Missing Data Imputation

Overview of General Issues and Solutions

Stef van Buuren, University of Utrecht & TNO

December 19, 2019; Building Multi-Source Databases for Comparative Analyses

Overview

- Missing data and harmonization
- Multiple imputation in a nutshell
- Alternatives for recoding
- Imputation of multilevel data

Why this course

- Missing data are everywhere
- Harmonization is an attempt to solve a missing data problem
- Ad-hoc fixes do not (always) work
- Multiple imputation is broadly applicable, yield correct statistical inferences
- Goal of the course: introduce `mice` as a way to think about data harmonization

Course materials

- URL to github site
- Materials: https://www.asc.ohio-state.edu/dataharmonization/wp-content/uploads/2019/12/Workshop-Missing-Data-Imputation-Materials-Kotnarowski-2019-FINAL.pdf

Reading materials

- Van Buuren, S. and Groothuis-Oudshoorn, C.G.M. (2011). `mice`: Multivariate Imputation by Chained Equations in `R`. Journal of Statistical Software, 45(3), 1–67. https://www.jstatsoft.org/article/view/v045i03
- Van Buuren, S. (2018). Flexible Imputation of Missing Data. Second Edition. Chapman & Hall/CRC, Boca Raton, FL. https://stefvanbuuren.name/fimd

Chapman & Hall/CRC
Interdisciplinary Statistics Series

Flexible Imputation of Missing Data

SECOND EDITION

Stef van Buuren

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Today's schedule

Slot	Time	What	Topic
A	10.00-11.30	L	Multiple imputation intro
	11.30-11.45		COFFEE/TEA
B	11.45-13:15	L	Imputation for harmonisation
	13.15-14.30		LUNCH
C	14.30-16.00	P	Lab session: Kotnarowski, IFiS
	13.15-14.30		COFFEE/TEA
D	16.15-17.30	P	Lab session: Kotnarowski, IFiS

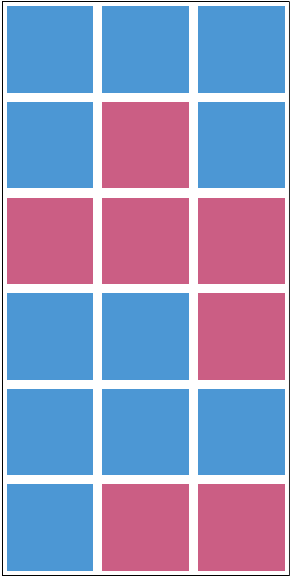
Definition of missing values

- Missing values are those values that are not observed
- Values do exist in theory, but we are unable to see them

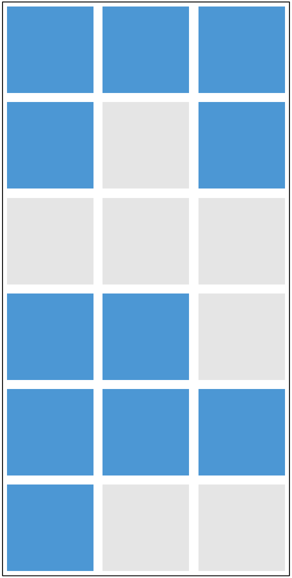
Some confusing terminology

- Complete data = Observed data + Unobserved data
- Incomplete data = Observed data
- Missing data = Unobserved data
- Complete cases = Subset of rows without missing values
- Complete variables = Subset of columns without missing values

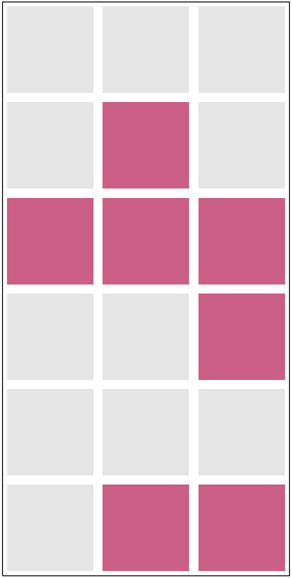
Complete data



Incomplete data = observed data



Missing data = unobserved data



Why values can be missing

Missingness can occur for a lot of reasons. For example

- power failure, bad luck
- death, dropout, refusal
- routing, experimental design
- join, merge, bind
- different variables per source
- different number of categories per source

Consequences of missing data

- Cannot calculate, not even the mean
- Less information than planned
- Enough statistical power?
- Different analyses, different n's
- Systematic biases in the analysis
- Appropriate confidence interval, P-values?

Missing data can severely complicate interpretation and analysis

Strategies to deal with missing data

- Prevention - impossible for ex-post analyses
- Weighting methods
- Likelihood methods, EM-algorithm
- Ad-hoc methods, e.g., single imputation, complete cases, recoding
- Multiple imputation

