Multiple Imputation in Practice (MIMP) S28

https://www.gerkovink.com/mimp

Stef van Buuren, Gerko Vink

July 11-14, 2022

Welcome

Course links

Summer School MIMP

Our teaching staff

- Stef van Buuren
- Gerko Vink
- Thom Volker
- Hanne Oberman
- Mingyang Cai

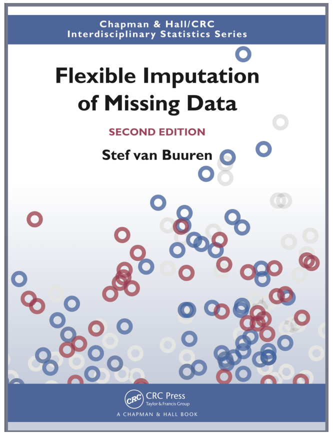
Overview

Why this course?

- Missing data are everywhere
- Ad-hoc fixes do not (always) work
- Multiple imputation is broadly applicable, yield correct statistical inferences, and there is good software
- **Goal**: Get comfortable with a powerful way of solving missing data problems
- We use the `mice` package in `R`

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



R

- Why `R`?

R software and examples

- Course site: https://www.gerkovink.com/mimp
- R install from https://cran.r-project.org
- R package: `mice 3.14.0`, https://cran.r-project.org/package=mice
- Development version: `mice 3.14.7`, https://github.com/amices/mice
- Documentation: https://amices.org/mice/
- Example code: https://github.com/stefvanbuuren/fimdbook/blob/master/R/fimd.R

Course schedule

Day	Location	Lecture.1	Practical.1	Lecture.2	Practical.2
		9am - 10.30am	10.45am - 12.15pm	1.15pm - 2.30pm	2.45pm - 4pm
Monday	Atlas	A	B	C	D
Tuesday	Atlas	E	F	G	H
Wednesday	Atlas	I	J	K	L
Thursday	Van Lier	M	N	O	P

Schedule for Monday, Jul 11

Slot	Type	Description	FIMD2
A	L	Introduction	Ch1
B	P	Ad-hoc methods and `mice`	`nhanes`
C	L	Multiple imputation, Univariate	Ch2, 3.1–3.7
D	P	Imputation with `mice`	`nhanes`

Schedule for Tuesday, Jul 12

Slot	Type	Description	FIMD2
E	L	Multivariate imputation	Ch4,5.6
F	P	Multivariate imputation in R	`mammalsleep`, `boys`
G	L	Modelling, derived variables	6.1-6.4
H	P	Imputation derived variables	`mammalsleep`, `boys`

Schedule for Wednesday, Jul 13

Slot	Type	Description	FIMD2
I	L	Combining inferences	Ch5
J	P	Analysis in R	
K	L	Sensitivity, reporting	3.8, 9.2, 12.2
L	P	Approach to sensitivity analysis	`leiden85`

Schedule for Thursday, Jul 14

Slot	Type	Description	FIMD2
M	L	Advanced features	various
N	P	Advanced features with in `mice`	
O	L	Capita selecta	
P	P	Get advice/support	

Introduction into missing data - MIMP A

Overview A

- Evolving views on missing data
- Why are missing data interesting?
- Terminology and concepts
- Strategies to deal with missing data

Overview A

- **Evolving views on missing data**
- Why are missing data interesting?
- Terminology and concepts
- Strategies to deal with missing data

Evolving views on missing data - 1970

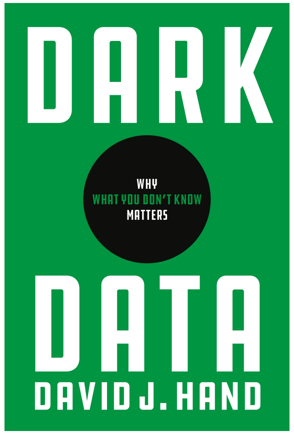
"Obviously the best way to treat missing data is not to have them."
— Orchard and Woodbury, 1972

Evolving views on missing data - 2000

"Sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data."
— Paul Allison, 2002

Evolving views on missing data - 2020

"Dark 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."
— David Hand, 2020

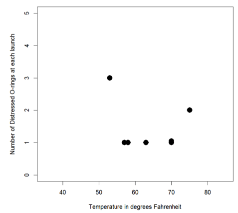


Challenger space shuttle - 28 Jan 1986 - 7 deaths

Challenger space shuttle - 28 Jan 1986 - 7 deaths

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

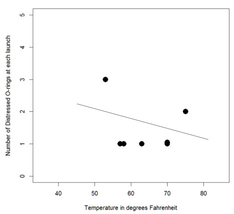
(a)



Challenger space shuttle - 28 Jan 1986 - 7 deaths

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

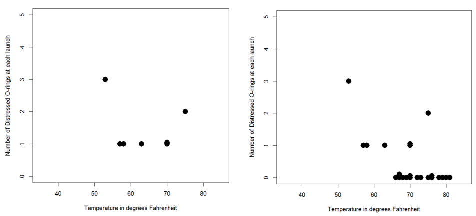
(a)



Challenger space shuttle - 28 Jan 1986 - 7 deaths

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

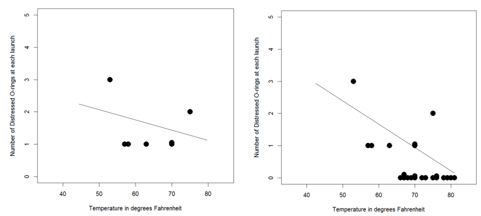
(a) (b)



