NSC R workshop

Stef van Buuren

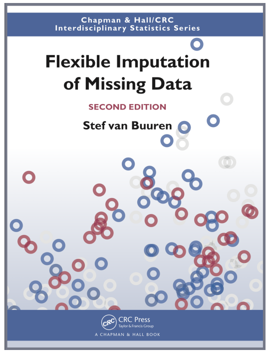
Apr 28, 2022 - ZOOM

Workshop topics

- Problem of missing data
- Strategies to deal with missing data
- Multiple imputation methodology to analyse incomplete data
- Using `R` package `mice`

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. Free text: https://stefvanbuuren.name/fimd Order book: https://www.crcpress.com/Flexible-Imputation-of-Missing-Data-Second-Edition/Buuren/p/book/9781138588318



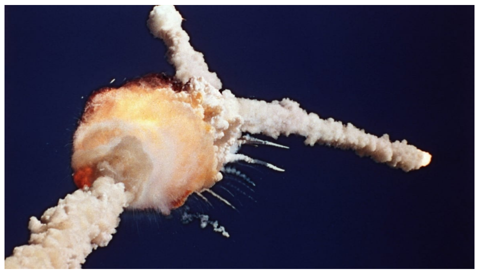
Motivation

- Real data are always incomplete
- Ad-hoc fixes do not (always) work
- Multiple imputation as principled and broadly applicable approach
- Goal: get comfortable with a powerful way to deal with incomplete data
- We use the `mice` package in `R`

What is missing data?

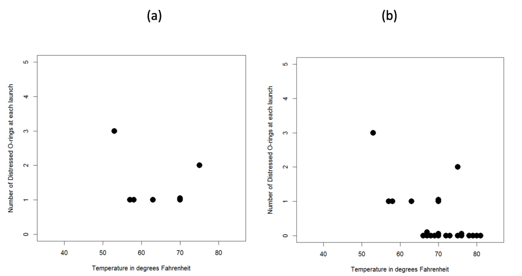
Missing data are concealed from us, and that very fact means we are at risk of misunderstanding, of drawing incorrect conclusions, and of making poor decisions.

Challenger space shuttle - 28 Jan 1986 - 7 deaths



Challenger space shuttle - 28 Jan 1986 - 7 deaths

Figure 1.1 (a) Data examined in the pre-launch teleconference; **(b)** Complete data.



Further characterization of missing values

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

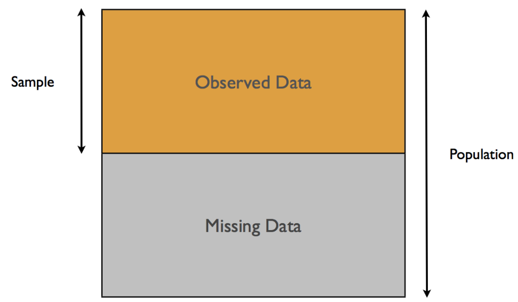
Evolving views on missing data

- "Obviously the best way to treat missing data is not to have them." (Orchard and Woodbury 1972)
- "Sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data" (Allison, 2002)
- "Missing data are the heart of statistics"

Missing data are the heart of statistics

- Taking a sample
- Estimating a causal effect
- Predicting future outcome
- Combining data from different sources

Sampling example



Reasons

Missing data can occur for a lot of reasons. For example

- death, dropout, refusal, concealed
- sampling, experimental design
- join, merge, bind
- too far away, too small to observe
- power failure, budget exhausted, bad luck

Why are missing values problematic?

- Cannot calculate, not even the mean
- Less information than planned
- 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
- Ad-hoc methods
- Weighting methods
- Likelihood methods, EM-algorithm
- Multiple imputation

Strategies to deal with missing data

- Prevention
- **Ad-hoc methods**
- Weighting methods
- Likelihood methods, EM-algorithm
- **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

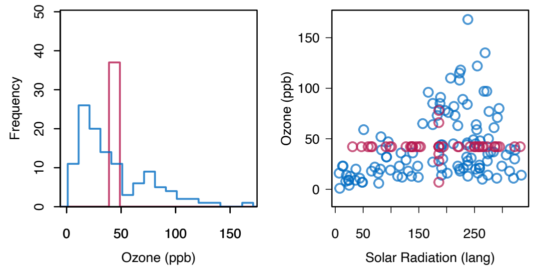
Listwise deletion, complete-case analysis