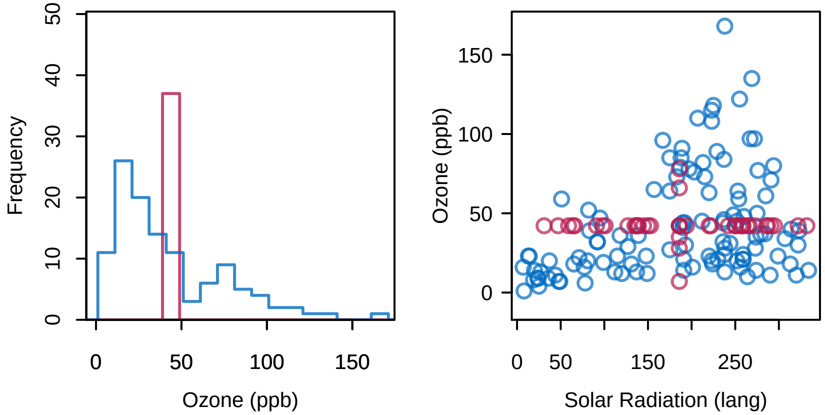
Listwise deletion, complete-case analysis

- Analyze only the complete records
- Advantages
  - Simple (default in most software)
  - Unbiased under MCAR
  - Conservative standard errors, significance levels
  - Two special properties in regression

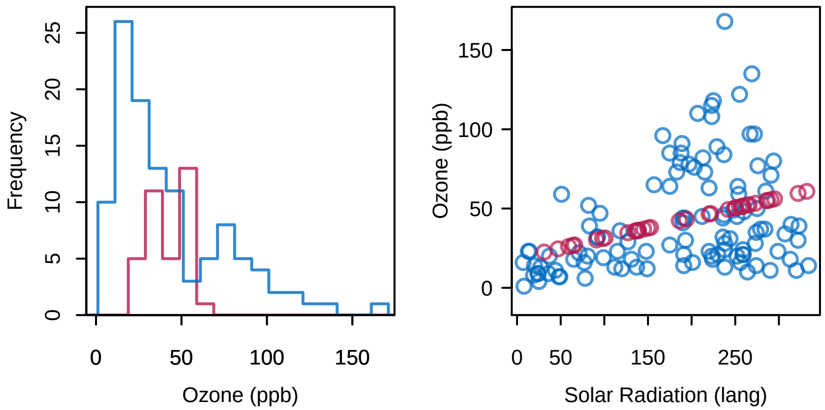
Listwise deletion, complete-case analysis

- Disadvantages
  - Wasteful
  - May not be possible
  - Larger standard errors
  - Biased under MAR, even for simple statistics like the mean
  - Inconsistencies in reporting

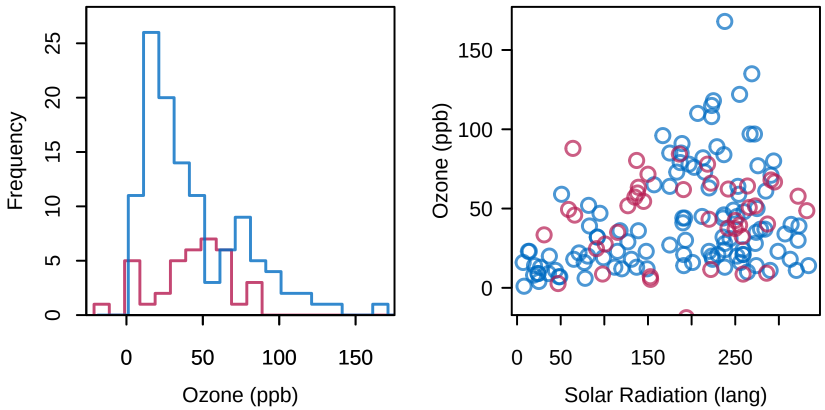
Mean imputation



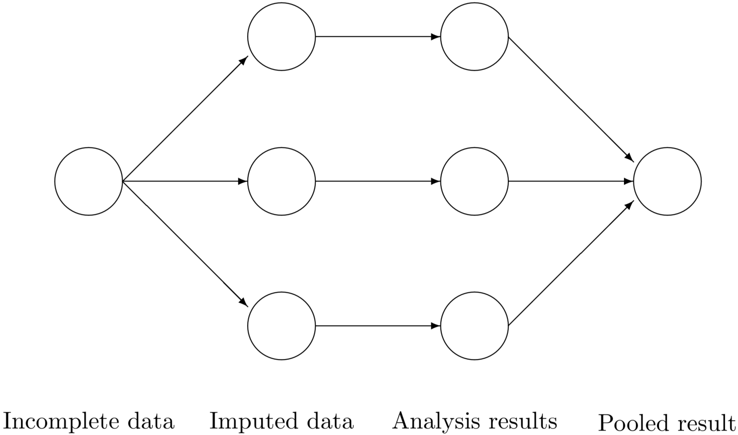
Regression imputation



Stochastic regression imputation



Multiple imputation



Acceptance of multiple imputation

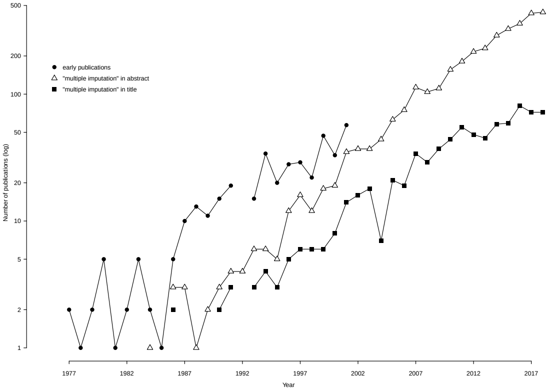


Figure 1: Source: Scopus (April 3, 2019)