Challenger space shuttle - 28 Jan 1986 - 7 deaths

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

(a) (b)



Dark data types (1/2)

- DD-Type 1: Data We Know Are Missing
- DD-Type 2: Data We Don't Know are Missing
- DD-Type 3: Choosing Just Some Cases
- DD-Type 4: Self-Selection
- DD-Type 5: Missing What Matters
- DD-Type 6: Data Which Might Have Been
- DD-Type 7: Changes with Time
- DD-Type 8: Definitions of Data
- DD-Type 9: Summaries of Data
- DD-Type 10: Measurement Error and Uncertainty

Dark data types (2/2)

- DD-Type 11: Feedback and Gaming
- DD-Type 12: Information Asymmetry
- DD-Type 13: Intentionally Darkened Data
- DD-Type 14: Fabricated and Synthetic Data
- DD-Type 15: Extrapolating beyond Your Data

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

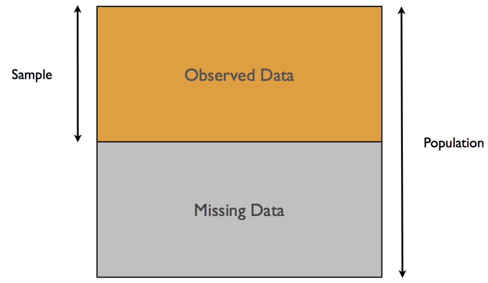
Overview A

- Evolving views on missing data
- **Why are missing data interesting?**
- Terminology and concepts
- Strategies to deal with missing data

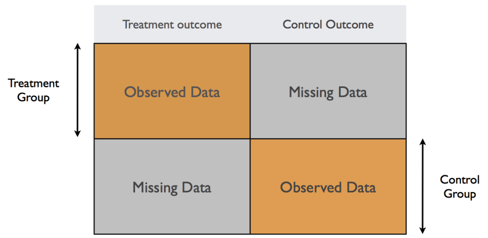
Why are missing data interesting?

- MISSING DATA ARE THE HEART OF STATISTICS
- Taking a sample
- Estimating a causal effect
- Predicting future outcome
- Combining data from different sources

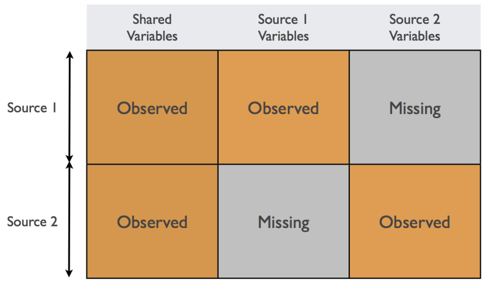
Sampling example



Experiment example



Matching example



Reasons

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

- death, dropout, refusal, concealed
- routing, 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
- 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

Overview A

- Evolving views on missing data
- Why are missing data interesting?
- **Terminology and concepts**
- Strategies to deal with missing data

Some confusing terminology

Complete data = Observed data + Unobserved data

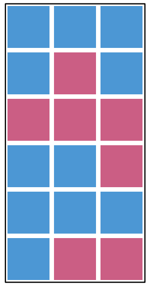
Incomplete data = Observed data

Missing data = Unobserved data

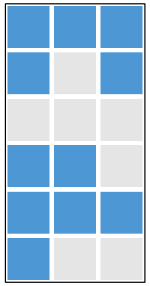
Complete cases = subset of rows in the observed data without missing values

Complete variables = subset of columns in the observed data without missing values

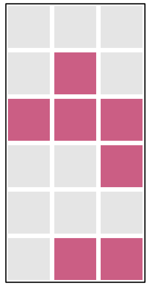
Complete data



Incomplete data = observed data



Missing data = unobserved data



Notation: Y, R, X

- Y random variable with missing data
- Y^{obs} true and observed values of Y
- Y^{mis} true but unobserved values of Y, missing values
- R response indicator
- $R = 1$ if Y is observed
- $R = 0$ if Y is missing
- X complete covariate

Missing data mechanism

- Process that governs which Y's are observed and which Y's are unobserved (Rubin, 1976)
- Sometimes we know this process (e.g.~experimental design, sampling)
- Alternatively, model by response probability $P(R|Y^{\text{obs}}, Y^{\text{mis}}, X)$
- Also called **missing data model**

MCAR: Missing Completely at Random

- Probability to be missing is not related to any data

$$P(R|Y^{\text{obs}}, Y^{\text{mis}}, X, \psi) = P(R|\psi)$$

- Examples
  - data transmission error
  - random sample

MAR: Missing at Random

- Probability to be missing depends on known data

$$P(R|Y^{\text{obs}}, Y^{\text{mis}}, X, \psi) = P(R|Y^{\text{obs}}, X, \psi)$$

- Examples
  - Income, where we have X related to wealth
  - Branch patterns (e.g. how old are your children?)

