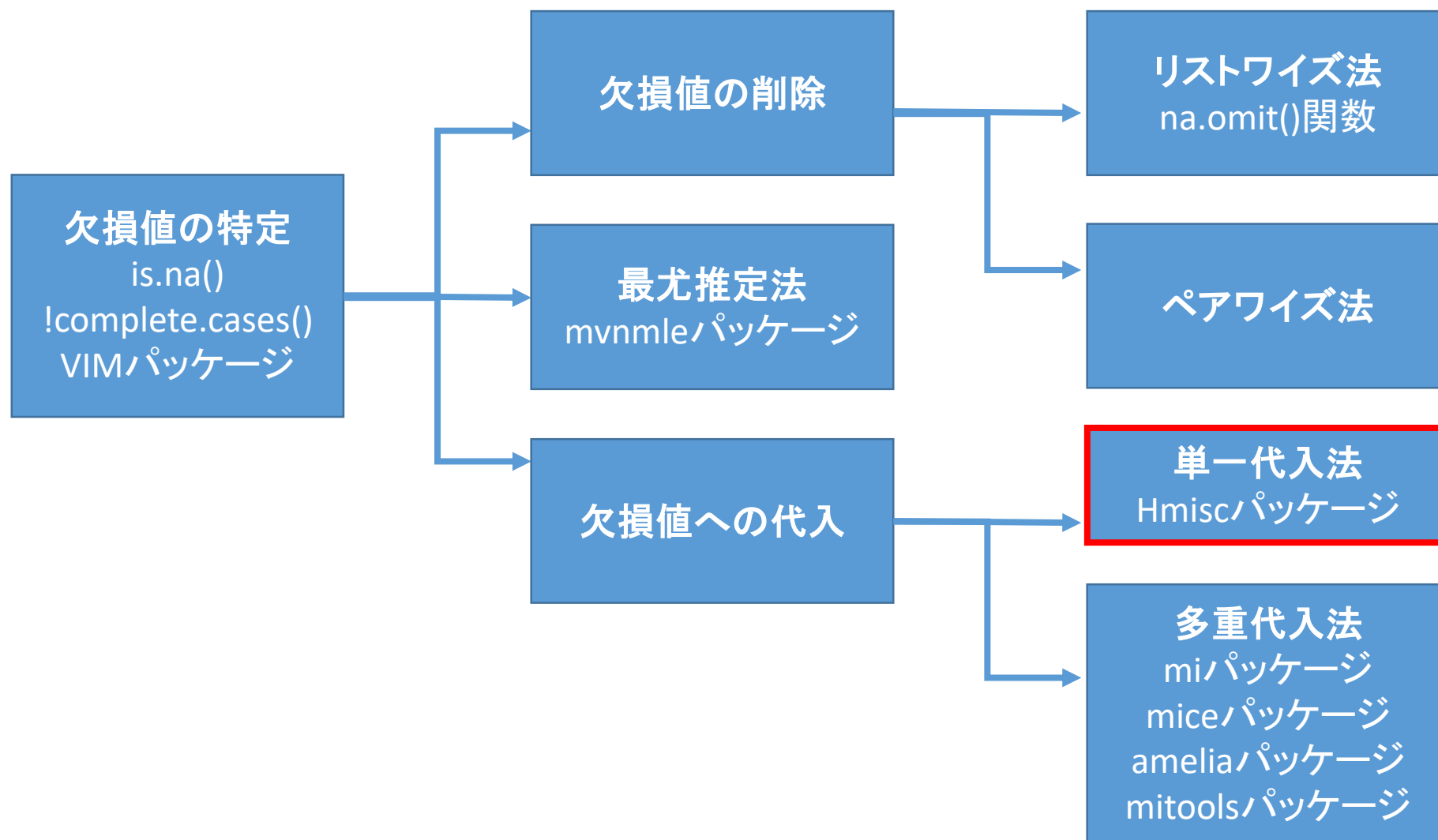
```
cov(airquality, use = "pairwise.complete.obs")
```

```
> cov(airquality, use = "pairwise.complete.obs")
```

	Ozone	Solar.R	Wind	Temp	Month	Day
Ozone	1088.200525	1056.583456	-70.9385307	218.521214	8.0089205	-3.8175412
Solar.R	1056.583456	8110.519414	-17.9459707	229.159754	-9.5222485	-119.0259802
Wind	-70.938531	-17.945971	12.4115385	-15.272136	-0.8897532	0.8488519
Temp	218.521214	229.159754	-15.2721362	89.591331	5.6439628	-10.9574303
Month	8.008921	-9.522248	-0.8897532	5.643963	2.0065359	-0.0999742
Day	-3.817541	-119.025980	0.8488519	-10.957430	-0.0999742	78.5797214

pairwise.complete.obsはpairwiseと省略も可能。

欠損値対応のフロー



単一代入法

- 欠損値を補完する最も簡単な方法は、平均値または中央値を代入することである。これを単一代入法(single imputation)と呼ぶ。
- 平均値を代入する単一代入法を特に、平均値代入法(mean imputation)と呼ぶ。
- 以下では、miceパッケージのmice関数を用いて、平均値代入法を実行する。