Pooled estimate $\bar{Q}$

$\hat{Q}_\ell$ is the estimate of the ℓ-th repeated imputation

$\hat{Q}_\ell$ contains k parameters, represented as a $k \times 1$ column vector

Pooled estimate $\bar{Q}$ is simply the average

$$\bar{Q} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{Q}_\ell$$

Within-imputation variance

Average of the complete-data variances as

$$\bar{U} = \frac{1}{m}\sum_{\ell=1}^{m}\bar{U}_\ell,$$

where $\bar{U}_\ell$ is the variance-covariance matrix of $\hat{Q}_\ell$ obtained for the ℓ-th imputation

$\bar{U}_\ell$ is the variance is the estimate, *not* the variance in the data

Within-imputation variance is large if the sample is small

Between-imputation variance

Variance between the m complete-data estimates is given by

$$B = \frac{1}{m-1}\sum_{\ell=1}^{m}(\hat{Q}_\ell - \bar{Q})(\hat{Q}_\ell - \bar{Q})',$$

where $\bar{Q}$ is the pooled estimate.

The between-imputation variance is large there many missing data

Total variance

The total variance is *not* simply $T = \bar{U} + B$

The correct formula is

$$\begin{aligned} T &= \bar{U} + B + B/m \\ &= \bar{U} + \left(1 + \frac{1}{m}\right) B \end{aligned} \quad (1)$$

for the total variance of $\bar{Q}_m$, and hence of $(Q - \bar{Q})$ if $\bar{Q}$ is unbiased

The term B/m is the simulation error

Three sources of variation

In summary, the total variance T stems from three sources:

1. $\bar{U}$, the variance caused by the fact that we are taking a sample rather than the entire population. This is the conventional statistical measure of variability;
2. B, the extra variance caused by the fact that there are missing values in the sample;
3. B/m, the extra simulation variance caused by the fact that $\bar{Q}_m$ itself is based on finite m.

Variance ratio's (1)

Proportion of the variation attributable to the missing data

$$\lambda = \frac{B + B/m}{T}$$

Relative increase in variance due to nonresponse

$$r = \frac{B + B/m}{\bar{U}}$$

These are related by $r = \lambda/(1 - \lambda)$.

Variance ratio's (2)

Fraction of information about Q missing due to nonresponse

$$\gamma = \frac{r + 2/(\nu + 3)}{1 + r}$$

This measure needs an estimate of the degrees of freedom ν (c.f. section 2.3.6)

Relation between γ and λ

$$\gamma = \frac{\nu + 1}{\nu + 3}\lambda + \frac{2}{\nu + 3}.$$

The literature often confuses γ and λ.

Statistical inference for $\bar{Q}$ (1)

The $100(1-\alpha)\%$ confidence interval of a $\bar{Q}$ is calculated as

$$\bar{Q} \pm t_{(\nu,1-\alpha/2)}\sqrt{T},$$

where $t_{(\nu,1-\alpha/2)}$ is the quantile corresponding to probability $1-\alpha/2$ of t_ν.

For example, use $t(10, 0.975) = 2.23$ for the 95% confidence interval for $\nu = 10$.

Statistical inference for $\bar{Q}$ (2)

Suppose we test the null hypothesis $Q = Q_0$ for some specified value Q_0. We can find the P-value of the test as the probability

$$P_s = \Pr\left[F_{1,\nu} > \frac{(Q_0 - \bar{Q})^2}{T}\right]$$

where $F_{1,\nu}$ is an F distribution with 1 and ν degrees of freedom.

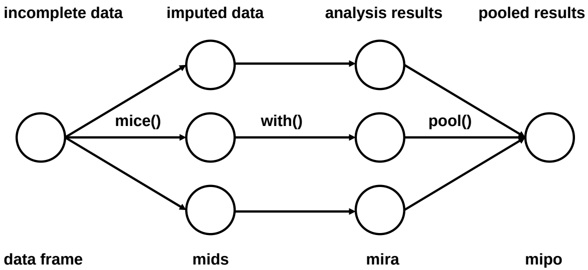
How large should m be?

Classic advice: $m = 3, 5, 10$. More recently: set m higher: 20–100.

Some advice:

- Use $m = 5$ or $m = 10$ if the fraction of missing information is low, $\gamma < 0.2$.
- Develop your model with $m = 5$. Do final run with m equal to percentage of incomplete cases.

Multiple imputation in mice



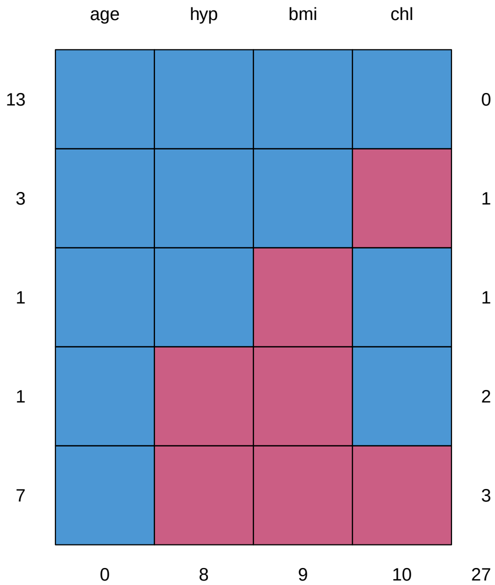
Inspect the data

```r
library("mice")
head(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
```