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

Mean imputation

- Replace the missing values by the mean of the observed data
- Advantages
  - Simple
  - Unbiased for the mean, under MCAR

Mean imputation



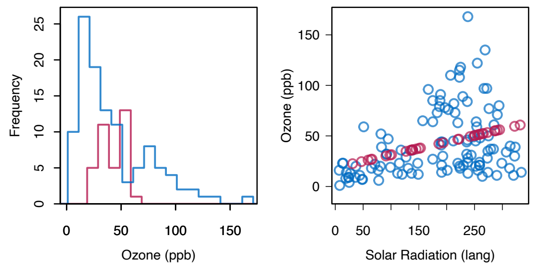
Mean imputation

- Disadvantages
  - Disturbs the distribution
  - Underestimates the variance
  - Biases correlations to zero
  - Biased under MAR
- AVOID (unless you know what you are doing)

Regression imputation

- Also known as **prediction**
  - Fit model for Y^{obs} under listwise deletion
  - Predict Y^{mis} for records with missing Y's
  - Replace missing values by prediction
- Advantages
  - Under MAR, unbiased estimates of regression coefficients
  - Good approximation to the (unknown) true data if explained variance is high
- Favourite among data scientists and machine learners

Regression imputation



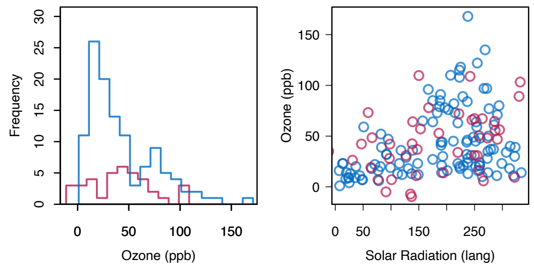
Regression imputation

- Disadvantages
  - Artificially increases correlations
  - Systematically underestimates the variance
  - Too optimistic P-values, too short confidence intervals
- AVOID. Harmful to statistical inference

Stochastic regression imputation

- Like regression imputation, but adds appropriate noise to the predictions to reflect uncertainty
- Advantages
  - Preserves the distribution of Y^{obs}
  - Preserves the correlation between Y and X in the imputed data

Stochastic regression imputation



Stochastic regression imputation

- Disadvantages
  - Symmetric and constant error restrictive
  - Single imputation incorrectly treats imputations as real data
  - Not so simple anymore

Overview of assumptions needed

	Unbiased			Standard Error
	Mean	Reg Weight	Correlation	
Listwise	MCAR	MCAR	MCAR	Too large
Pairwise	MCAR	MCAR	MCAR	Complicated
Mean	MCAR	–	–	Too small
Regression	MAR	MAR	–	Too small
Stochastic	MAR	MAR	MAR	Too small
LOCF	–	–	–	Too small
Indicator	–	–	–	Too small

Multiple imputation

- Imputes each missing value m times
- Variation between the m imputed values reflects our ignorance about the true value

Acceptance of multiple imputation

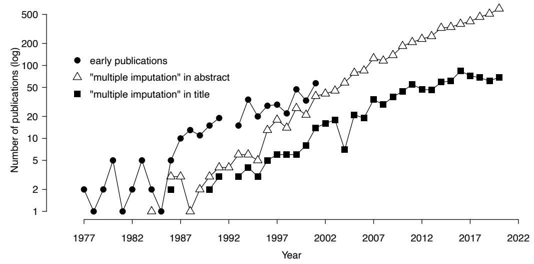
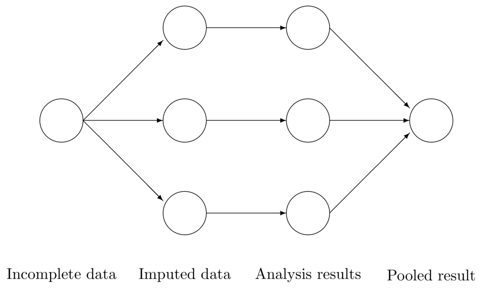


Figure 1: Source: Scopus (May 27, 2021)

Multiple imputation



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.