#employee_IQ_JP.csvというファイルをカレントディレクトリに配置

```
employee.IQ.JP <- read.csv("employee_IQ_JP.csv", row.names =  
NULL,colClasses = c(rep("integer", 3), "factor", "integer", "factor",  
"integer", "factor"))
```

平均値代入法の実行

```
imp <- mice(subset(employee.IQ.JP, select = c(IQ, MAR)), method =  
"mean", m = 1, maxit = 1)
```

```
imp
```

```
> employee.IQ.JP <- read.csv("employee_IQ_JP.csv", row.names = N
> imp <- mice(subset(employee.IQ.JP, select = c(IQ, MAR)), metho
```

```
iter imp variable
  1    1    MAR
```

```
> imp
```

Multiply imputed data set

Call:

```
mice(data = subset(employee.IQ.JP, select = c(IQ, MAR)), m = 1,
      method = "mean", maxit = 1)
```

Number of multiple imputations: 1

Missing cells per column:

IQ	MAR
0	5

Imputation methods:

IQ	MAR
"mean"	"mean"

VisitSequence:

MAR
2

PredictorMatrix:

	IQ	MAR
IQ	0	0
MAR	1	0

Random generator seed value: NA

多重代入法の実行回数。今回は指定しておらず、mice関数のデフォルト値であるm=1が使用されている。

各項目で欠損しているサンプルの件数。IQは欠損値がなく、MARは欠損値5個。

欠損値の補充に使用した方法を示している。次ページにmethodとして指定可能な手法を示す。

補完する列の順序を示す。今回はIQは補完する必要がないので、MARだけを補完する。

欠損が生じている各変数を補完するために使用された予測変数を現す行列。1が付けられた変数は予測に使用。0は不使用。

乱数の種を使用したかどうかを表す。

削除または、代入方法

手法名	概要
リストワイズ法	欠損値を持つ行を削除
ペアワイズ法	相関係数や共分散等の算出の際2変数いずれかが欠損値をもつサンプルを削除
単一代入法	平均値や中央値など単一の値を欠損値へ代入。平均値を代入すれば平均代入法
回帰代入法	欠損値のないサンプルに回帰分析を行い、欠損値を含む項目の推定式を元に欠損値を補充。
確率的回帰代入法	回帰代入法により推定した値にランダムに誤差を与えて、欠損値を補充。
完全情報最尤推定法	サンプルごとに、欠損値パターンに応じた、尤度関数を仮定して、最尤推定を実施して、得られる多変量正規分布を用いて、平均値や分散共分散行列を推定。
多重代入法(※5)	欠損値に代入した、データセットを複数作成し、各データセットに対して、分析を実施し、その結果を統合することにより、欠損値を補充。

※5: S.v.Buuren. Flexible imputation of missing data. Chapman and Hall/CRC, 2012.

この本において、miceパッケージを中心として、Rの実装を交えながら多重代入法について詳しく書いてある。

[https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

[FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

[FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

回帰代入法

#データの読み込み

```
employee.IQ.JP <- read.csv("employee_IQ_JP.csv", row.names = NULL,  
  colClasses = c(rep("integer", 3), "factor", "integer", "factor",  
  "integer", "factor"))
```