Inspect missing data pattern

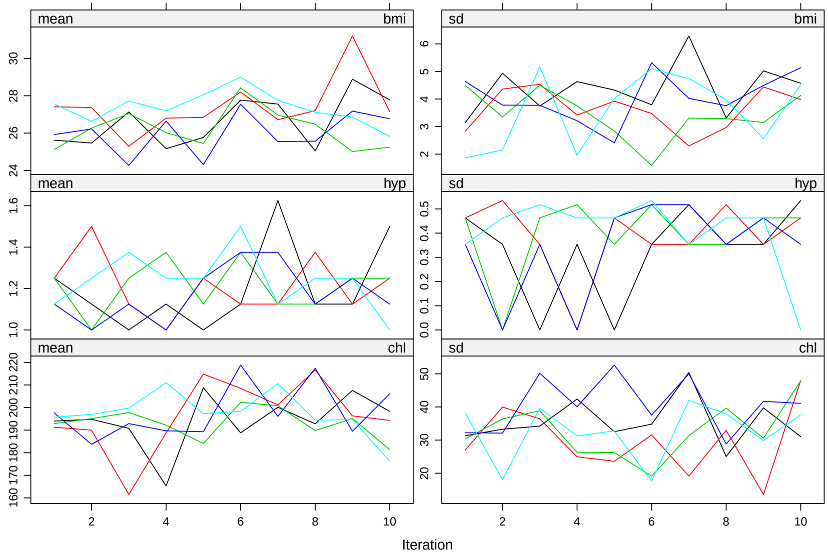
```r
md.pattern(nhanes)
```



Multiply impute the data

```
imp <- mice(nhanes, print = FALSE, maxit=10, seed = 24415)
```

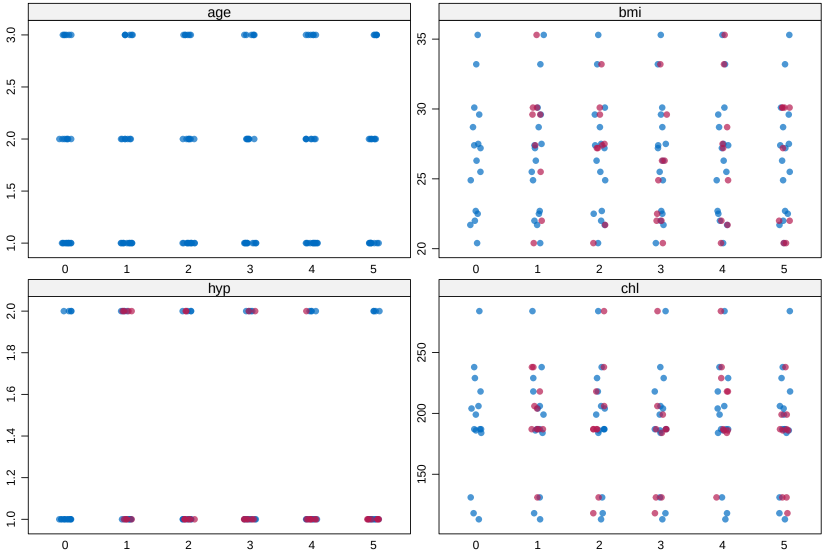
Inspect the trace lines for convergence



Stripplot of observed and imputed data

```r
stripplot(imp, pch = 20, cex = 1.2)
```

