chap4_5output

```
### Multivariate Imputation by Chained Equations
## The 'mice' package in R
## Based on  van Buuren and Groothuis-Oudshoorn (2011)
## https://www.jstatsoft.org/article/view/v045i03/v45i03.pdf


#install.packages("mice")
library(mice)
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
# toy data from the National Health and Nutrition Examination Survey
# Age group (1=20-39, 2=40-59, 3=60+)
# bmi Body mass index (kg/m**2)
# Hypertensive (1=no,2=yes)
# chl Total serum cholesterol (mg/dL)
data(nhanes)
dim(nhanes)
```

```
## [1] 25  4
```

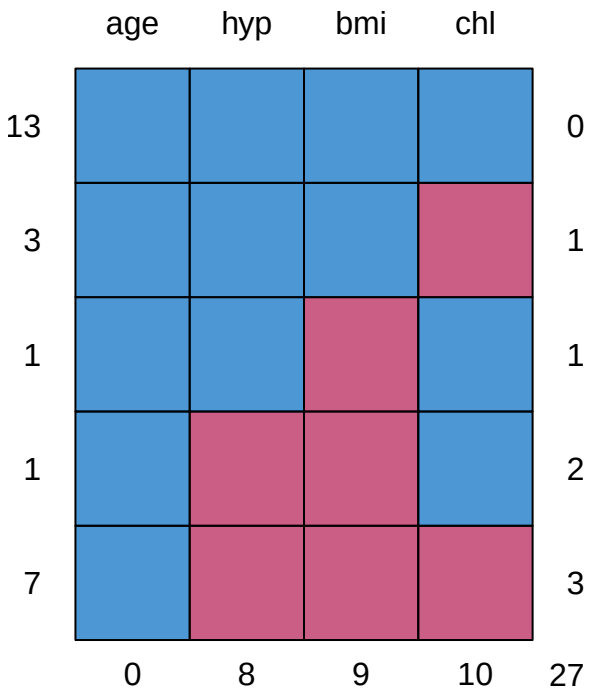
```
nhanes
```

```
##    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
## 11   1   NA  NA  NA
## 12   2   NA  NA  NA
## 13   3 21.7   1 206
## 14   2 28.7   2 204
## 15   1 29.6   1  NA
## 16   1   NA  NA  NA
## 17   3 27.2   2 284
## 18   2 26.3   2 199
## 19   1 35.3   1 218
## 20   3 25.5   2  NA
```

```
## 21   1   NA  NA  NA
## 22   1 33.2   1 229
## 23   1 27.5   1 131
## 24   3 24.9   1  NA
## 25   2 27.4   1 186
```

```
nhanes$age <- as.factor(nhanes$age)
nhanes$hyp <- as.factor(nhanes$hyp)

# Summarize response patterns
md.pattern(nhanes)
```




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

```
#   1st column: number of rows with that response patt.
#   Last colmn: number of missing entries in that patt.
#   Last row: number of rows with missingness in that col.


# Creating imputations
tempdata <- mice(nhanes,
            m = 5, # default number of imputations
            method = c("","norm","logreg","norm"),
            maxit = 6, # iterations per run
            printFlag=TRUE,  #change to FALSE for silent computation
            seed = 23109)
```

```
##
```

```
##  iter imp variable
##   1   1  bmi  hyp  chl
##   1   2  bmi  hyp  chl
##   1   3  bmi  hyp  chl
##   1   4  bmi  hyp  chl
##   1   5  bmi  hyp  chl
##   2   1  bmi  hyp  chl
##   2   2  bmi  hyp  chl
##   2   3  bmi  hyp  chl
##   2   4  bmi  hyp  chl
##   2   5  bmi  hyp  chl
##   3   1  bmi  hyp  chl
##   3   2  bmi  hyp  chl
##   3   3  bmi  hyp  chl
##   3   4  bmi  hyp  chl
##   3   5  bmi  hyp  chl
##   4   1  bmi  hyp  chl
##   4   2  bmi  hyp  chl
##   4   3  bmi  hyp  chl
##   4   4  bmi  hyp  chl
##   4   5  bmi  hyp  chl
##   5   1  bmi  hyp  chl
##   5   2  bmi  hyp  chl
##   5   3  bmi  hyp  chl
##   5   4  bmi  hyp  chl
##   5   5  bmi  hyp  chl
##   6   1  bmi  hyp  chl
##   6   2  bmi  hyp  chl
##   6   3  bmi  hyp  chl
##   6   4  bmi  hyp  chl
##   6   5  bmi  hyp  chl
```

```
# method(s): none (""), normal linear model ("norm"), logistic regression ("logreg")
#   other methods are available, such as polyreg, pmm: predictive mean matching, mean etc

class(tempdata) # mids: multiple imputed dataset
```

```
## [1] "mids"
```

```
summary(tempdata)
```

```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##      age      bmi      hyp      chl
##       ""   "norm" "logreg"   "norm"
## PredictorMatrix:
##     age bmi hyp chl
## age   0   1   1   1
## bmi   1   0   1   1
## hyp   1   1   0   1
## chl   1   1   1   0
```