#回帰式の推定

```
fit.lm <- lm(MAR ~ IQ, data = employee.IQ.JP)
```

#欠損値の予測

```
pred <- predict(fit.lm, subset(employee.IQ.JP, is.na(MAR)))
```


#欠損値の補完

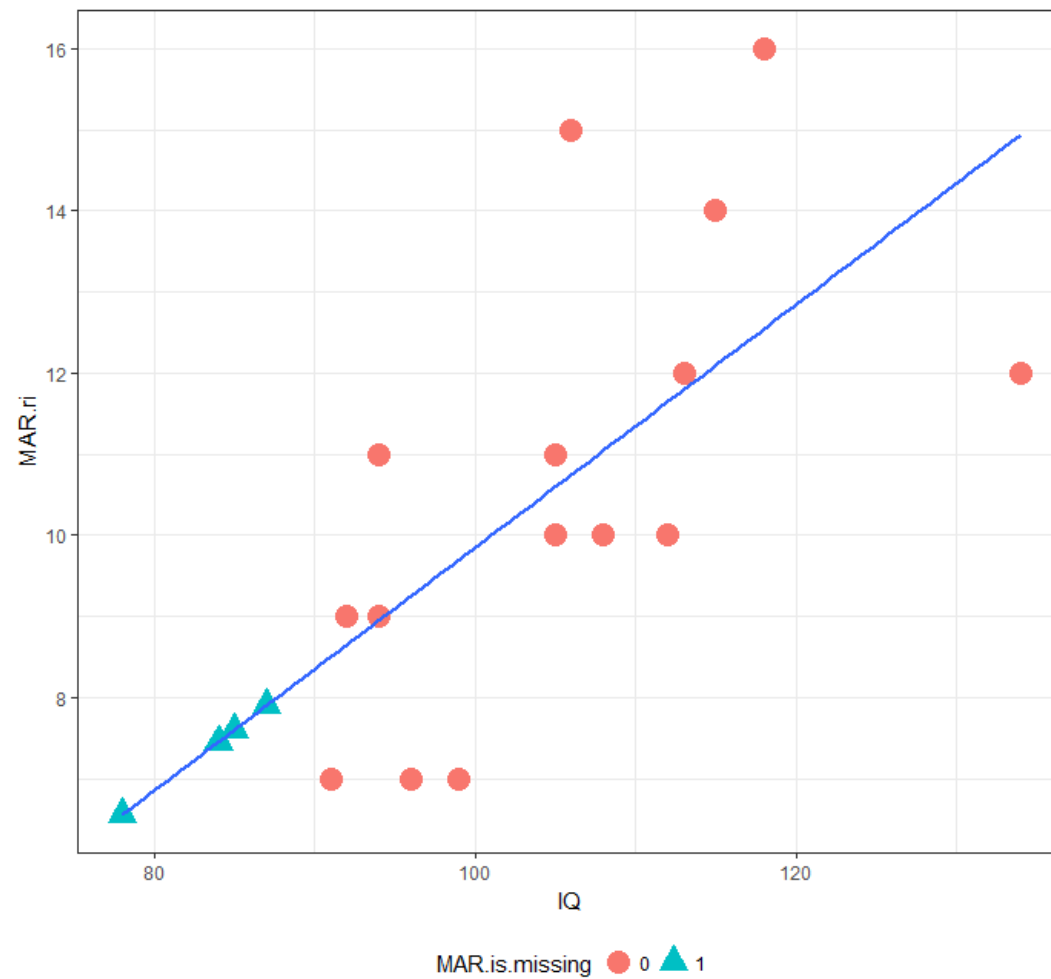
```
employee.IQ.JP$MAR.ri <- employee.IQ.JP$MAR
```

```
employee.IQ.JP$MAR.ri[is.na(employee.IQ.JP$MAR)] <- pred
```

散布図のプロット

```
p <- ggplot(data = employee.IQ.JP, aes(x = IQ, y = MAR.ri)) +  
  geom_point(aes(colour  
    = MAR.is.missing, group = MAR.is.missing, shape = MAR.is.missing),  
    size = 5) +  
  geom_smooth(method = lm, se = FALSE) + theme_bw() %+replace%  
  theme(legend.position= "bottom")  
print(p)
```

回帰代入法で補完された散布図



削除または、代入方法

手法名	概要
リストワイズ法	欠損値を持つ行を削除
ペアワイズ法	相関係数や共分散等の算出の際2変数いずれかが欠損値をもつサンプルを削除
単一代入法	平均値や中央値など単一の値を欠損値へ代入。平均値を代入すれば平均代入法
回帰代入法	欠損値のないサンプルに回帰分析を行い、欠損値を含む項目の推定式を元に欠損値を補充。
確率的回帰代入法	回帰代入法により推定した値にランダムに誤差を与えて、欠損値を補充。
完全情報最尤推定法	サンプルごとに、欠損値パターンに応じた、尤度関数を仮定して、最尤推定を実施して、得られる多変量正規分布を用いて、平均値や分散共分散行列を推定。
多重代入法(※5)	欠損値に代入した、データセットを複数作成し、各データセットに対して、分析を実施し、その結果を統合することにより、欠損値を補充。

※5: S.v.Buuren. Flexible imputation of missing data. Chapman and Hall/CRC, 2012.

この本において、miceパッケージを中心として、Rの実装を交えながら多重代入法について詳しく書いてある。

[https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

[FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

[FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

確率的回帰代入法

#データの読み込み

```
employee.IQ.JP <- read.csv("employee_IQ_JP.csv", row.names = NULL,  
  colClasses = c(rep("integer", 3), "factor", "integer", "factor",  
  "integer", "factor"))
```

#確率的回帰代入法の実行

```
imp <- mice(subset(employee.IQ.JP, select = c(IQ, MAR)), method =  
"norm.nob", m = 1, maxit = 1, seed = 123)
```