MNAR: Missing Not at Random

- Probability to be missing depends on unknown data

$$P(R|Y^{\text{obs}}, Y^{\text{mis}}, X, \psi) \text{ does not simplify}$$

- Examples
  - Income, without covariates related to income
  - Body weight report

Overview A

- Evolving views on missing data
- Why are missing data interesting?
- Terminology and concepts
- **Strategies to deal with missing data**

Strategies to deal with missing data

- **Prevention**
- Ad-hoc methods, e.g., single imputation, complete cases
- Weighting methods
- Likelihood methods, EM-algorithm
- Multiple imputation

Prevention

- Design: Time intervals, Number of variables, Pilot study
- Collection: Incentives, Match interviewer-respondent, Quick follow-up, Retrieve missing data
- Measures: Use short forms, Minimize intrusive measures, Clarity, Layout
- Treatment: Minimize burden and intensity
- Data entry: Double coding

Strategies to deal with missing data

- Prevention
- **Ad-hoc methods, e.g., single imputation, complete cases**
- 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: Special properties

- For any regression with missing in X, estimates under listwise deletion are unbiased as long as the missingness does not depend on Y. Includes even some cases of MNAR (Glynn & Laird, 1986; Little 1992).
- In logistic regression: With missing in Y or X (but not both), parameter estimates under listwise deletion are unbiased as long as the missingness depends only on Y (and not on X) (except for the intercept) (Vach 1994). This property is widely exploited in case-control studies in epidemiology.
- See FIMD2 2.7 https://stefvanbuuren.name/fimd/sec-when.html

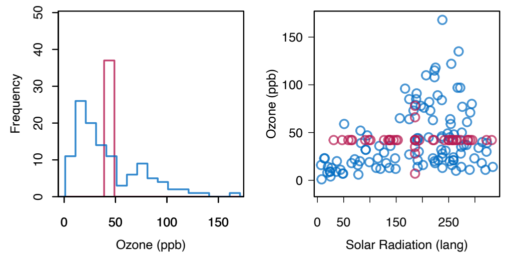
Listwise deletion, complete-case analysis

- 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



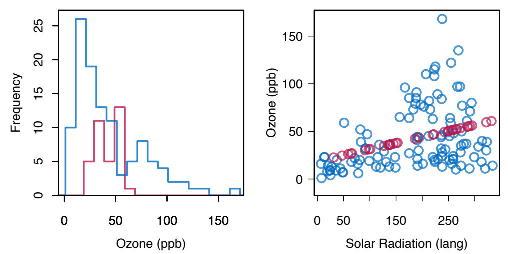
Mean imputation

- 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



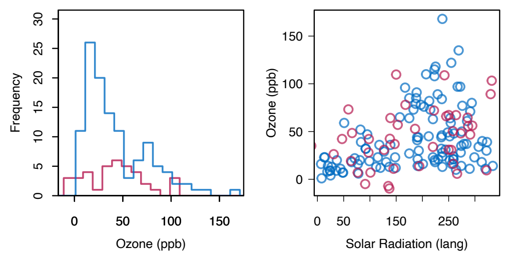
Regression imputation

- Disadvantages
  - Artificially increases correlations
  - Systematically underestimates the variance
  - Too optimistic P-values and 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: does not take uncertainty imputed data into account, and incorrectly treats them as real
  - Not so simple anymore

Indicator method

- Also known as *dummy variable adjustment*
- Complete-data model: $Y = X\beta + \epsilon$, missing data in X
- Pseudocode: `recode X(missing(X)=1, else=0)` into R
- `recode X(missing(X)=mean(X),else=copy)` into Z
- Fit $Y = Z\beta + R\gamma + \epsilon$ instead of $Y = X\beta + \epsilon$
- Advantages
  - Simple
  - Can increase efficiency of the treatment estimate in randomized trails, even under some MNAR cases

Indicator method

- Disadvantages
  - Biased estimates, even under MCAR
  - Incorrect P-values and confidence intervals
- AVOID, unless you have a good reason not to

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

Strategies to deal with missing data

- Prevention
- Ad-hoc methods, e.g., single imputation, complete cases
- **Weighting methods**
- Likelihood methods, EM-algorithm
- Multiple imputation

Weighting

- Take the complete cases
- Re-weight any statistic to the distribution of the covariates in the population
- Advantages
  - Simple (one set of weights for all incomplete variables)
  - In SPSS: WEIGHT command
  - Reduces bias under MAR assumption
  - Standard methodology in official statistics
- Disadvantages
  - Discards data, increases the variance
  - Weights may not be available
  - Needs special variance estimators
  - Limited to unit non-response

Strategies to deal with missing data

- Prevention
- Ad-hoc methods, e.g., single imputation, complete cases
- Weighting methods
- **Likelihood methods, EM-algorithm**
- Multiple imputation

Maximum likelihood

- EM: Expectation-Maximization algorithm
- Direct ML
- Full Information Maximum Likelihood (FIML)
- Iterative methods to estimate parameters that effectively ignore the missing data
- Advantages:
  - Optimizes likelihood calculation directly
  - Many applications, widely accepted
  - Theoretically grounded
  - Easy to apply (when there is software)
- Disadvantages
  - Local minima, slow convergence
  - Difficult to apply outside standard models

Maximum likelihood software

- Mixed models: Proc Mixed (SAS), MLWin
- Structural models: AMOS, Mplus, Mx
- Rasch analyse: RUMM2030

Strategies to deal with missing data

- Prevention
- Ad-hoc methods, e.g., single imputation, complete cases
- Weighting methods
- Likelihood methods, EM-algorithm
- **Multiple imputation**

Multiple imputation

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

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

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
- Listwise deletion up to 5% of missing data per variable
- Weighting and likelihood methods are generally valid, but may be complex
- Multiple imputation is an all-round general purpose method

Multiple imputation, univariate - MIMP C

Overview C

- General idea of multiple imputation
- Statistical inference on multiply-imputed data
- Creating univariate imputations
- How to evaluate imputation methods
- Drawing from the observed data
- Categorical and other variable types