```
# PredictorMatrix: rows correspond to incomplete target variables.
#   Ones indicate the column variables used as a predictors
#   to impute the row variable.  Can be changed with
```

```
#   'predictorMatrix' argument in mice()


# Imputations are stored on a per-variable basis
#   rows: observations with missingness in that column
#   columns: separate imputations
tempdata$imp  #see all imputed values
```

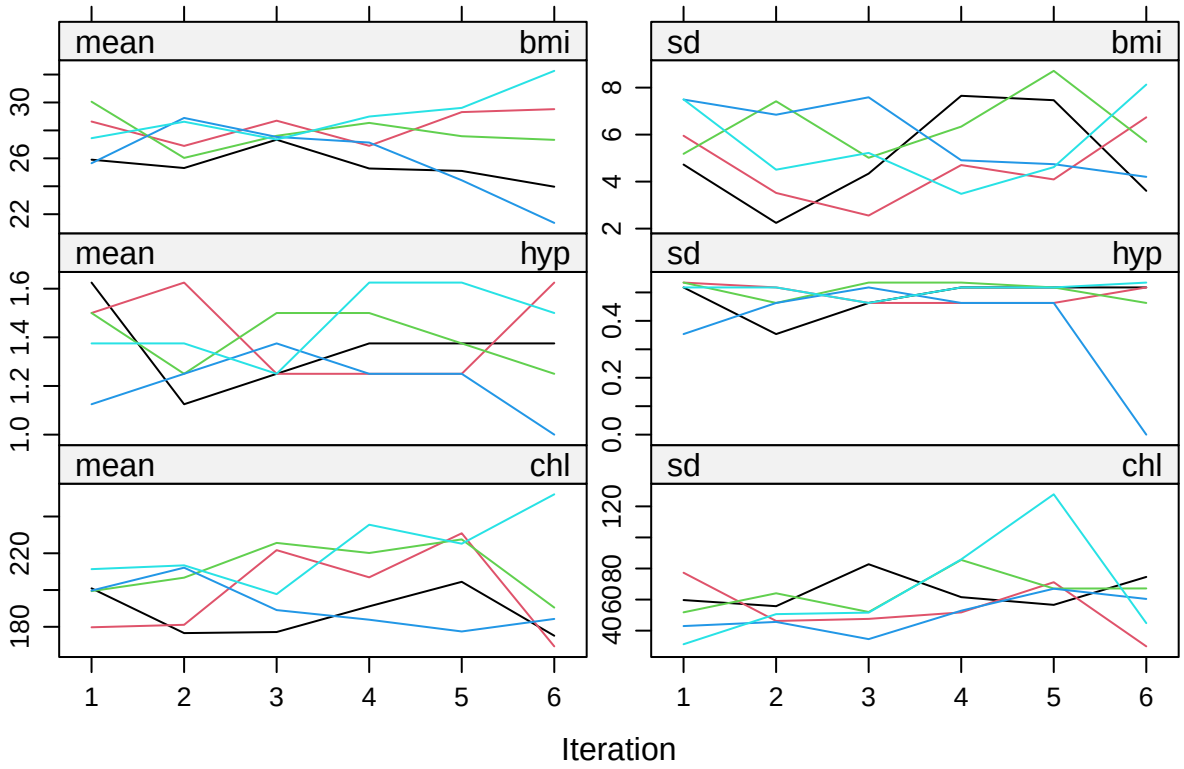
```
## $age
## [1] 1 2 3 4 5
## <0 rows> (or 0-length row.names)
##
## $bmi
##           1        2        3        4        5
## 1  22.68719 28.38807 25.82014 23.08041 32.15747
## 3  24.47157 33.79057 27.25442 23.21025 33.15505
## 4  25.50336 29.03205 21.44568 19.19416 26.09024
## 6  18.87292 24.56495 17.71806 17.38981 15.70918
## 10 27.57300 15.00394 27.01723 13.89974 32.24690
## 11 24.71649 37.33227 32.21611 19.62590 44.01054
## 12 22.10806 33.22510 34.95862 24.18204 35.32510
## 16 19.64369 29.10144 34.00004 27.71064 31.21618
## 21 30.14468 35.25720 25.48149 24.09605 40.45736
##
## $hyp
##    1 2 3 4 5
## 1  1 1 1 1 1
## 4  2 2 2 1 1
## 6  2 2 1 1 1
## 10 1 1 2 1 2
## 11 1 2 1 1 2
## 12 2 2 1 1 2
## 16 1 1 1 1 1
## 21 1 2 1 1 2
##
## $chl
##            1        2         3         4        5
## 1  109.31981 162.1278 132.81780 156.43813 217.9387
## 4  228.07009 168.6656 135.70714 248.26871 275.8466
## 10 261.66615 164.7256 205.93946 137.37889 206.2159
## 11 122.82630 216.9734 237.38655  83.84746 266.9129
## 12 176.84679 175.4796 255.73956 213.09277 256.2657
## 15 141.60201 128.0421 164.93018 189.25531 230.1992
## 16  49.28745 128.5351 251.14292 172.43865 177.0715
## 20 215.36597 150.7891 253.45892 261.91859 311.1217
## 21 154.41365 208.1451  53.60677 124.54127 261.0926
## 24 292.05923 190.1568 213.96484 255.62991 317.9877
```

```
tempdata$imp$bmi  #see imputed values for bmi
```

```
##           1        2        3        4        5
## 1  22.68719 28.38807 25.82014 23.08041 32.15747
## 3  24.47157 33.79057 27.25442 23.21025 33.15505
## 4  25.50336 29.03205 21.44568 19.19416 26.09024
```

```
