Overview of how to deal with missing data (ft. multiple imputation mice)

11.07.2025

Glötzl Alexander
FAKULTÄT FÜR PHYSIK

Universität Regensburg

Missing values:

airquality data set in R:

```
> 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
13    11     290  9.2   66     5  13
14    14     274 10.9   68     5  14
15    18      65 13.2   58     5  15
16    14     334 11.5   64     5  16
```

MCAR:

missing **completely** at random

MAR:

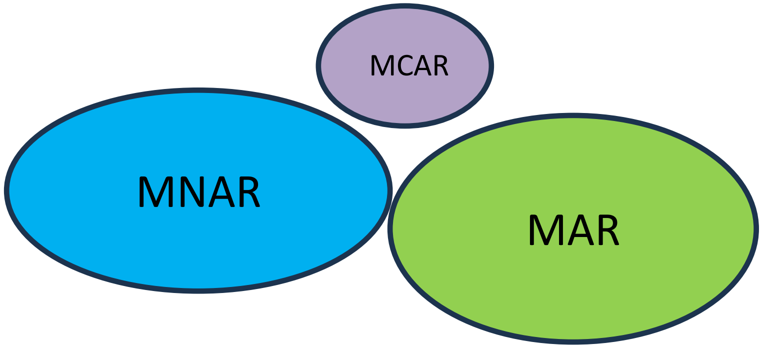
missing at random

- missingness depends on **observed** features, not unobserved

MNAR:

missing **not** at random

- we don't know why some data are missing



- Listwise deletion:

```
> 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
13    11     290  9.2   66     5  13
14    14     274 10.9   68     5  14
15    18      65 13.2   58     5  15
16    14     334 11.5   64     5  16
```

advantages:
- convenient
- unbiased mean, variance & regression weights (only under MCAR)

disadvantages:
- increasingly wasteful with more and more features
- 'time' as feature makes pattern recognition harder

```
fit <- lm(Ozone ~ Wind, data = airquality, na.action = na.omit)
```

- Pairwise deletion:

```
> 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
13    11     290  9.2   66     5  13
14    14     274 10.9   68     5  14
15    18      65 13.2   58     5  15
16    14     334 11.5   64     5  16
```

advantages:

- unbiased mean & variance (only under MCAR)
- is less wasteful

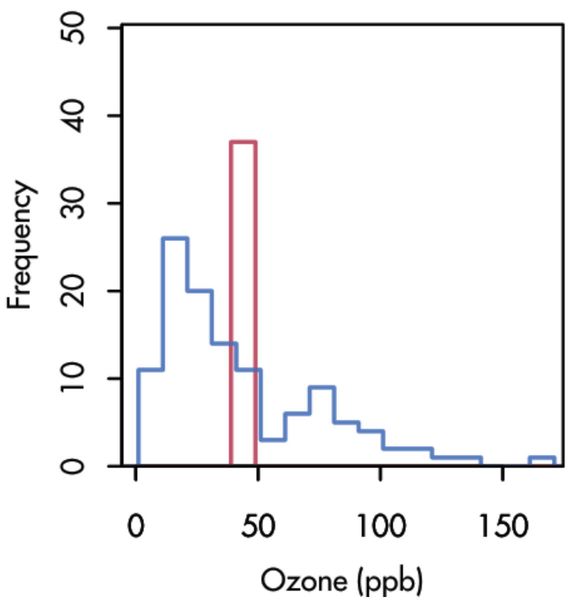
disadvantages:

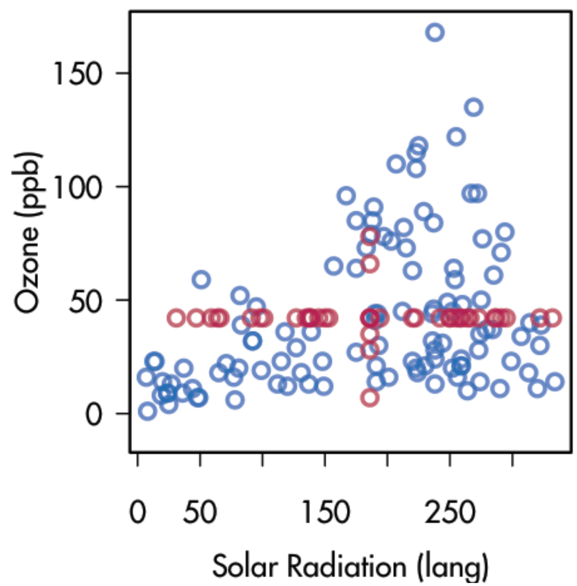
- calculation of standard error is not clear
- problems for highly correlated features
- only works with normally distr. numerical data (not with different types of data)

```
data <- airquality[, c("Ozone", "Solar.R", "Wind")]
cv <- cov(data, use = "pairwise")
library(lavaan)
fit <- lavaan("Ozone ~ 1 + Wind + Solar.R
               Ozone ~~ Ozone",
              sample.mean = mu, sample.cov = cv,
              sample.nobs = sum(complete.cases(data)))
```