#欠損値の補完

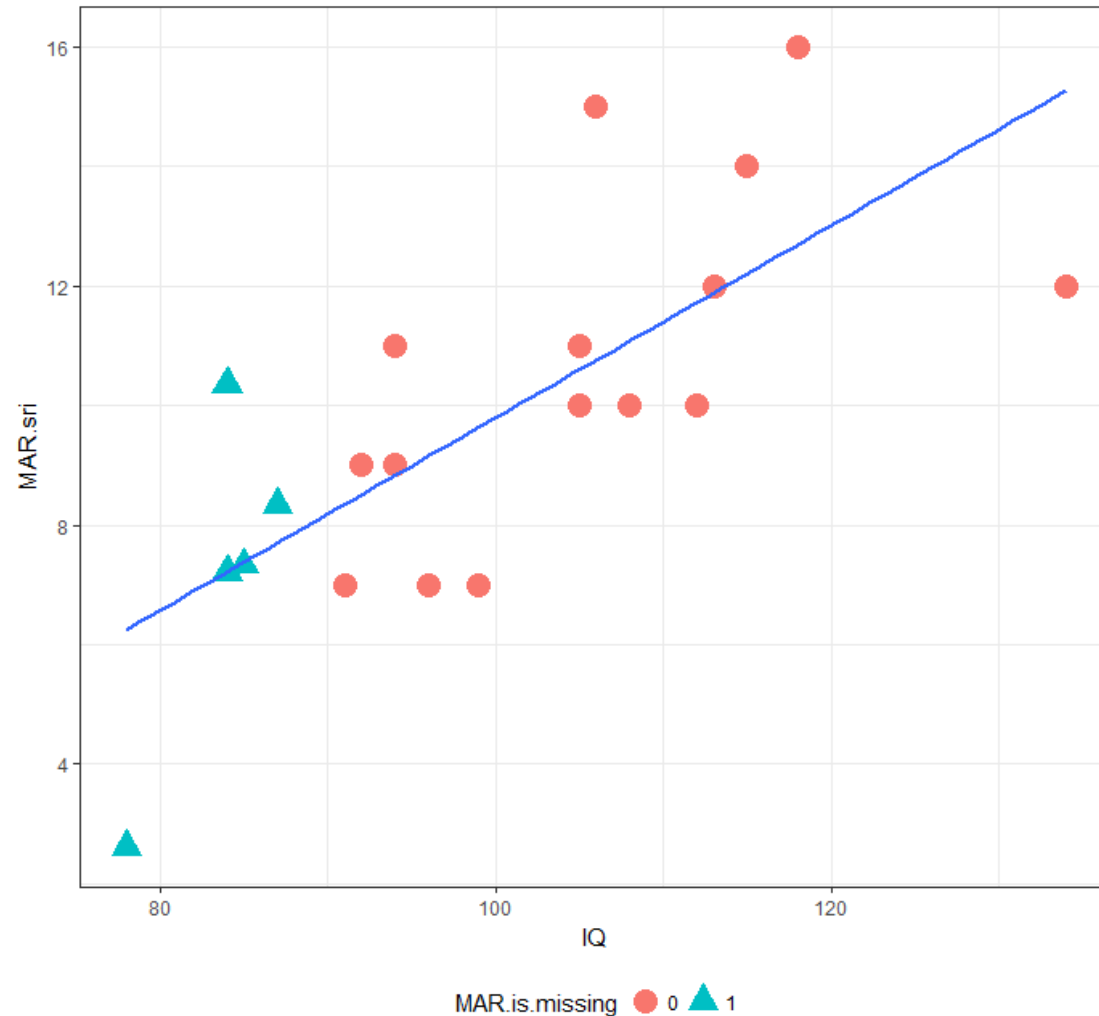
```
employee.IQ.JP$MAR.sri <- employee.IQ.JP$MAR  
employee.IQ.JP$MAR.sri[is.na(employee.IQ.JP$MAR)] <-  
unlist(imp$imp$MAR)
```

```
> imp <- mice(subset(employee.IQ.JP, select  
  
  iter imp variable  
    1    1    MAR
```

補完されたデータでの散布図作成

```
p <- ggplot(data = employee.IQ.JP, aes(x = IQ, y = MAR.sri)) +  
  geom_point(aes(colour= MAR.is.missing, group = MAR.is.missing,  
    shape = MAR.is.missing), size = 5) + geom_smooth(method = lm,  
    se = FALSE) + theme_bw() %+replace% theme(legend.position =  
  "bottom")  
print(p)
```

確率的回帰代入法で補完されたデータのプロット



削除または、代入方法

手法名	概要
リストワイズ法	欠損値を持つ行を削除
ペアワイズ法	相関係数や共分散等の算出の際2変数いずれかが欠損値をもつサンプルを削除
単一代入法	平均値や中央値など単一の値を欠損値へ代入。平均値を代入すれば平均代入法
回帰代入法	欠損値のないサンプルに回帰分析を行い、欠損値を含む項目の推定式を元に欠損値を補充。
確率的回帰代入法	回帰代入法により推定した値にランダムに誤差を与えて、欠損値を補充。
完全情報最尤推定法	サンプルごとに、欠損値パターンに応じた、尤度関数を仮定して、最尤推定を実施して、得られる多変量正規分布を用いて、平均値や分散共分散行列を推定。
多重代入法(※5)	欠損値に代入した、データセットを複数作成し、各データセットに対して、分析を実施し、その結果を統合することにより、欠損値を補充。

※5: S.v.Buuren. Flexible imputation of missing data. Chapman and Hall/CRC, 2012.