Stripplot of observed and imputed data

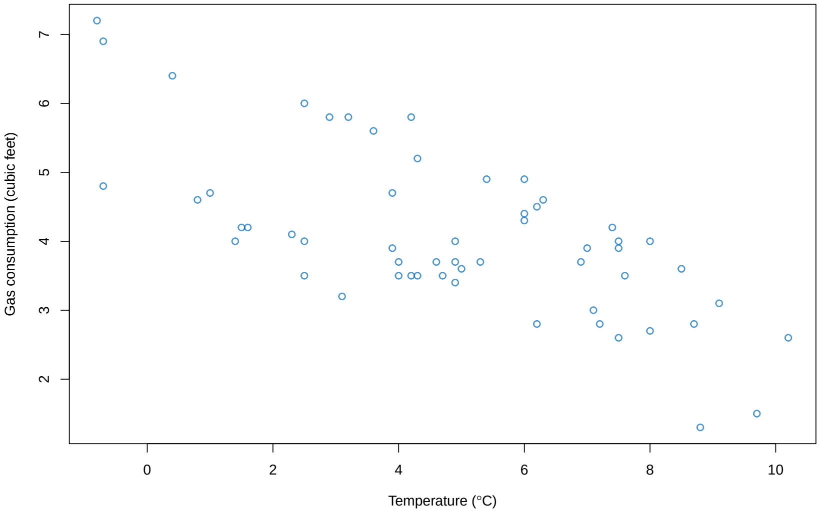


Fit the complete-data model

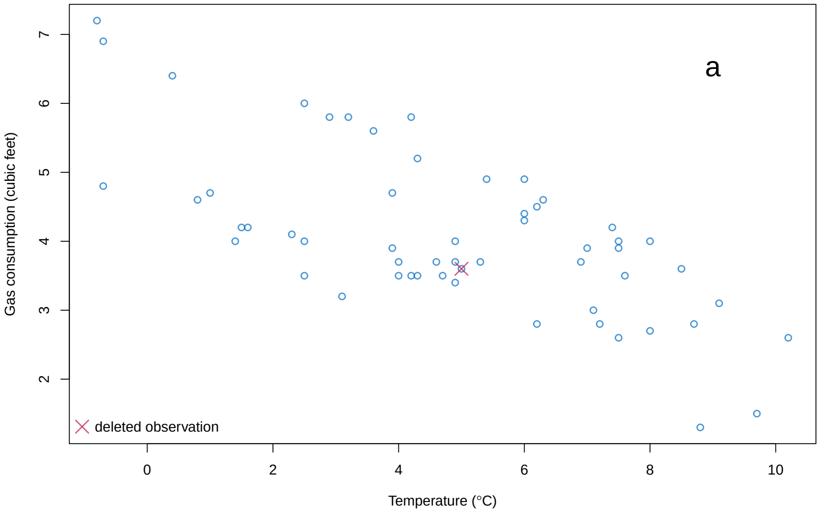
```r
fit <- with(imp, lm(bmi ~ age))
est <- pool(fit)
summary(est)
```

```
##             estimate std.error statistic   df  p.value
## (Intercept)    30.69      2.09     14.70 13.4 1.16e-09
## age            -2.35      1.01     -2.33 17.4 3.23e-02
```