## 6  18.87292 24.56495 17.71806 17.38981 15.70918
## 10 27.57300 15.00394 27.01723 13.89974 32.24690
## 11 24.71649 37.33227 32.21611 19.62590 44.01054
## 12 22.10806 33.22510 34.95862 24.18204 35.32510
## 16 19.64369 29.10144 34.00004 27.71064 31.21618
## 21 30.14468 35.25720 25.48149 24.09605 40.45736
```

```
# Monitor convergence of the pseudo-Gibbs sampler
plot(tempdata)
```




```
# Obtain the different completed datasets
# First one:
complete(tempdata,1)
```

```
##    age      bmi hyp       chl
## 1    1 22.68719   1 109.31981
## 2    2 22.70000   1 187.00000
## 3    1 24.47157   1 187.00000
## 4    3 25.50336   2 228.07009
## 5    1 20.40000   1 113.00000
## 6    3 18.87292   2 184.00000
## 7    1 22.50000   1 118.00000
## 8    1 30.10000   1 187.00000
## 9    2 22.00000   1 238.00000
## 10   2 27.57300   1 261.66615
## 11   1 24.71649   1 122.82630
## 12   2 22.10806   2 176.84679
## 13   3 21.70000   1 206.00000
## 14   2 28.70000   2 204.00000
## 15   1 29.60000   1 141.60201
## 16   1 19.64369   1  49.28745
```

```
## 17   3 27.20000   2 284.00000
## 18   2 26.30000   2 199.00000
## 19   1 35.30000   1 218.00000
## 20   3 25.50000   2 215.36597
## 21   1 30.14468   1 154.41365
## 22   1 33.20000   1 229.00000
## 23   1 27.50000   1 131.00000
## 24   3 24.90000   1 292.05923
## 25   2 27.40000   1 186.00000
```

```
# Third one:
complete(tempdata,3)
```

```
##    age      bmi hyp       chl
## 1    1 25.82014   1 132.81780
## 2    2 22.70000   1 187.00000
## 3    1 27.25442   1 187.00000
## 4    3 21.44568   2 135.70714
## 5    1 20.40000   1 113.00000
## 6    3 17.71806   1 184.00000
## 7    1 22.50000   1 118.00000
## 8    1 30.10000   1 187.00000
## 9    2 22.00000   1 238.00000
## 10   2 27.01723   2 205.93946
## 11   1 32.21611   1 237.38655
## 12   2 34.95862   1 255.73956
## 13   3 21.70000   1 206.00000
## 14   2 28.70000   2 204.00000
## 15   1 29.60000   1 164.93018
## 16   1 34.00004   1 251.14292
## 17   3 27.20000   2 284.00000
## 18   2 26.30000   2 199.00000
## 19   1 35.30000   1 218.00000
## 20   3 25.50000   2 253.45892
## 21   1 25.48149   1  53.60677
## 22   1 33.20000   1 229.00000
## 23   1 27.50000   1 131.00000
## 24   3 24.90000   1 213.96484
## 25   2 27.40000   1 186.00000
```