- Mean imputation:

```
> colMeans(airquality, na.rm=TRUE)
     Ozone    Solar.R       Wind       Temp
 42.129310 185.931507   9.957516  77.882353
```





advantages:

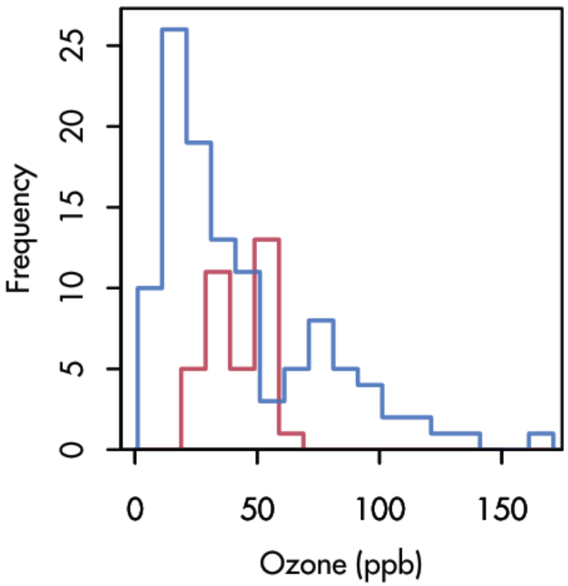
- fast & simple

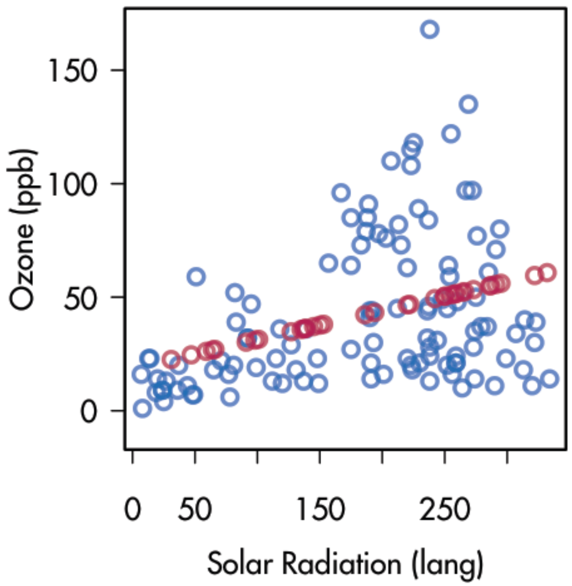
disadvantages:

- underestimates variance
- disturbs correlations
- bias towards any estimate except the mean

```
> colSums(is.na(airquality))
  Ozone Solar.R    Wind    Temp
     37       7       0       0
```

- (univariate) Regression imputation:





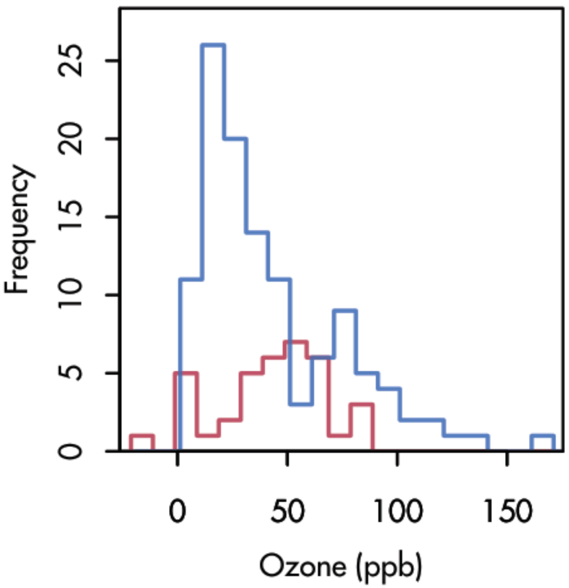
advantages:

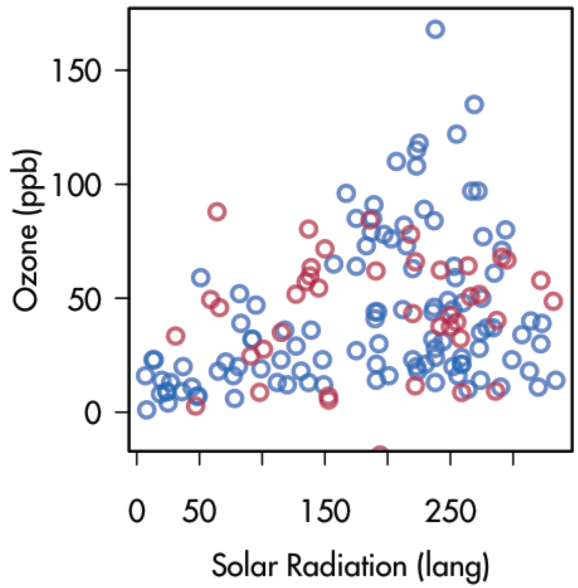
- unbiased mean & regression weights
- predicts the **most likely** value

disadvantages:

- underestimates variance
- overestimates correlations
- **no uncertainty** in imputation

- Stochastic regression imputation:





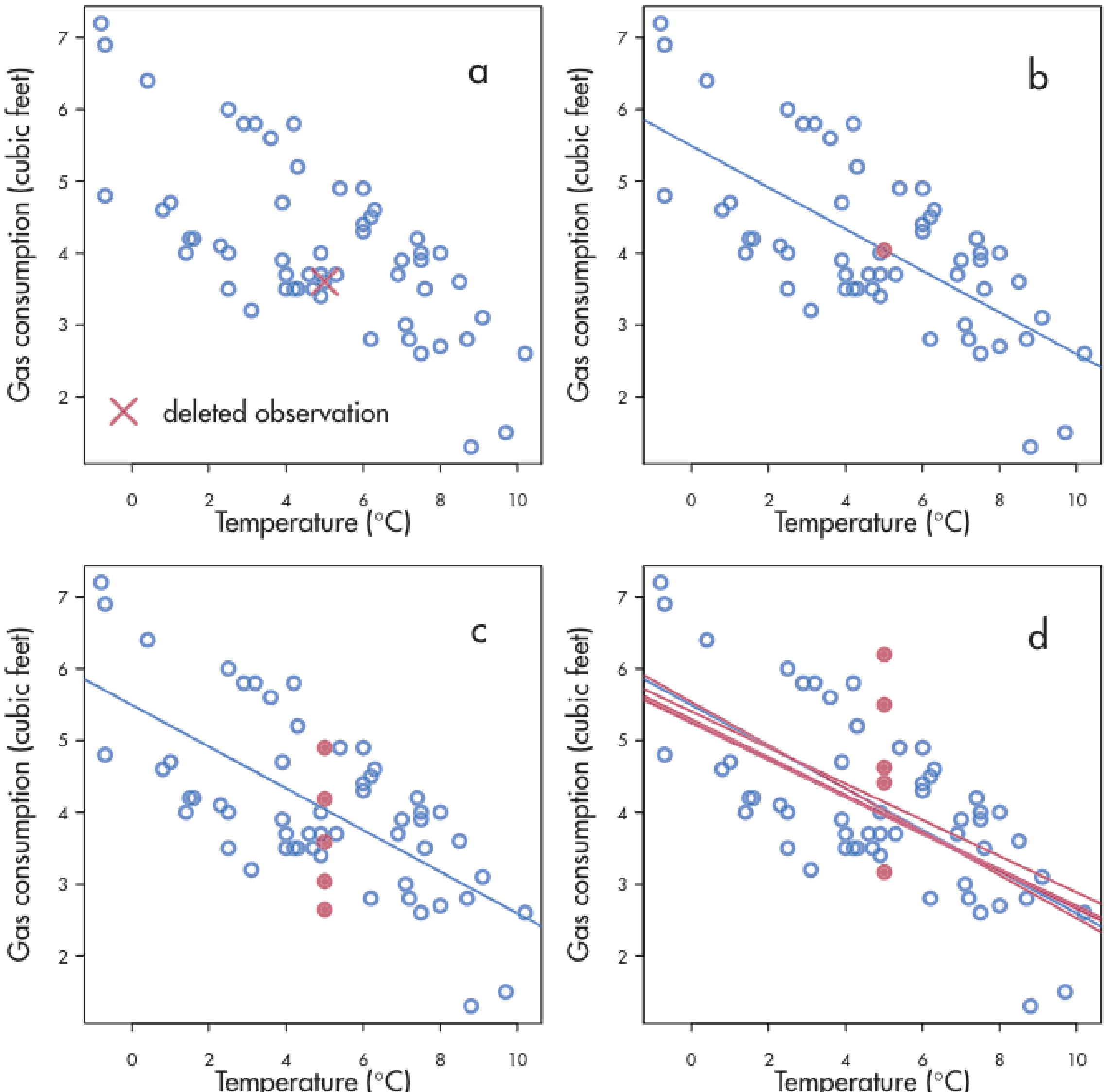
advantages:

- improves on previous <u>regression imputation</u> in regards to variance and correlation

disadvantages:

- there are **no negative** ozone levels
- regression line is not cone shaped

Imputation on univariate missing data:

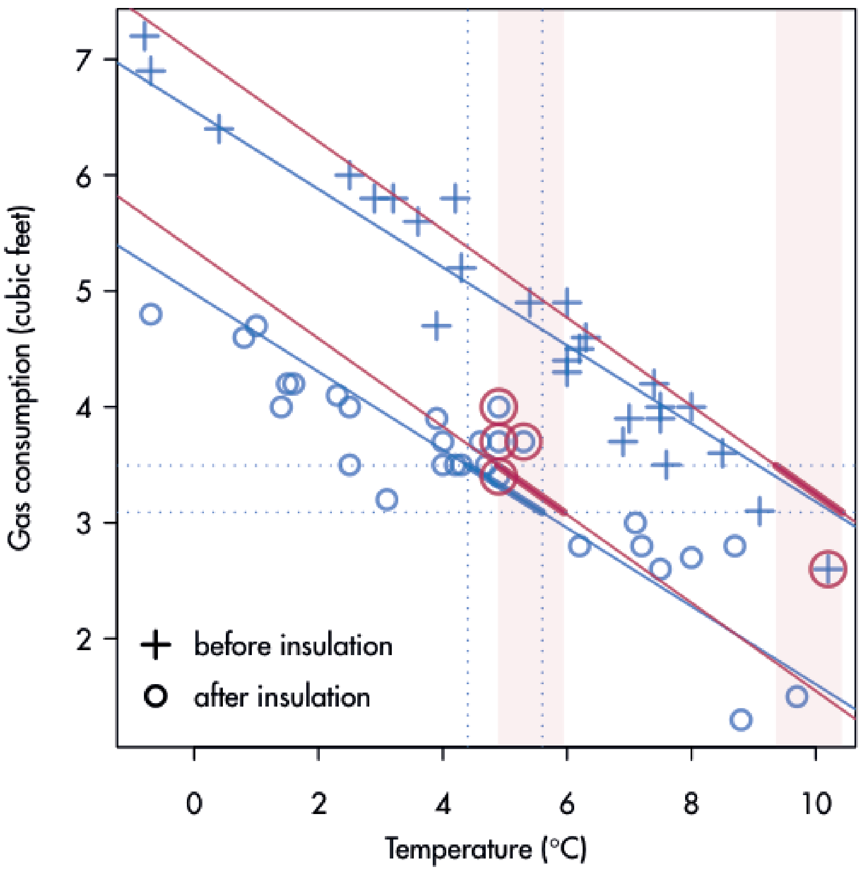


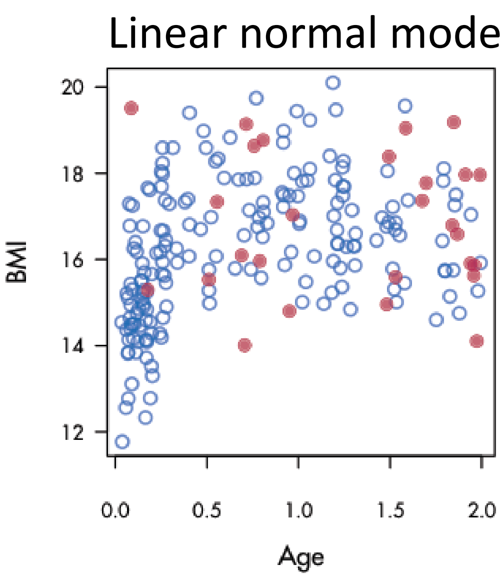
(a) no imputation
(b) regression imputation