Overview C

- **General idea of multiple imputation**
- Statistical inference on multiply-imputed data
- Creating univariate imputations
- How to evaluate imputation methods
- Drawing from the observed data
- Categorical and other variable types

Multiple imputation

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

Acceptance of multiple imputation

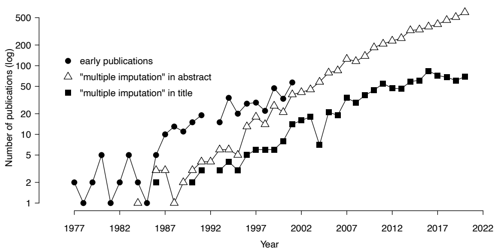
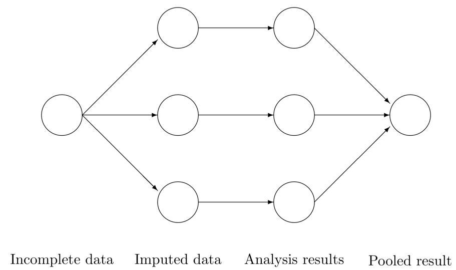


Figure 1: Source: Scopus (May 27, 2021)

Multiple imputation steps



Estimand

- Q is a quantity of scientific interest in the population.
- Q can be a vector of population means, population regression weights, population variances, and so on.
- Q may not depend on the particular sample, thus Q cannot be a standard error, sample mean, p-value, and so on.

Goal of multiple imputation

- Estimate Q by $\hat{Q}$ or $\bar{Q}$ accompanied by a valid estimate of its uncertainty.
- What is the difference between $\hat{Q}$ or $\bar{Q}$?
  - $\hat{Q}$ and $\bar{Q}$ both estimate Q
  - $\hat{Q}$ accounts for the sampling uncertainty
  - $\bar{Q}$ accounts for the sampling **and** missing data uncertainty

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

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

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.

Degrees of freedom (1)

With missing data, n is effectively lower. Thus, the degrees of freedom in statistical tests need to be adjusted.

The *old* formula assumes $n = \infty$:

$$\begin{aligned} \nu_{\text{old}} &= (m-1)\left(1+\frac{1}{r^2}\right) \\ &= \frac{m-1}{\lambda^2} \end{aligned} \tag{2}$$

Degrees of freedom (2)

The new formula is

$$\nu = \frac{\nu_{\text{old}}\nu_{\text{obs}}}{\nu_{\text{old}}+\nu_{\text{obs}}}. \tag{3}$$

where the estimated observed-data degrees of freedom that accounts for the missing information is

$$\nu_{\text{obs}} = \frac{\nu_{\text{com}}+1}{\nu_{\text{com}}+3}\nu_{\text{com}}(1-\lambda). \tag{4}$$

with $\nu_{\text{com}} = n - k$.

Overview C

- General idea of multiple imputation
- **Statistical inference on multiply-imputed data**
- Creating univariate imputations
- How to evaluate imputation methods
- Drawing from the observed data
- Categorical and other variable types

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?

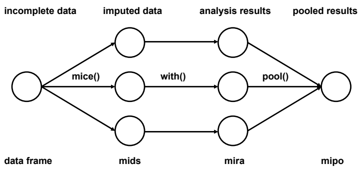
Classic advice: $m = 3, 5, 10$. More recently: set m higher: 20–100.

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.

Example of imputation-analysis-pooling steps

Multiple imputation 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
```

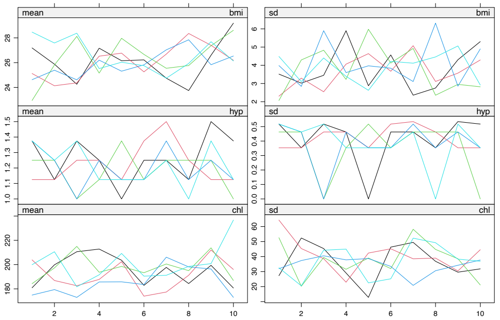
Inspect missing data pattern

```
md.pattern(nhanes, plot = FALSE)
```

```
   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
```

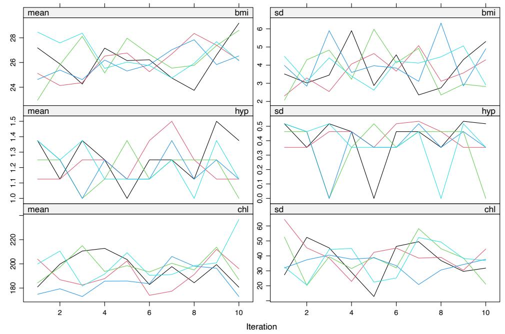
Multiply impute the missing data

```
imp <- mice(nhanes, print = FALSE, maxit = 10, seed = 1)
plot(imp)
```



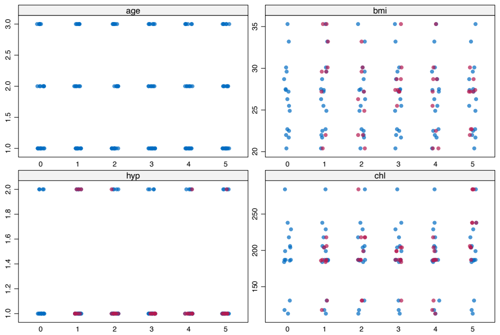
Inspect the tracelines for convergence