Multiple imputation

- Advantages
  - Correct point and variance estimates
  - Splits missing data problem from complete-data analysis
  - Theoretical properties well established
  - Flexible, widely applicable
  - Extensible to MNAR
- Disadvantages
  - Need to create and work with multiple imputed data sets
  - May not always be most efficient

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?

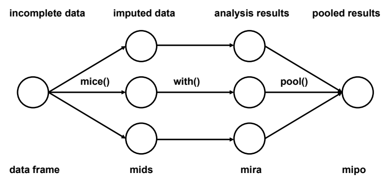
Classic advice: $m = 3, 5, 10$. More recently: set m higher: 20–100.

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`

Generic workflow in `mice`



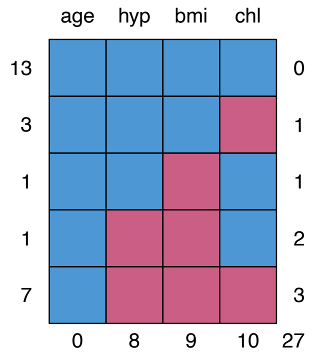
Inspect the data

```
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

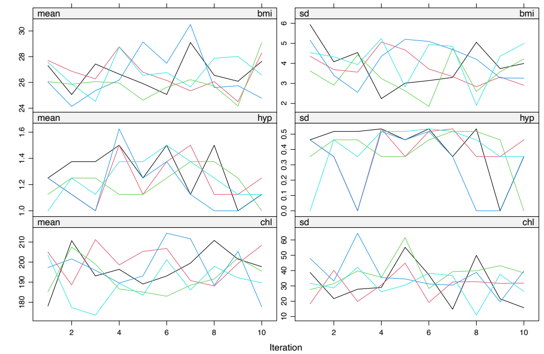
```
md.pattern(nhanes)
```



Multiply impute the data

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

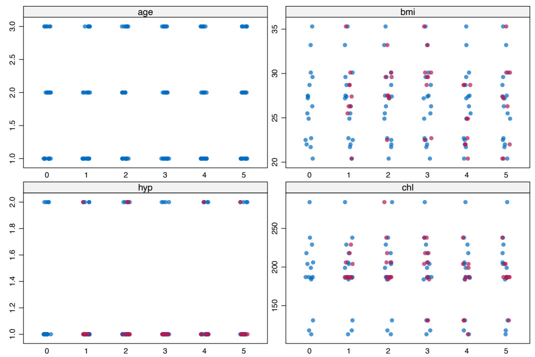
Inspect the trace lines for convergence



Stripplot of observed and imputed data

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

Stripplot of observed and imputed data



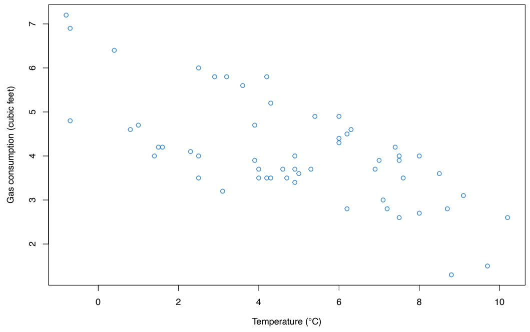
Fit the complete-data model

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