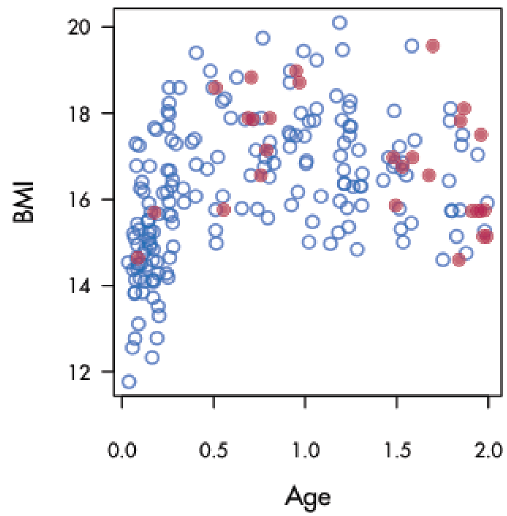
(c) stochastic regression imputation
(d) bayesian inference + noise

$$p(\theta \mid \mathbf{X}, \alpha) = \frac{p(\mathbf{X} \mid \theta, \alpha)p(\theta \mid \alpha)}{p(\mathbf{X} \mid \alpha)}$$

- predictive mean matching







advantages:

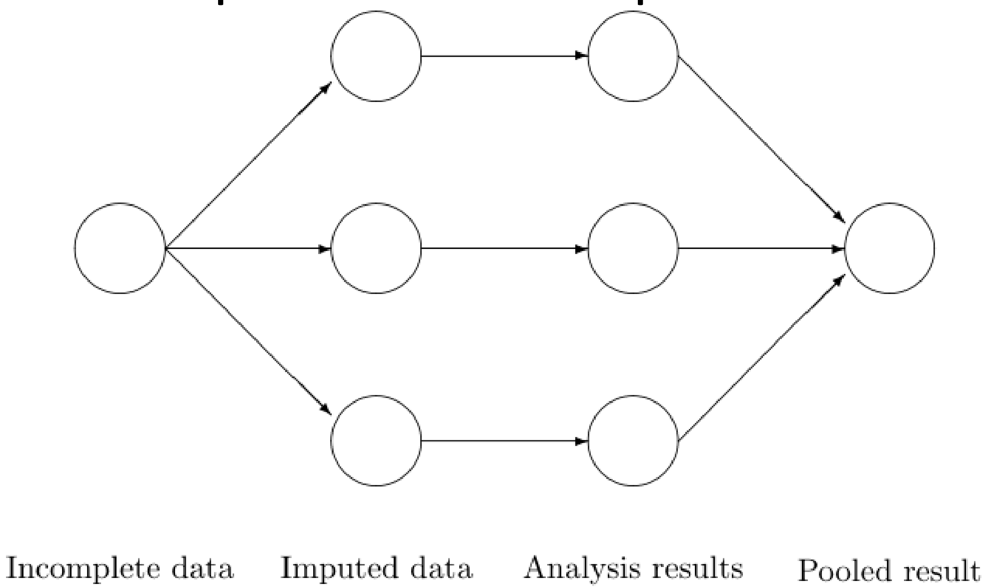
- samples real values

disadvantages:

- if sample size small, duplication of same donor may occur many times

- Multiple Imputation by Chained Equations (MICE)

example with m=3 imputations



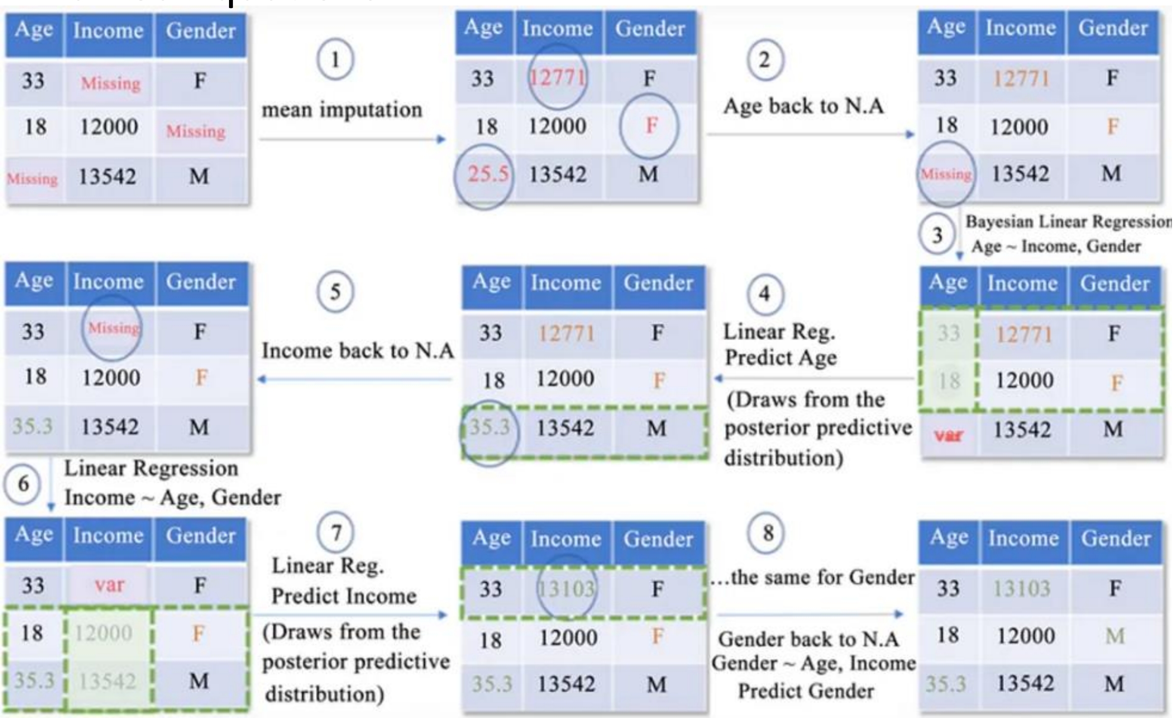
advantages:
- works under MAR
- preserves uncertainty of imputations
- number of imputations m = 5 or 10 mostly sufficient