```
plot(imp)
```



Stripplot of observed and imputed data

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



Analyse and pool

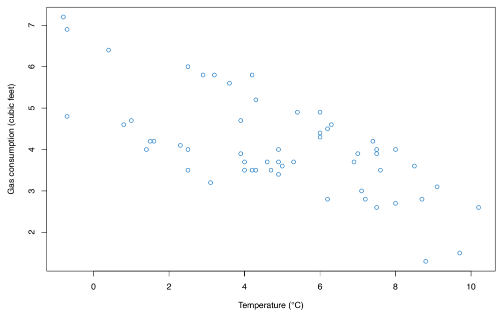
```
fit <- with(imp, lm(bmi ~ age))
est <- pool(fit)
summary(est)
```

```
         term estimate std.error statistic   df  p.value
1 (Intercept)    30.40      2.25     13.49 10.4 6.34e-08
2         age    -2.02      1.11     -1.83 12.6 9.09e-02
```

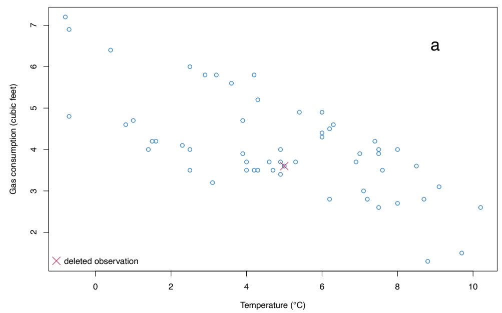
Overview C

- General idea of multiple imputation
- Statistical inference on multiply-imputed data
- **Creating univariate imputations**
- How to evaluate imputation methods
- Drawing from the observed data
- Categorical and other variable types

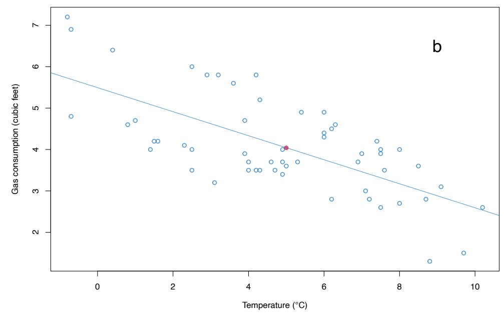
Relation between temperature and gas consumption



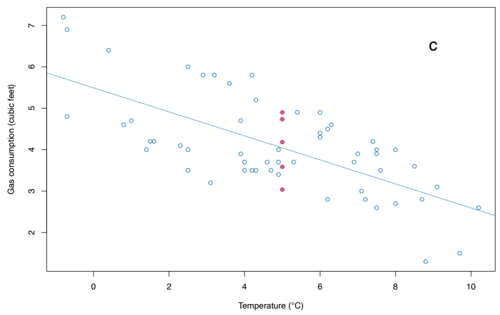
We delete gas consumption of observation 47



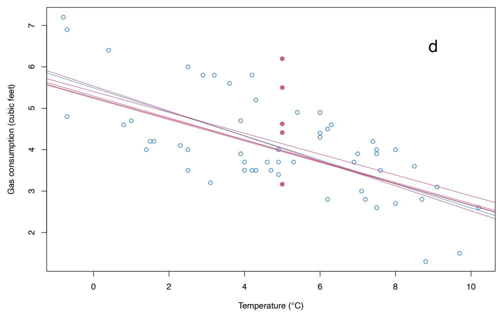
Predict imputed value from regression line



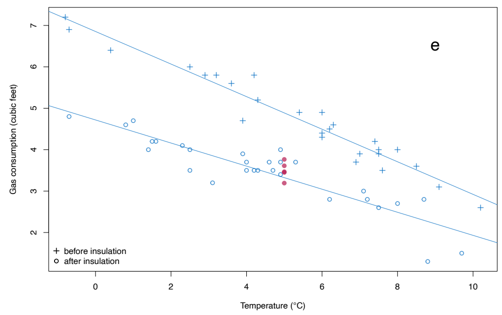
Predicted value + noise



Predicted value + noise + parameter uncertainty



Imputation based on two predictors



Overview C

- General idea of multiple imputation
- Statistical inference on multiply-imputed data
- Creating univariate imputations
- **How to evaluate imputation methods**
- Drawing from the observed data
- Categorical and other variable types

How to evaluate imputation methods

- https://stefvanbuuren.name/fimd/sec-evaluation.html
- Four evaluation criteria
- Example code

How to evaluate imputation methods: bias

- *Raw bias (RB) and percent bias (PB).*
- The raw bias of the estimate $\bar{Q}$ is defined as the difference between the expected value of the estimate and truth: $\mathrm{RB} = \mathrm{E}(\bar{\mathrm{Q}}) - \mathrm{Q}$.
- RB should be close to zero.
- Bias can also be expressed as percent bias: $\mathrm{PB} = 100 \times |(\mathrm{E}(\bar{\mathrm{Q}}) - \mathrm{Q})/\mathrm{Q}|$.
- For acceptable performance we use an upper limit for PB of 5%.

How to evaluate imputation methods: coverage

- *Coverage rate (CR).*
- The coverage rate (CR) is the proportion of confidence intervals that contain the true value. The actual rate should be equal to or exceed the nominal rate. If CR falls below the nominal rate, the method is too optimistic, leading to false positives.