```
# How to analyze your multiple imputed datasets?

# 'with' each completed dataset, we want to run
#   a linear regression, for example
fit <- with(tempdata, lm(chl ~ age + bmi))

class(fit) # mira: multiple imputed repeated analysis
```

```
## [1] "mira"   "matrix"
```

```
# The results of the different analyses
print(fit)
```

```
## call :
## with.mids(data = tempdata, expr = lm(chl ~ age + bmi))
##
```

```
## call1 :
## mice(data = nhanes, m = 5, method = c("", "norm", "logreg", "norm"),
##     maxit = 6, printFlag = TRUE, seed = 23109)
##
## nmis :
## age bmi hyp chl
##   0   9   8  10
##
## analyses :
## [[1]]
##
## Call:
## lm(formula = chl ~ age + bmi)
##
## Coefficients:
## (Intercept)         age2         age3          bmi
##     -66.349       72.247      110.106        7.983
##
##
## [[2]]
##
## Call:
## lm(formula = chl ~ age + bmi)
##
## Coefficients:
## (Intercept)         age2         age3          bmi
##      40.392       46.504       48.460        4.254
##
##
## [[3]]
##
## Call:
## lm(formula = chl ~ age + bmi)
##
## Coefficients:
## (Intercept)         age2         age3          bmi
##     -82.207       56.291       92.809        8.764
##
##
## [[4]]
##
## Call:
## lm(formula = chl ~ age + bmi)
##
## Coefficients:
## (Intercept)         age2         age3          bmi
##     -14.104       53.954      105.243        6.572
##
##
## [[5]]
##
## Call:
## lm(formula = chl ~ age + bmi)
##
```

```
## Coefficients:
## (Intercept)         age2         age3          bmi
##     -16.522       41.765      122.667        6.677
```

```
# combine them using Rubin's rules
pool_fit <- pool(fit)

class(pool_fit) # mipo: multiple imputed pooled outcomes
```

```
## [1] "mipo"       "data.frame"
```

```
summary(pool_fit)
```

```
##          term   estimate std.error  statistic       df    p.value
## 1 (Intercept) -27.758035 67.991128 -0.4082597 4.411040 0.70213366
## 2        age2  54.152269 20.194075  2.6815919 7.896481 0.02818315
## 3        age3  95.856967 35.777260  2.6792708 2.718487 0.08346651
## 4         bmi   6.849997  2.373524  2.8860027 4.198274 0.04222933
```

```
# Issues not covered here:
#   - Functions of variables: interactions, polinomial terms
#   - Selecting predictors for each model
# See van Buuren and Groothuis-Oudshoorn (2011) for details.

# Regarding the unrelated conditional models used in MICE:
# "It is not yet clear what the consequences of incompatibility
#    are on the quality of the imputations" (van Buuren and Groothuis-Oudshoorn, 2011)

# Other MICE-related references
#   - The 'mi' R package: http://www.stat.columbia.edu/~gelman/research/published/mipaper.pdf
#   - Raghunathan et al (2001):

#https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001001/article/5857-eng.pdf?st=6HKI46qF
```