disadvantages:
- not super fast with increasing imputations m

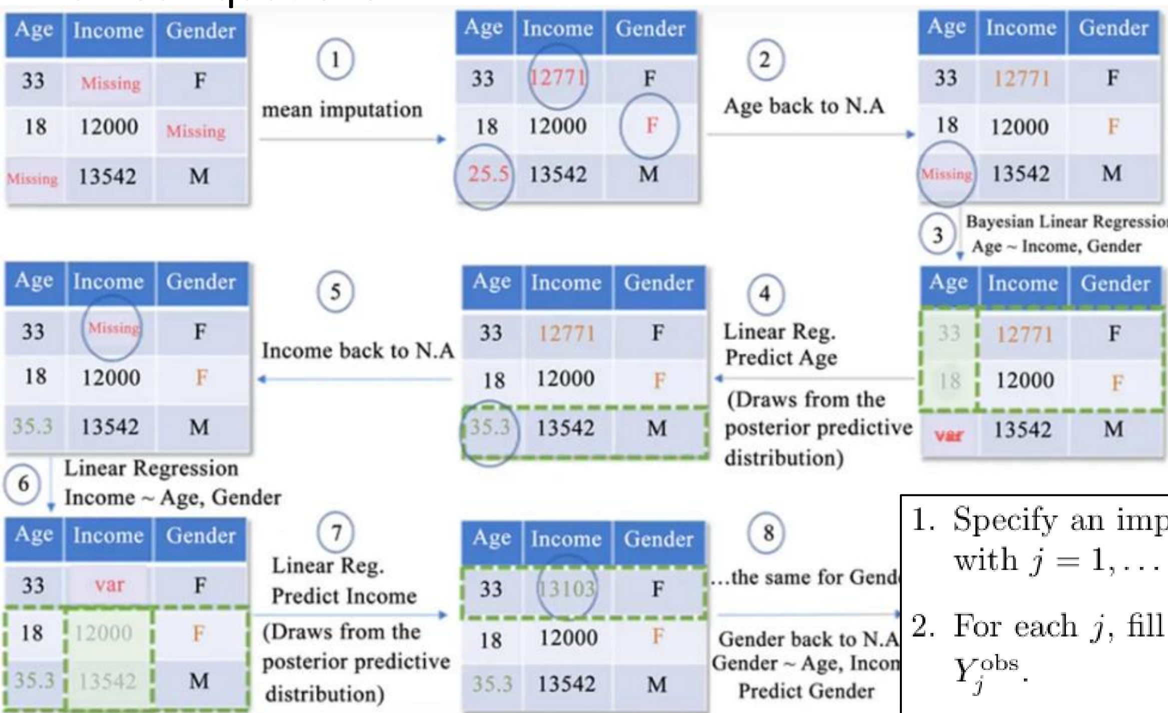
```
imp <- mice(airquality, seed = 1, m = 20, print = FALSE)
fit <- with(imp, lm(Ozone ~ Wind + Temp + Solar.R))
summary(pool(fit))

             estimate std.error statistic    df        p.value
(Intercept) -62.7055   21.1973     -2.96 106.3 0.003755025718
Wind         -3.0839    0.6281     -4.91  91.7 0.000003024665
Temp          1.5988    0.2311      6.92 115.4 0.000000000271
Solar.R       0.0573    0.0217      2.64 112.8 0.009489765888
```

Chained Equations:



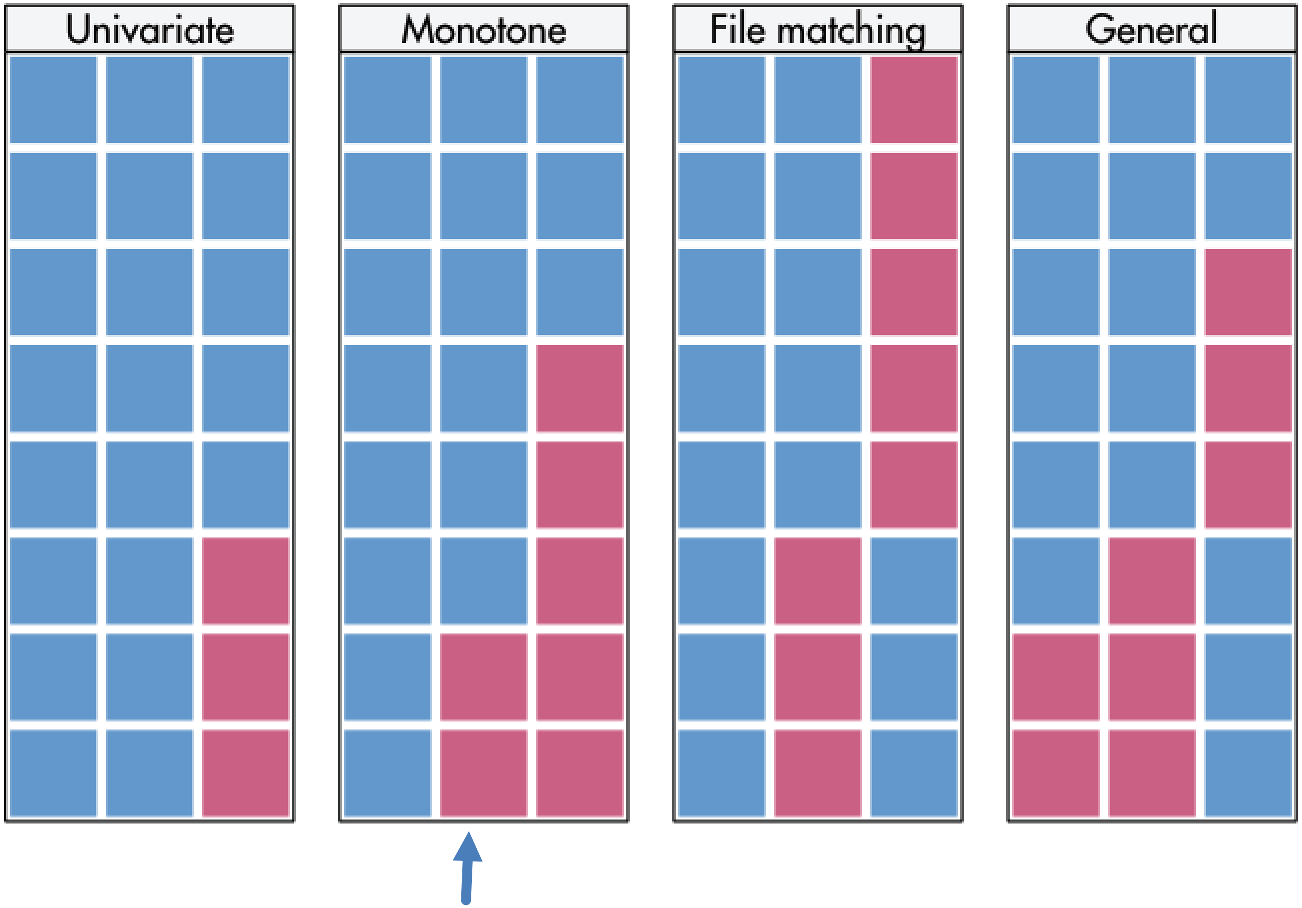
Chained Equations:



MICE algorithm

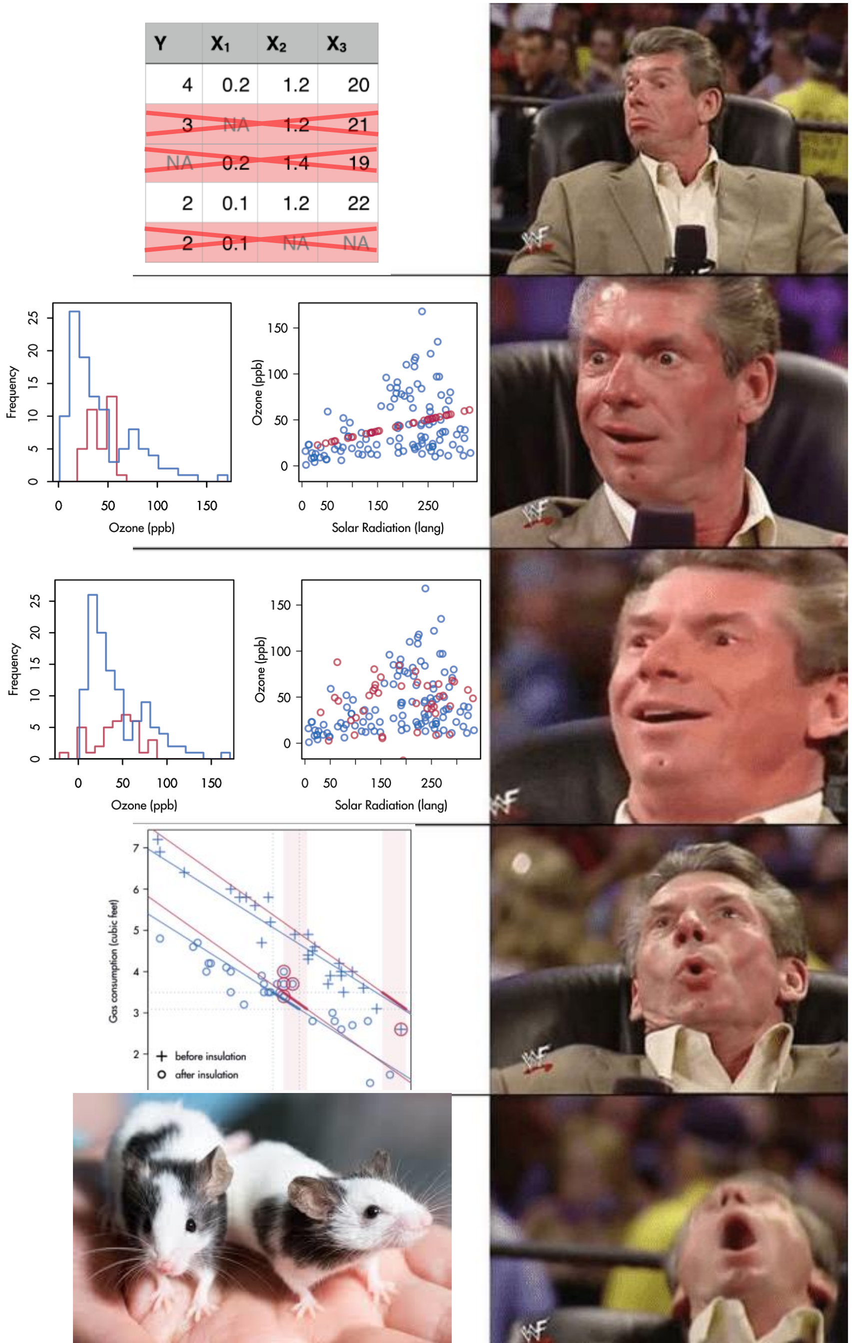
1. Specify an imputation model $P(Y_j^{\mathrm{mis}} | Y_j^{\mathrm{obs}}, Y_{-j}, R)$ for variable Y_j with $j = 1, \dots, p$.
2. For each j, fill in starting imputations $\dot{Y}_j^0$ by random draws from Y_j^{obs}.
3. Repeat for $t = 1, \dots, M$.
4. Repeat for $j = 1, \dots, p$.
5. Define $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \dots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \dots, \dot{Y}_p^{t-1})$ as the currently complete data except Y_j.
6. Draw $\dot{\phi}_j^t \sim P(\phi_j^t | Y_j^{\mathrm{obs}}, \dot{Y}_{-j}^t, R)$.
7. Draw imputations $\dot{Y}_j^t \sim P(Y_j^{\mathrm{mis}} | Y_j^{\mathrm{obs}}, \dot{Y}_{-j}^t, R, \dot{\phi}_j^t)$.
8. End repeat j.
9. End repeat t.

- Connectivity in multivariate data



algo: 'monotone data imputation'

```
imp <- mice(data, visit = "monotone", maxit = 1, m = 2,
            print = FALSE)
```



Y X1 X2 X3
4 0.2 1.2 20
3 NA 1.2 21
NA 0.2 1.4 19
2 0.1 1.2 22
2 0.1 NA NA
Frequency
Ozone (ppb)
Ozone (ppb)
Solar Radiation (lang)
Frequency
Ozone (ppb)
Ozone (ppb)
Solar Radiation (lang)
Gas consumption (cubic feet)
before insulation
after insulation

Thank you