- A CR below 90 percent for a nominal 95 percent interval indicates poor quality.
- A high CR (e.g., 0.99) may indicate that confidence interval is too wide, so the method is inefficient and leads to inferences that are too conservative.
- Inferences that are "too conservative" are generally regarded a lesser sin than "too optimistic".

How to evaluate imputation methods: efficiency

- *Average width (AW)*
- The AW of the confidence interval is an indicator of statistical efficiency.
- The length should be as small as possible, but not so small that the CR will fall below the nominal level.

How to evaluate imputation methods: RMSE

- *Root mean squared error (RMSE).* The $\mathrm{RMSE} = \sqrt{(\mathrm{E}(\bar{\mathrm{Q}}) - \mathrm{Q})^2}$ is a compromise between bias and variance, and evaluates $\bar{Q}$ on both accuracy and precision.
- The RMSE is widely used in machine learning and data.
- Less useful to evaluate multiple imputation methods.

What can go wrong with the RMSE?

Suppose we measure the average discrepancy between the true and imputed values by the RMSE:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n_{\mathrm{mis}}}\sum_{i=1}^{n_{\mathrm{mis}}}(y_i^{\mathrm{mis}} - \dot{y}_i)^2} \quad (5)$$

- Minimizing this criterion alone selects methods that ignore the uncertainty of the prediction.
- Amplifies the relations between the data and leads to too optimistic *P*-values.
- Except in trivial cases, imputation methods cannot reconstruct the true data!
- Bottom line: **Do not use this RMSE**

Four techniques for normal data

1. *Predict*: $\dot{y} = \hat{\beta}_0 + X_{\mathrm{mis}}\hat{\beta}_1$ (`mice.impute.norm.predict()`)
2. *Predict + noise*: $\dot{y} = \hat{\beta}_0 + X_{\mathrm{mis}}\hat{\beta}_1 + \dot{\epsilon}$ (`mice.impute.norm.nob()`)
3. *Bayesian multiple imputation*: $\dot{y} = \dot{\beta}_0 + X_{\mathrm{mis}}\dot{\beta}_1 + \dot{\epsilon}$, where $\dot{\beta}_0$, $\dot{\beta}_1$ and $\dot{\sigma}$ are random draws from their posterior distribution (`mice.impute.norm()`)
4. *Bootstrap multiple imputation*: $\dot{y} = \dot{\beta}_0 + X_{\mathrm{mis}}\dot{\beta}_1 + \dot{\epsilon}$, where $\dot{\beta}_0$, $\dot{\beta}_1$ and $\dot{\sigma}$ are the least squares estimates calculated from a bootstrap sample taken from the observed data (`mice.impute.norm.boot()`)

Simulation results for four normal methods (missing x)