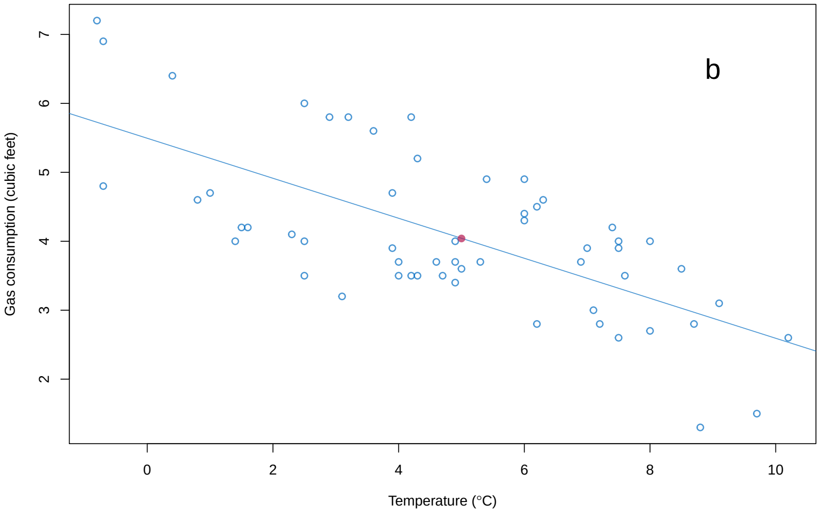
Temperature and gas consumption



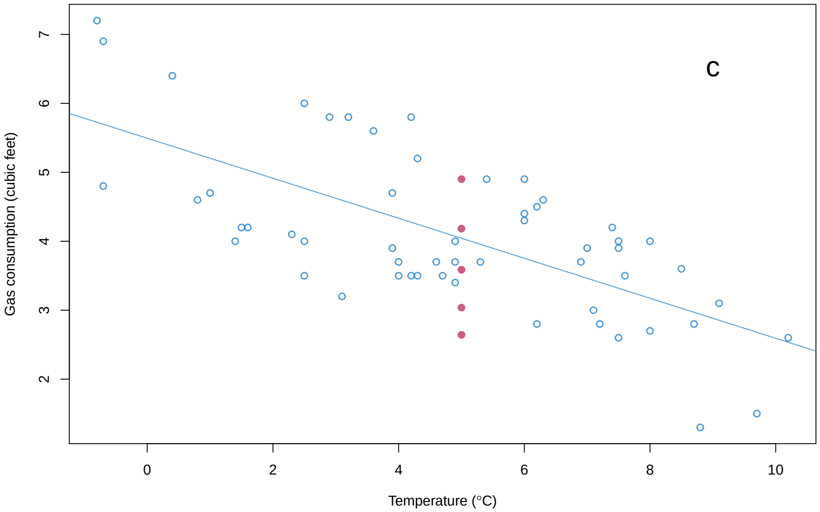
Delete gas consumption of day 47



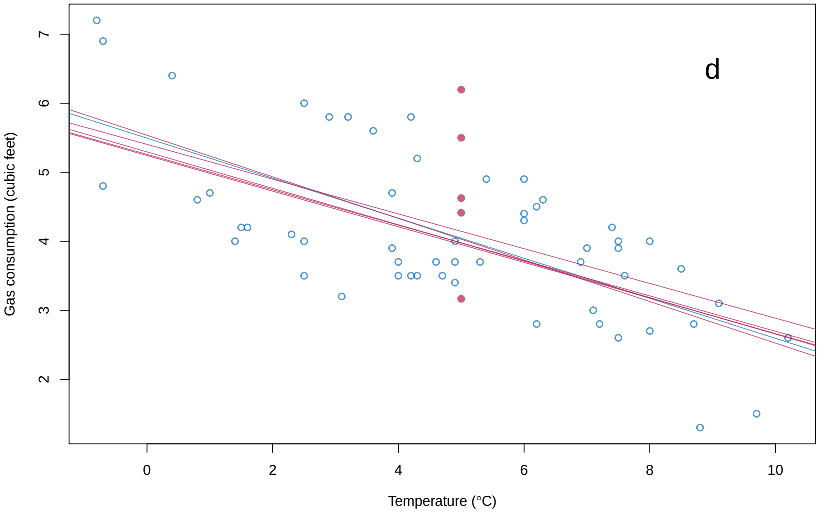
Predict value from regression line



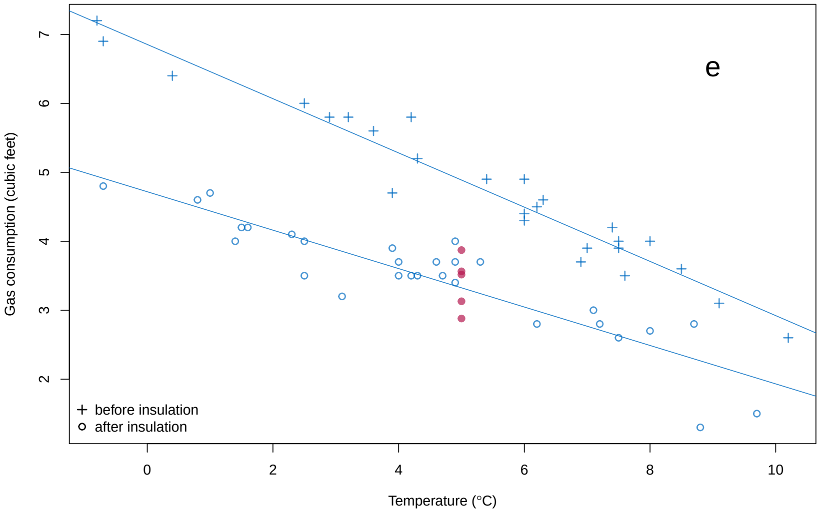
Predict value + add noise



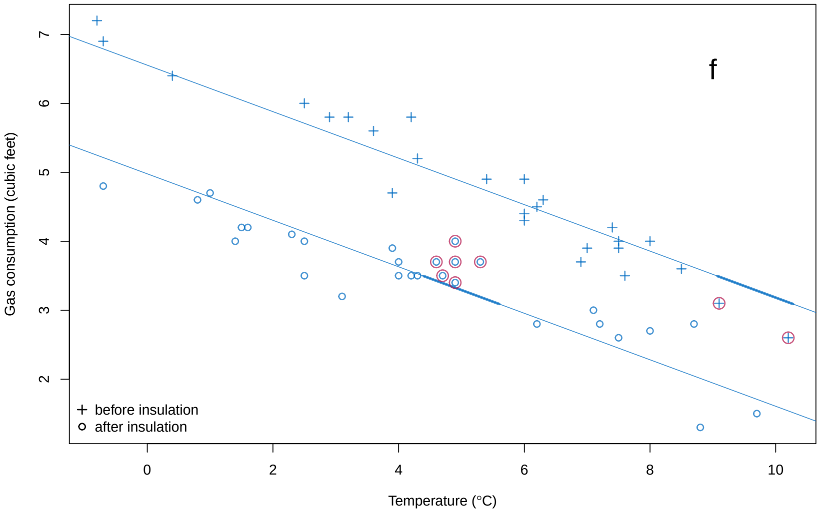