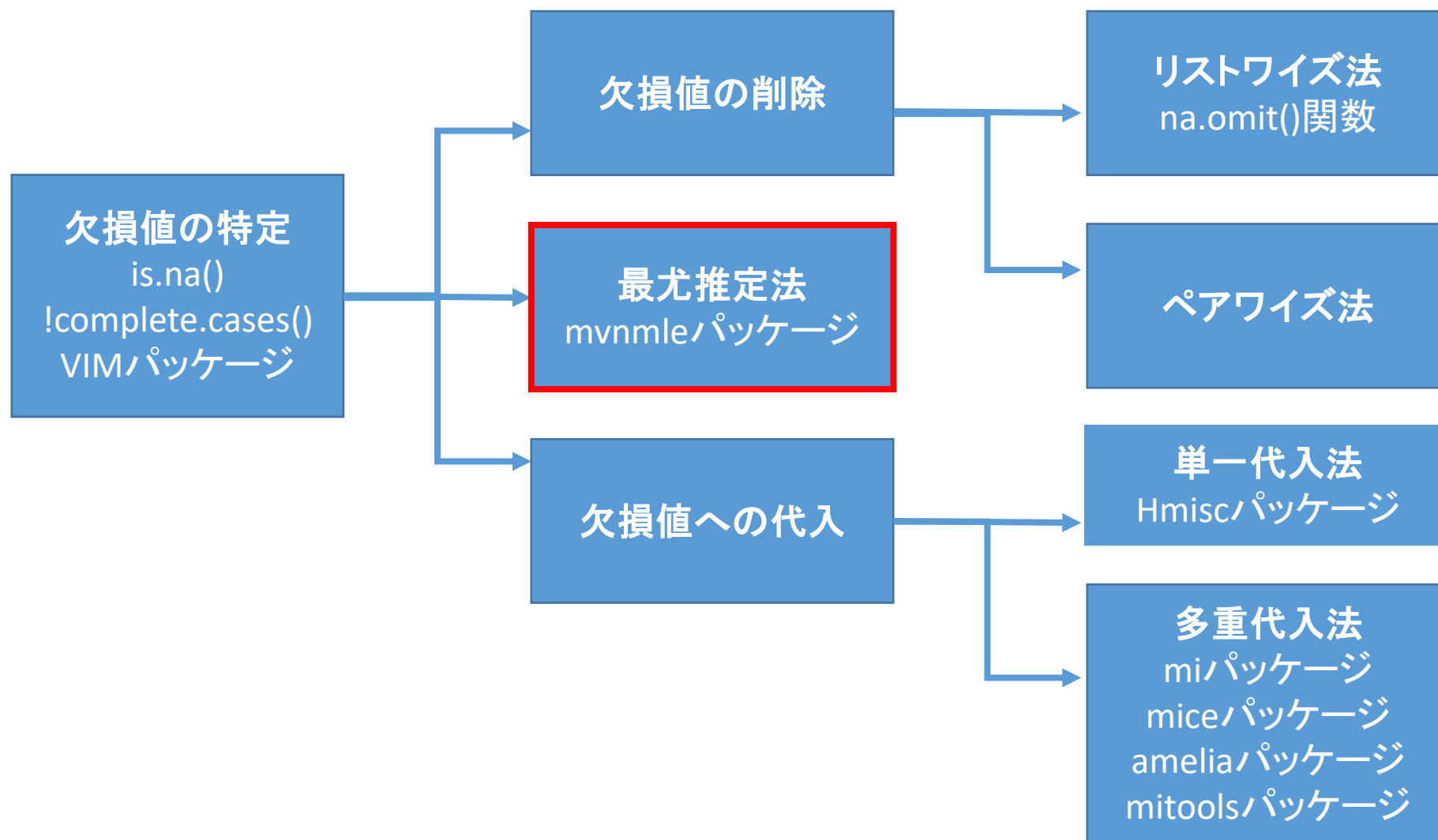
この本において、miceパッケージを中心として、Rの実装を交えながら多重代入法について詳しく書いてある。

[https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

[FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

[FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

欠損値対応のフロー



完全情報最尤推定法

```
install.packages("mvnmle", quiet = TRUE, dependencies=T)  
library(mvnmle)
```

#データの読み込み

```
employee.IQ.JP <- read.csv("employee_IQ_JP.csv", row.names = NULL,  
  colClasses = c(rep("integer", 3), "factor", "integer", "factor",  
  "integer", "factor"))
```