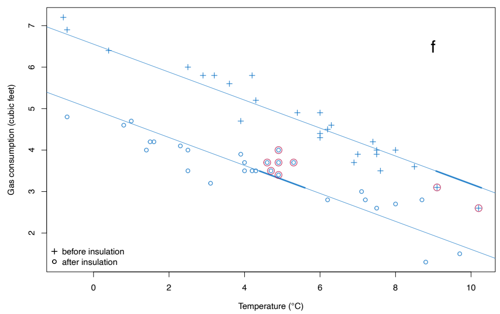
Method	Bias	% Bias	Coverage	CI Width	RMSE
`norm.predict`	-0.1007	34.7	0.359	0.160	0.118
`norm.nob`	0.0006	0.2	0.924	0.202	0.056
`norm`	0.0075	2.6	0.955	0.254	0.058
`norm.boot`	-0.0014	0.5	0.946	0.238	0.058
Listwise deletion	-0.0001	0.0	0.946	0.251	0.063

- https://stefvanbuuren.name/fimd/sec-linearnormal.html#sec:perflin

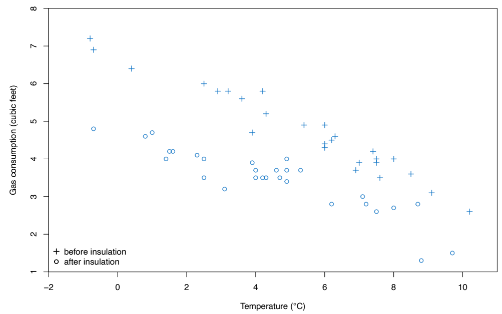
Overview C

- General idea of multiple imputation
- Statistical inference on multiply-imputed data
- Creating univariate imputations
- How to evaluate imputation methods
- **Drawing from the observed data**
- Categorical and other variable types

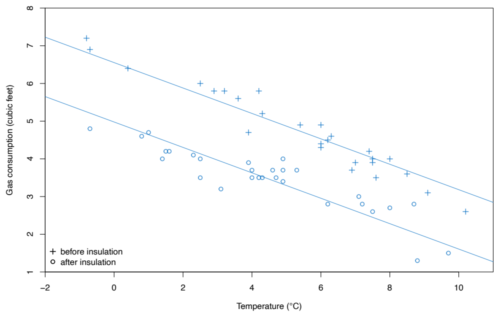
Drawing from the observed data



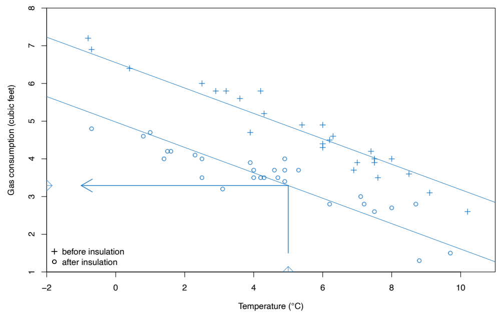
Predictive mean matching



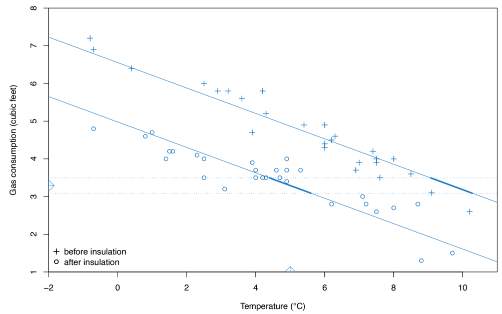
PMM: Add two regression lines



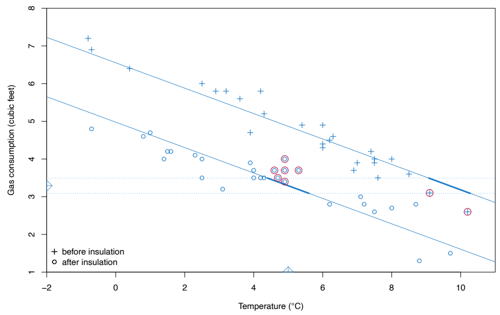
PMM: Predicted given 5°,C, 'after insulation'



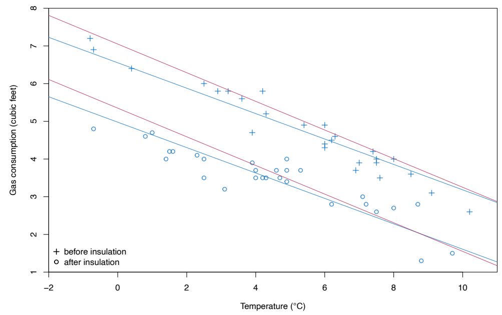
PMM: Define a matching range $\hat{y} \pm \delta$



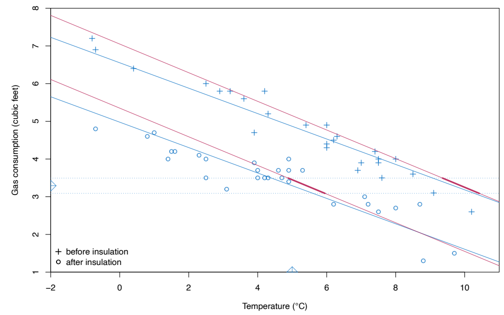
PMM: Select potential donors



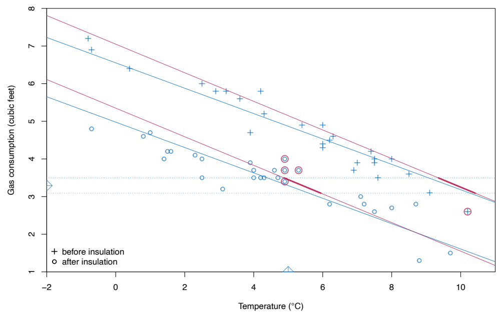
PMM: Bayesian PMM: Draw a line



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



PMM: Select potential donors



Overview C

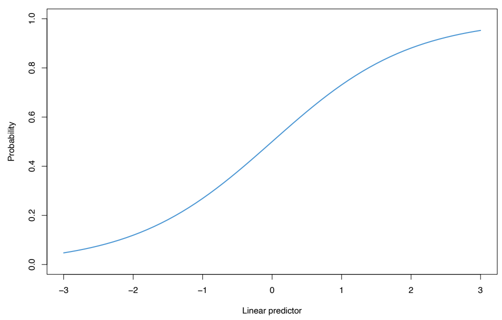
- General idea of multiple imputation
- Statistical inference on multiply-imputed data
- Creating univariate imputations
- How to evaluate imputation methods
- Drawing from the observed data
- **Categorical and other variable types**

Imputation of a binary variable

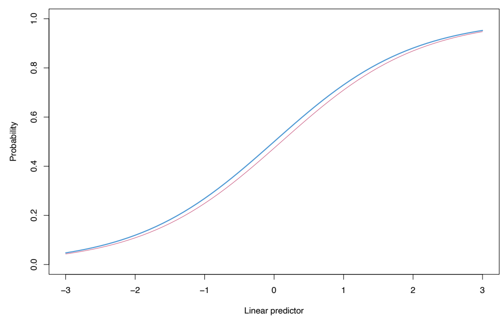
- Logistic regression

$$\Pr(y_i = 1 | X_i, \beta) = \frac{\exp(X_i \beta)}{1 + \exp(X_i \beta)}$$

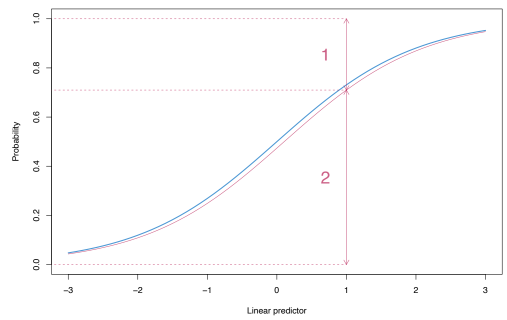
Fit logistic model



Draw parameter estimate



Read off the probability



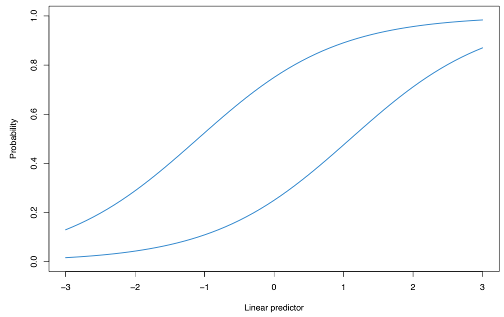
Impute ordered categorical variable

- K ordered categories $k = 1, \ldots, K$
- *ordered logit model*, or
- *proportional odds model*

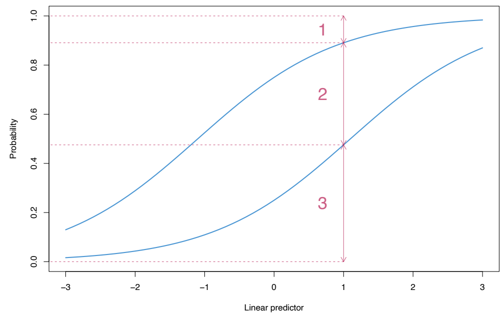
$$\Pr(y_i = k|X_i, \beta) = \frac{\exp(\tau_k + X_i\beta)}{\sum_{k=1}^{K} \exp(\tau_k + X_i\beta)}$$

-

Fit ordered logit model



Read off the probability



Built-in imputation functions

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

Multivariate imputation, MICE algorithm - MIMP E