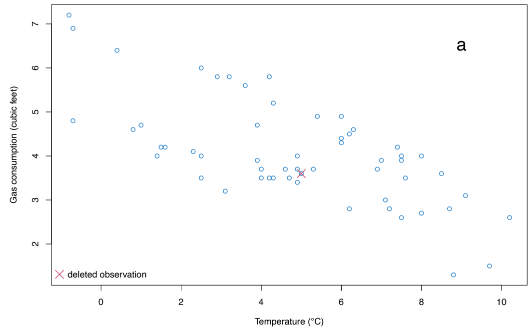
```
##          term estimate std.error statistic   df p.value
## 1 (Intercept)     30.5      2.45     12.46  7.2 3.94e-0
## 2         age     -2.1      1.12     -1.87 10.8 8.89e-0
```

Creating univariate imputations

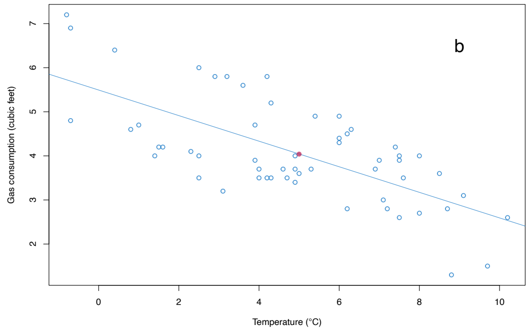
Relation between temperature and gas consumption



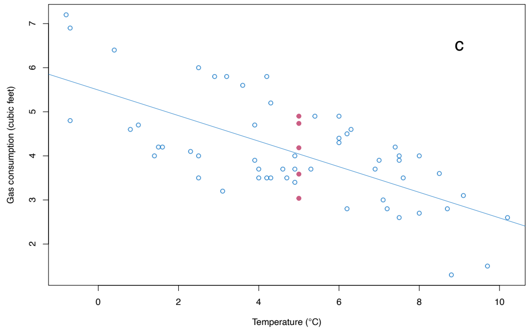
We delete gas consumption of observation 47



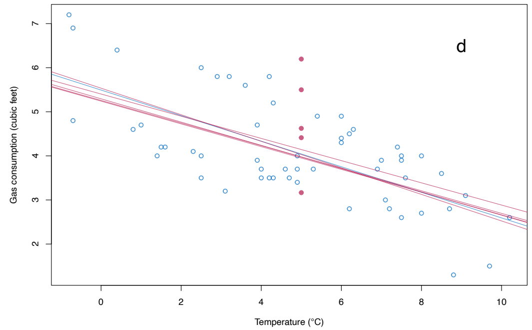
Predict imputed value from regression line



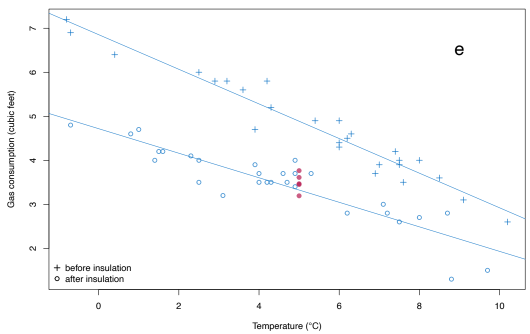
Predicted value + noise



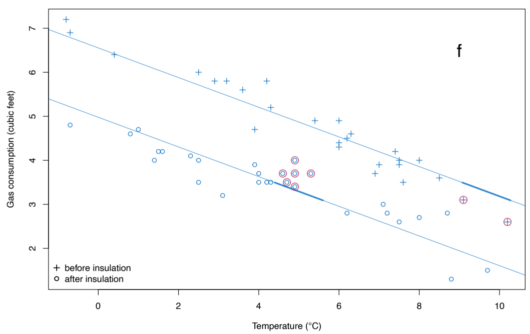
Predicted value + noise + parameter uncertainty



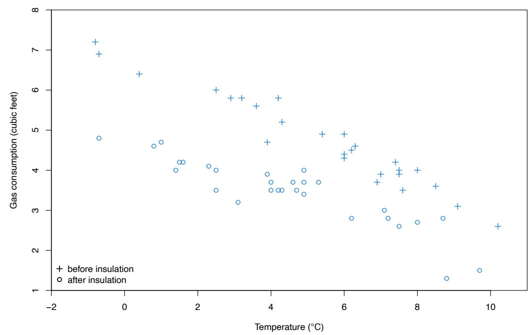
Imputation based on two predictors



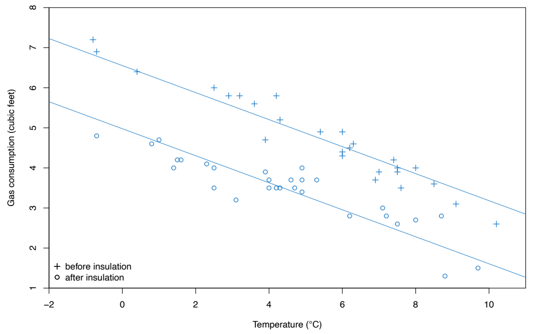
Drawing from the observed data



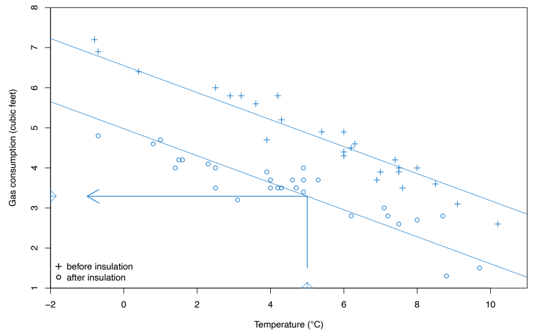
Predictive mean matching



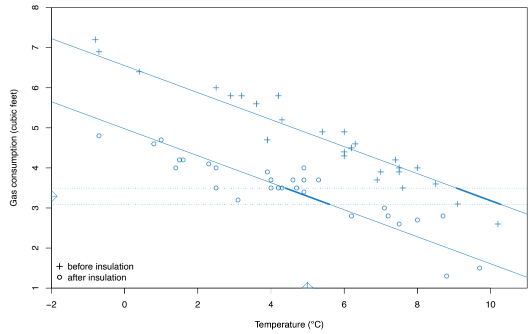
PMM: Add two regression lines



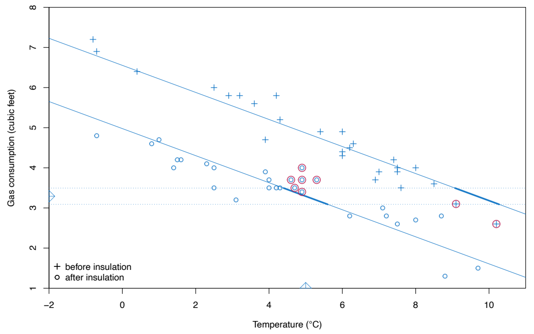
PMM: Predicted given 5°,C, ‘after insulation’



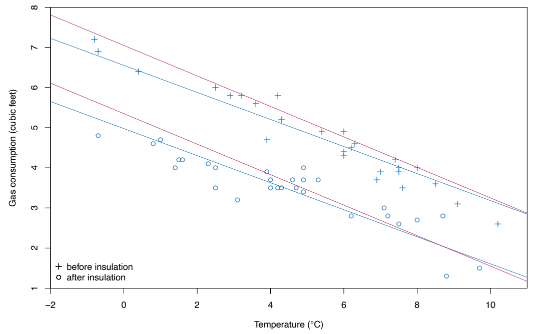
PMM: Define a matching range $\hat{y} \pm \delta$



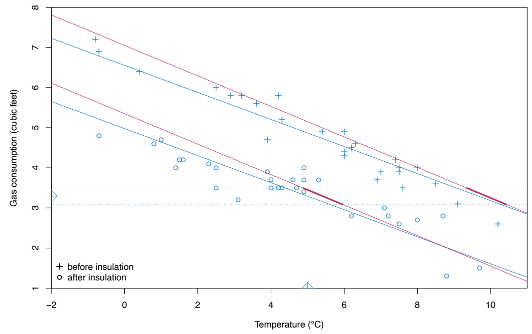