#完全情報最尤推定法の実行

```
mle.emp <- mlest(subset(employee.IQ.JP, select = c(IQ, MAR)))
```

```
mle.emp
```

```
> mle.emp
$muhat
[1] 99.99989  9.84867

$sigmahat
      [,1]      [,2]
[1,] 189.60050 28.369839
[2,]  28.36984  8.617752

$value
[1] 162.0294

$gradient
[1] 1.517721e-07 -8.888394e-07  2.926208e-07 -2.586376e-06  1.449507e-06

$stop.code
[1] 1

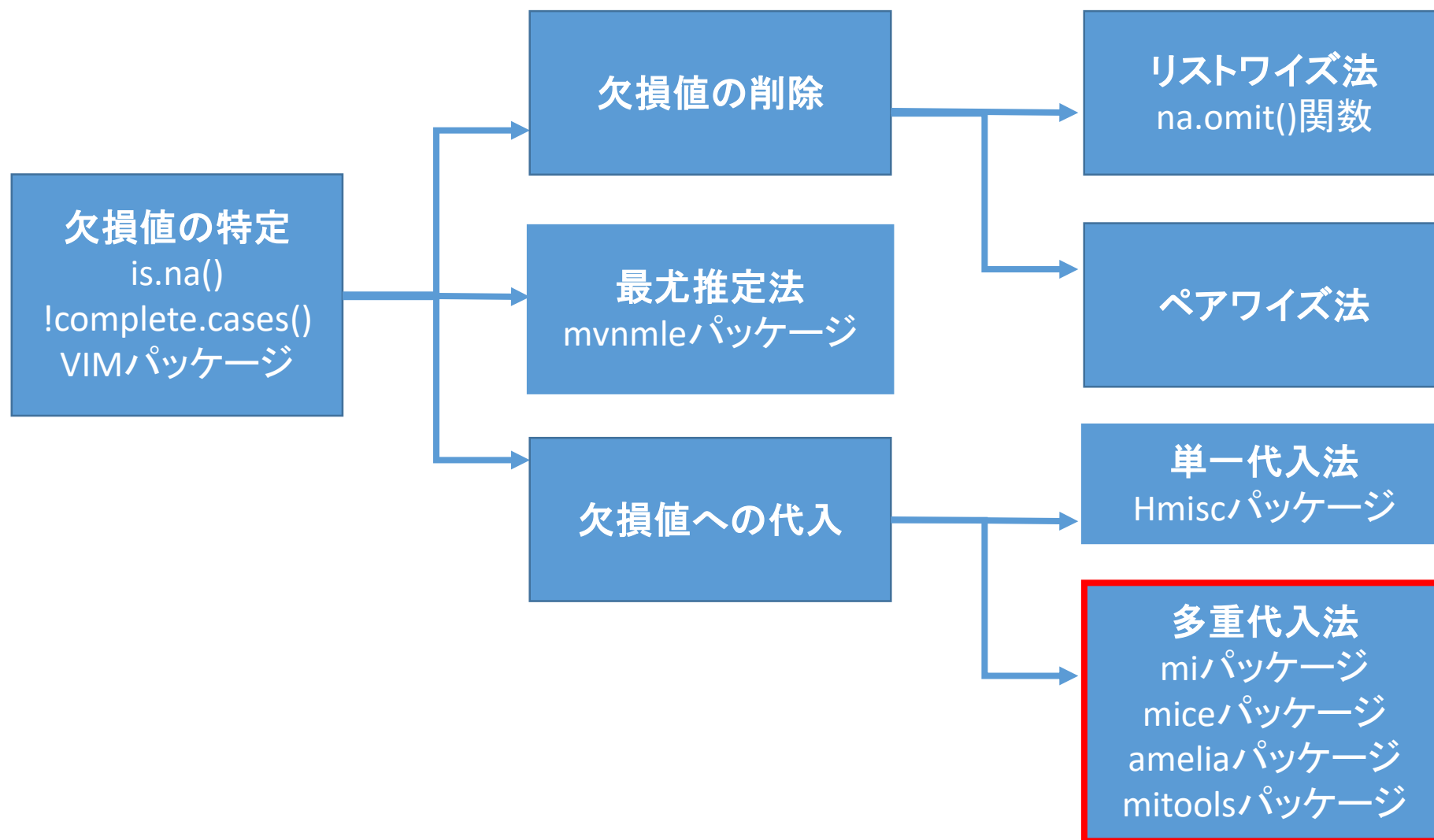
$iterations
[1] 36
```

#平均値の算出

```
mean(employee.IQ.JP$JobPerformance)
```

```
> mean(employee.IQ.JP$JobPerformance)
[1] 10.35
```

欠損値対応のフロー



削除または、代入方法

手法名	概要
リストワイズ法	欠損値を持つ行を削除
ペアワイズ法	相関係数や共分散等の算出の際2変数いずれかが欠損値をもつサンプルを削除
単一代入法	平均値や中央値など単一の値を欠損値へ代入。平均値を代入すれば平均代入法
回帰代入法	欠損値のないサンプルに回帰分析を行い、欠損値を含む項目の推定式を元に欠損値を補充。
確率的回帰代入法	回帰代入法により推定した値にランダムに誤差を与えて、欠損値を補充。
完全情報最尤推定法	サンプルごとに、欠損値パターンに応じた、尤度関数を仮定して、最尤推定を実施して、得られる多変量正規分布を用いて、平均値や分散共分散行列を推定。
多重代入法(※5)	欠損値に代入した、データセットを複数作成し、各データセットに対して、分析を実施し、その結果を統合することにより、欠損値を補充。

※5: S.v.Buuren. Flexible imputation of missing data. Chapman and Hall/CRC, 2012.