Predict + noise + parameter draw



Two predictors

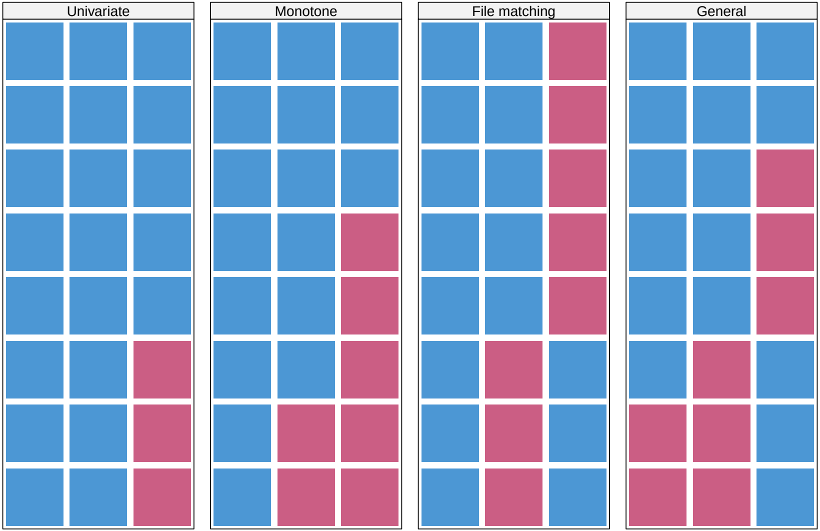


Drawing from observed data



Multivariate missing data

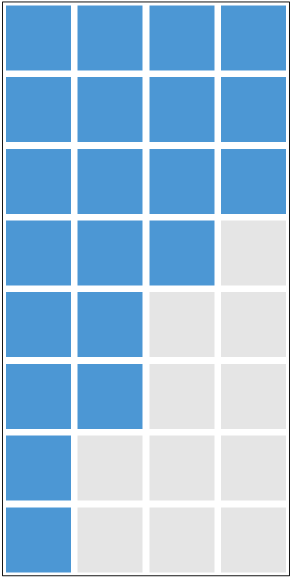
Missing data patterns



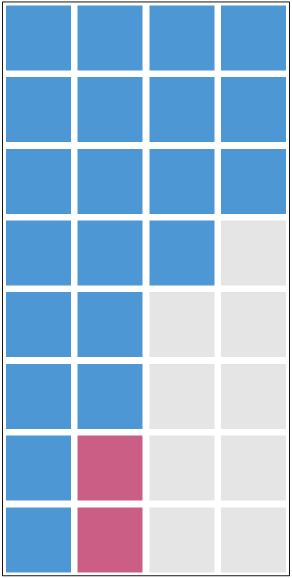
Three general strategies

- Monotone data imputation
- Joint modeling
- Fully conditional specification (FCS)