PMM: Select potential donors



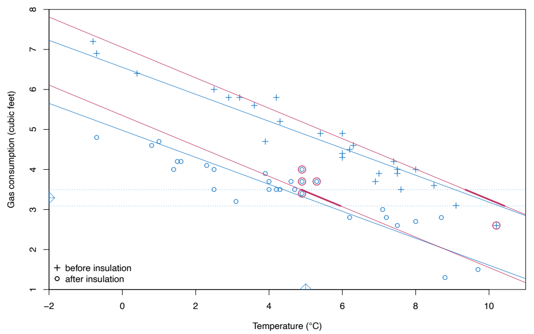
PMM: Bayesian PMM: Draw a line



PMM: Define a matching range $\hat{y} \pm \delta$



PMM: Select potential donors



Built-in imputation functions

https://amices.org/mice/reference/index.html

Creating multivariate imputations, MICE algorithm

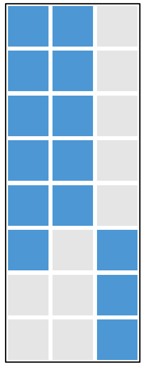
Issues in multivariate imputation

- The predictors Y_{-j} themselves can contain missing values;
- "Circular" dependence can occur, where Y_j^{mis} depends on Y_h^{mis}, and vice versa;
- Especially with large p and small n, collinearity or empty cells can occur;
- Derived variables;
- The ordering of the rows and columns can be meaningful, e.g., as in longitudinal data;
- Imputation can create impossible combinations, such as pregnant grandfathers.

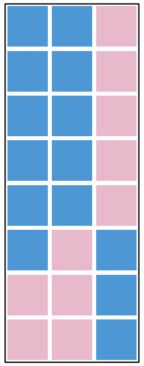
Fully conditional specification (FCS), MICE algorithm

- Imputes multivariate missing data on a variable-by-variable basis
- Requires a specification of an imputation model for each incomplete variable
- Creates imputations per variable in an iterative fashion

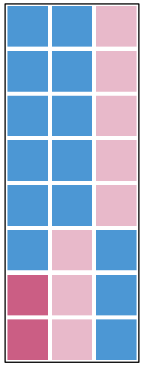
Imputation by fully conditional specification