この本において、miceパッケージを中心として、Rの実装を交えながら多重代入法について詳しく書いてある。

[https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

[FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

[FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false](https://books.google.co.jp/books?hl=en&lr=&id=M89TDSml-FoC&oi=fnd&pg=PP1&dq=S.v.Buuren.+Flexible+imputation+of+missing+data.&ots=BbxNjnQudg&sig=6wbZv6FL_M4g-xk8xbJ2V9F02gw#v=onepage&q&f=false)

多重代入法

```
employee.IQ.JP <- read.csv("employee_IQ_JP.csv", row.names = NULL,  
  colClasses = c(rep("integer", 3), "factor", "integer", "factor", "integer",  
    "factor"))
```

#欠損値を補完したデータセットの作成

```
imp <- mice(subset(employee.IQ.JP, select=c(IQ, MAR)), seed=123,  
print=FALSE)
```

#分析

```
fit <- with(imp, lm(MAR ~ IQ))
```

#分析結果の統合

```
pool.fit <- pool(fit)
```

#統合した分析結果の要約の表示

```
sum.pf <- summary(pool.fit)
```

```
tab <- round(sum.pf, 3)
```

```
tab[, c(1:3, 5)]
```

```
> tab[, c(1:3, 5)]
```

	est	se	t	Pr(> t)
(Intercept)	-3.005	3.541	-0.848	0.411
IQ	0.131	0.035	3.771	0.002

#欠損値の補完

```
slope <- pool.fit$qbar[2]
```

```
intercept <- pool.fit$qbar[1]
```

```
imputed <- (slope * employee.IQ.JP$IQ + intercept)
```

```
employee.IQ.JP$MAR.mi <- employee.IQ.JP$MAR
```

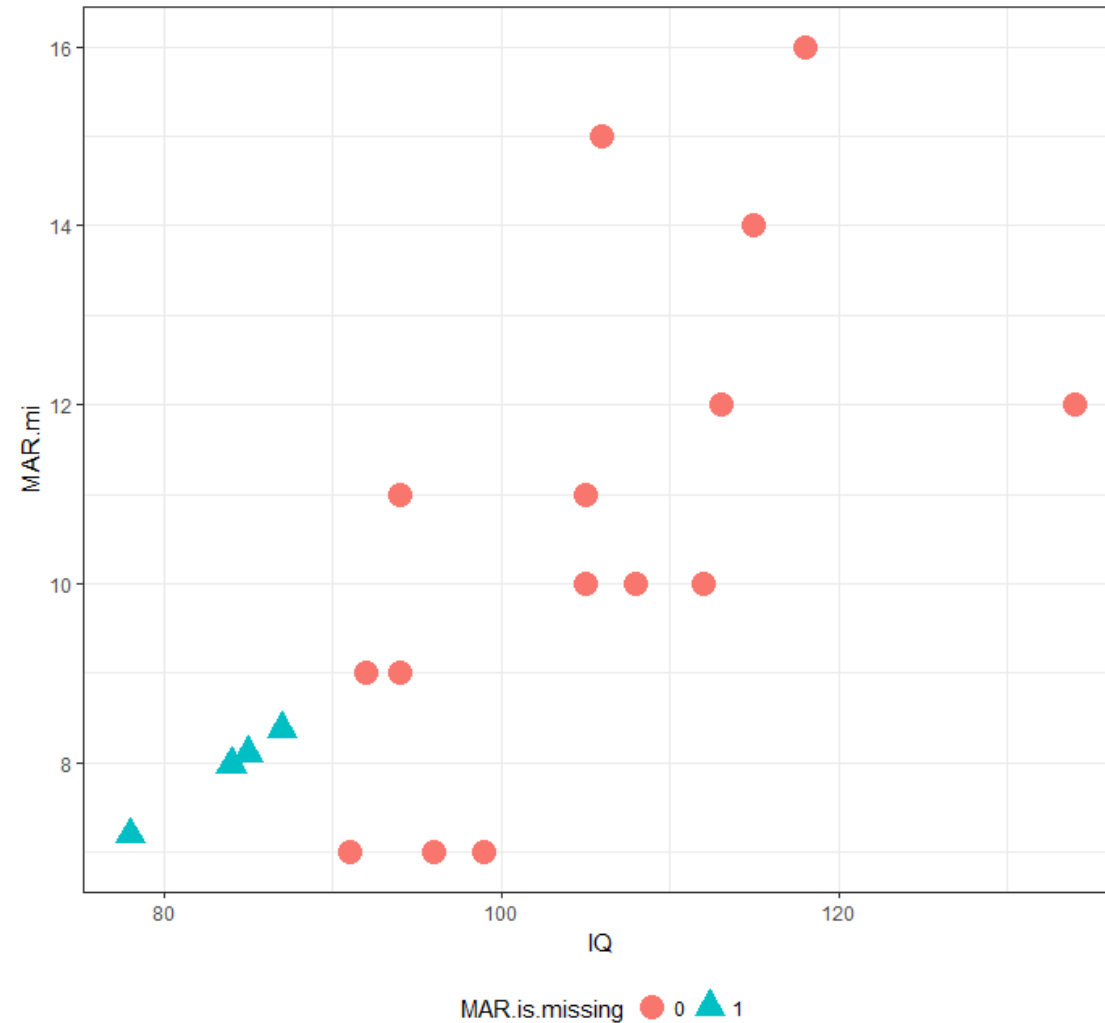
```
is.missing <- is.na(employee.IQ.JP$MAR.mi)
```

```
employee.IQ.JP$MAR.mi[is.missing] <- imputed[is.missing]
```