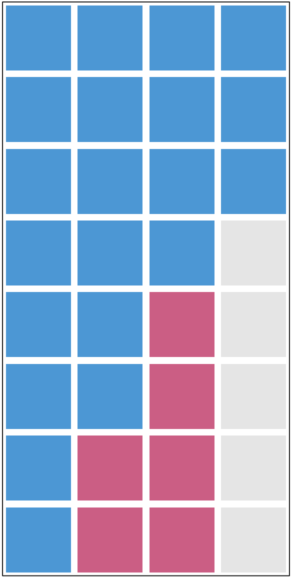
Imputation of monotone pattern



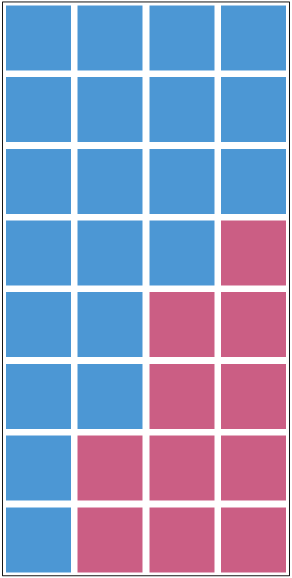
Imputation of monotone pattern



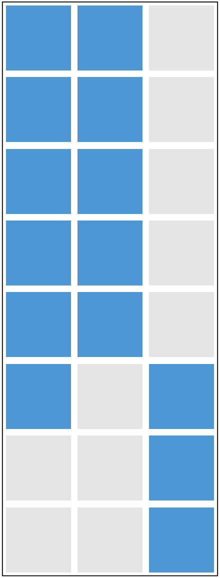
Imputation of monotone pattern



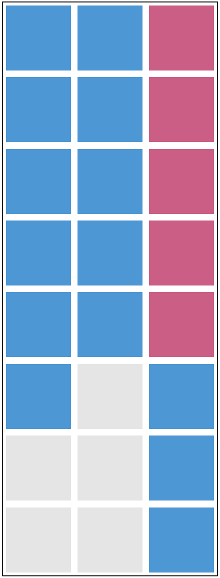
Imputation of monotone pattern



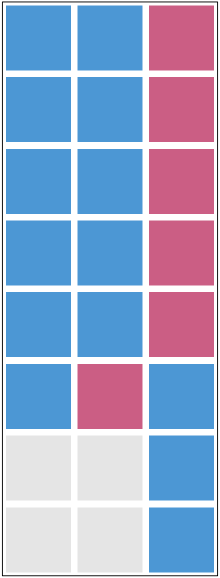
Imputation by joint modelling



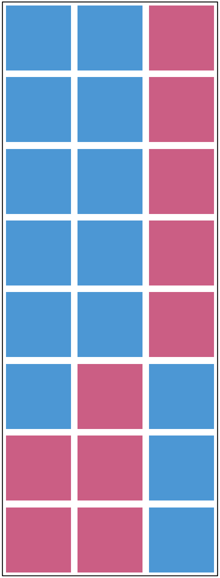
Imputation by joint modelling



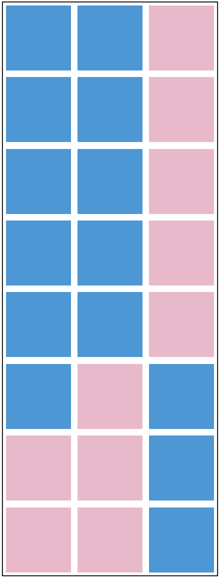
Imputation by joint modelling



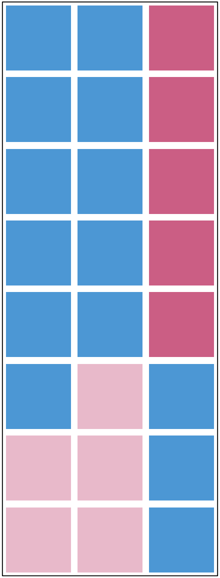
Imputation by joint modelling



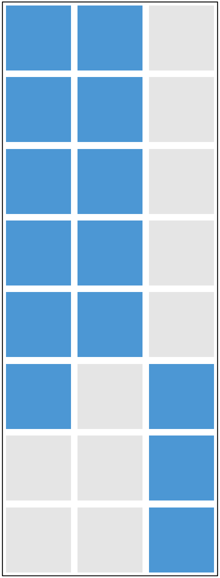
Joint modelling - next iteration



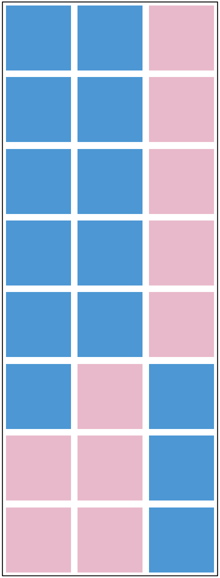
Joint modelling - next iteration



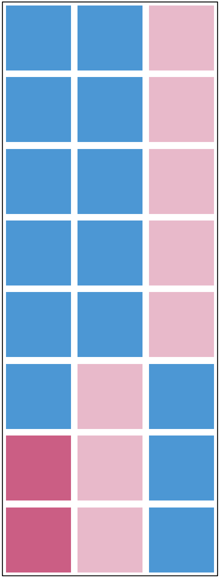
Fully conditional specification



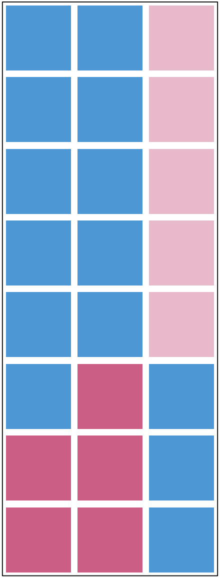
Fully conditional specification



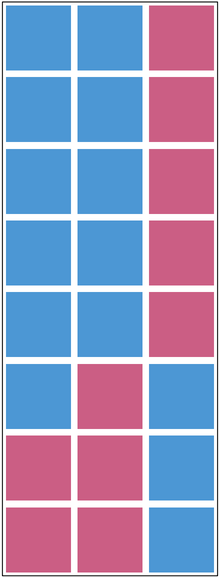
Fully conditional specification



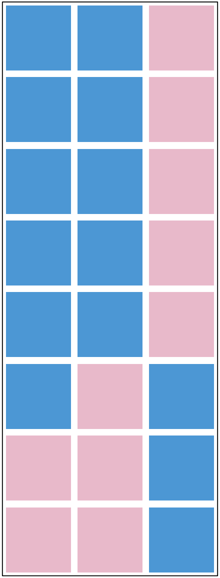
Fully conditional specification



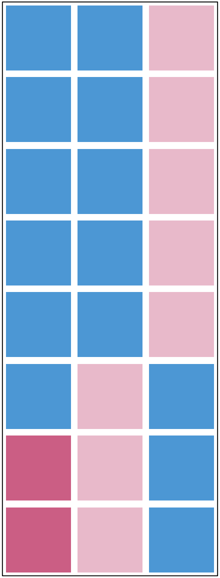
Fully conditional specification



Fully conditional specification - next



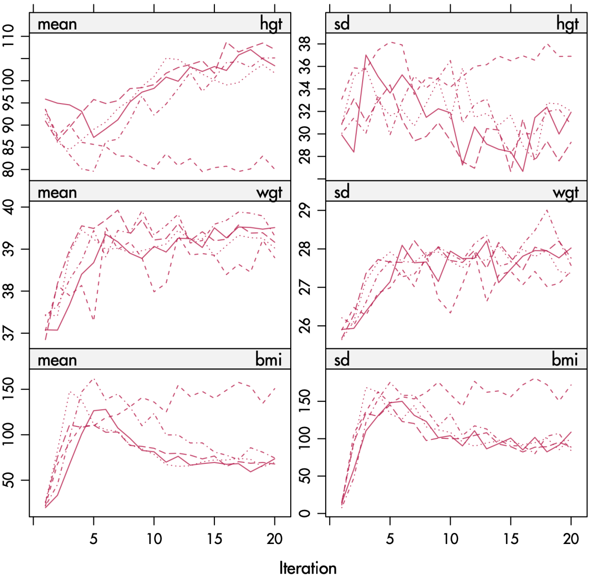
Fully conditional specification - next



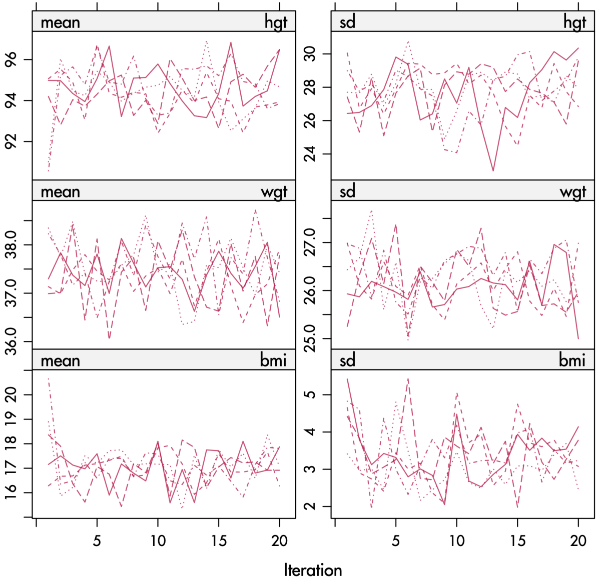
How many iterations?

- Quick convergence
- 5–10 iterations is adequate for most problems
- More iterations is λ is high
- Inspect the generated imputations
- Monitor convergence to detect anomalies

Non-convergence



Convergence



Conclusion

- A general problem, a general solution
- `mice` package: >50,000 downloads per month
- Highly useful for data combination