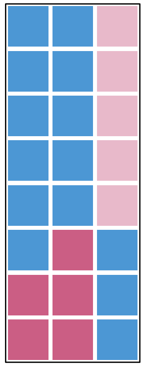
Imputation by fully conditional specification



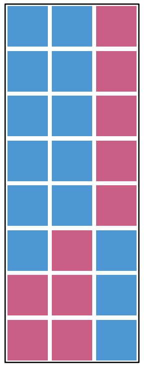
Imputation by fully conditional specification



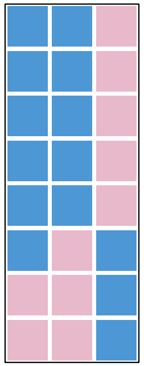
Imputation by fully conditional specification



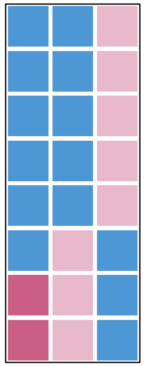
Imputation by fully conditional specification



Imputation by fully conditional specification - next iteration



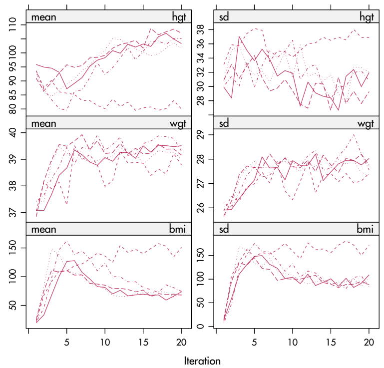
Imputation by fully conditional specification - next iteration



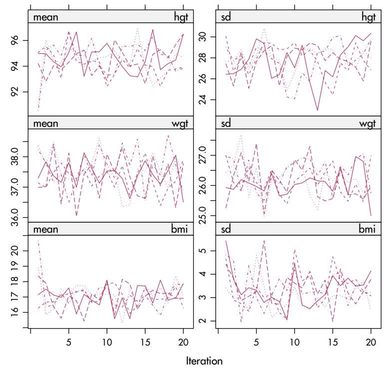
How many iterations?

- 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



Number of iterations

Watch out for situations where

- the correlations between the Y_j's are high;
- the missing data rates are high; or
- constraints on parameters across different variables exist.

That's it!

More R code and examples

- GitHub site: https://github.com/amices/mice

Conclusion

- Missing data are a fact of life, and actually interesting
- There are many ways to treat missing data, only few are valid
- Always try to prevent missing data
- Use ad-hoc methods with caution
- Multiple imputation is an all-round general purpose method
- Many applications possible