補完したデータを散布図としてプロット

```
p <- ggplot(data = employee.IQ.JP, aes(x = IQ, y = MAR.mi)) +  
  geom_point(aes(colour = MAR.is.missing, group = MAR.is.missing,  
    shape = MAR.is.missing), size = 5) + theme_bw() %+replace%  
  theme(legend.position = "bottom")  
print(p)
```

多重代入法で補完した結果のplot



#補完したデータによる平均値

```
mean(employee.IQ.JP$MAR.mi)
```

#元データの平均値

```
mean(employee.IQ.JP$JobPerformance)
```

```
> mean(employee.IQ.JP$MAR.mi)
```

```
[1] 9.981429
```

```
> mean(employee.IQ.JP$JobPerformance)
```

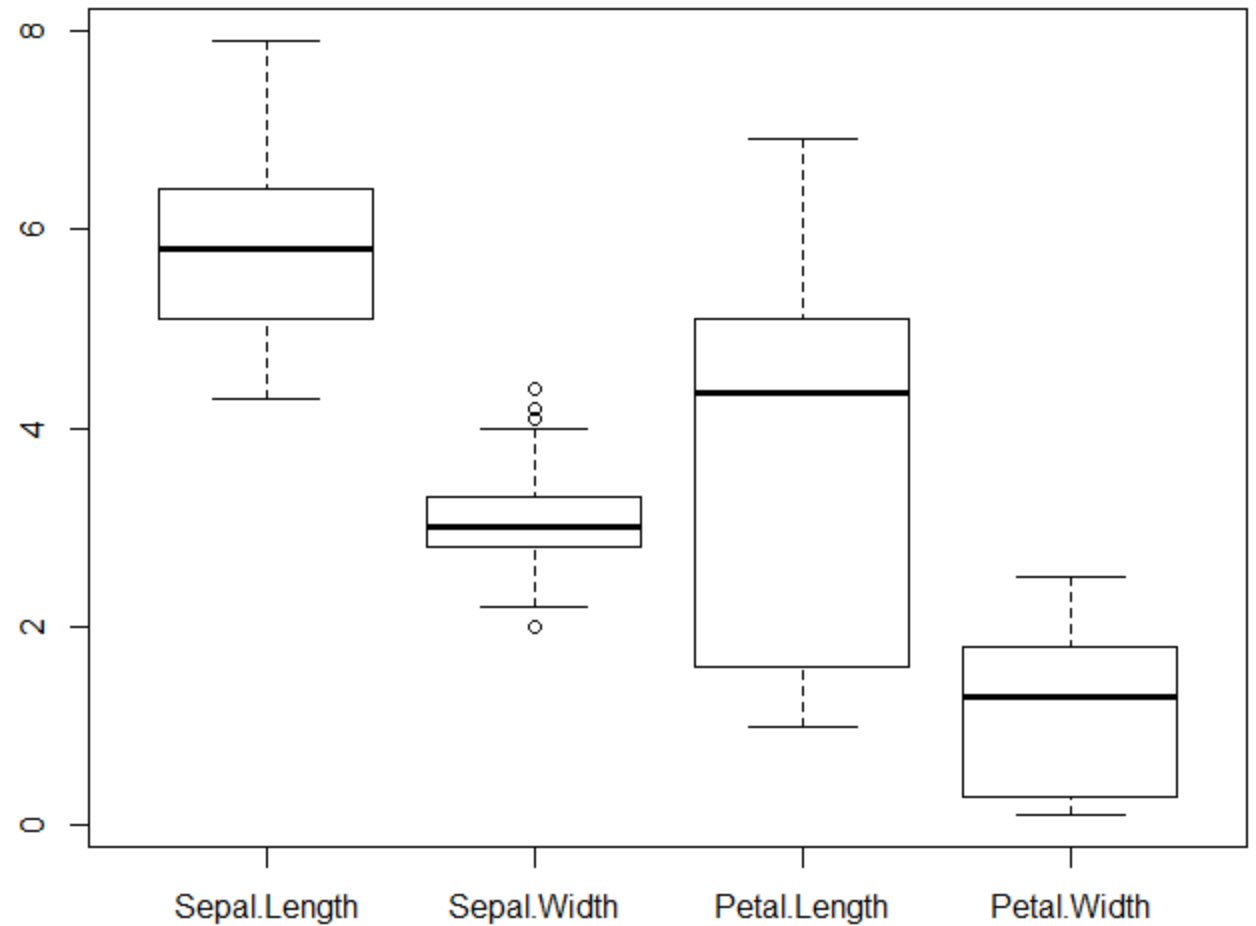
```
[1] 10.35
```

外れ値について


```
bp.iris <- boxplot(subset(iris, select = -Species))
```

```
bp.iris$out
```

```
> bp.iris$out  
[1] 4.4 4.1 4.2 2.0
```



箱ひげ図の見方