Overview E

- Multivariate missing data
- Three imputation approaches
- MICE algorithm
- Assessment of convergence
- Compatibility

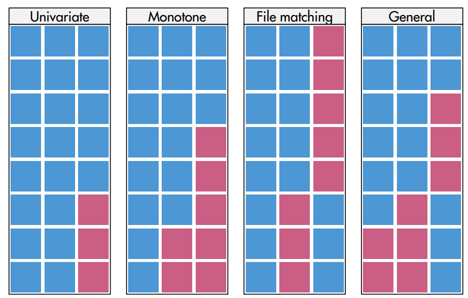
Overview E

- **Multivariate missing data**
- Three imputation approaches
- MICE algorithm
- Assessment of convergence
- Compatibility

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;
- Variables are often of different types (e.g., binary, unordered, ordered, continuous);
- Especially with large p and small n, collinearity or empty cells can occur;
- The ordering of the rows and columns can be meaningful, e.g., as in longitudinal data;
- The relation between Y_j and predictors Y_{-j} can be complex, e.g., nonlinear, or subject to censoring processes;
- Imputation can create impossible combinations, such as pregnant grandfathers.

Missing data patterns



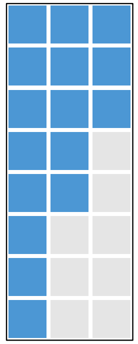
Overview E

- Multivariate missing data
- **Three imputation approaches**
- MICE algorithm
- Assessment of convergence
- Compatibility

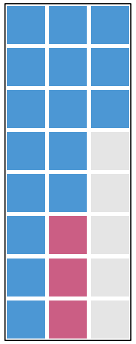
Three general strategies

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

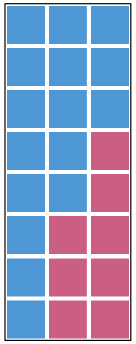
Monotone data imputation - 1



Monotone data imputation - 2



Monotone data imputation - 3



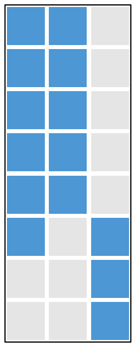
Monotone data imputation - Steps

1. Sort the data $Y_{j,\mathrm{obs}}$ with $j = 1, \dots, p$ according to their missingness.
2. Draw $\dot{\phi}_1 \sim P(Y_{1,\mathrm{obs}}|X)\}$
3. Impute $\dot{Y}_1 \sim P(Y_{1,\mathrm{mis}}|X, \dot{\phi}_1)$
4. Draw $\dot{\phi}_2 \sim P(Y_{2,\mathrm{obs}}|X, \dot{Y}_1)$
5. Impute $\dot{Y}_2 \sim P(Y_{2,\mathrm{mis}}, \dot{Y}_1, \dot{\phi}_2)\}$
6. $\vdots$

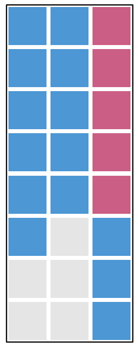
Monotone data imputation - Pro's and con's

- Pro's
  - Fast
  - Flexible
- Con's
  - Only possible for monotone pattern

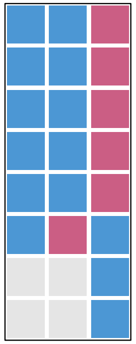
Joint modelling - 1



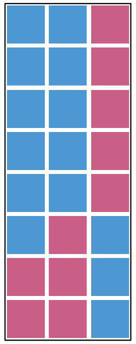
Joint modelling - 2



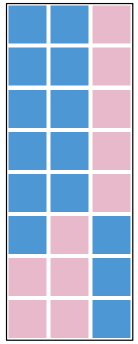
Joint modelling - 3



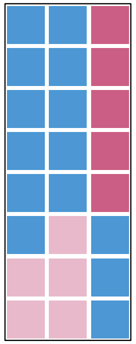
Joint modelling - 4



Joint modelling - next iteration - 5



Joint modelling - next iteration - 6



Joint modelling - Steps

1. Specify joint model $P(Y, X, R)$
2. Derive $P(Y_{\mathrm{mis}}|Y_{\mathrm{obs}}, X, R)$
3. Use MCMC techniques to draw imputations $\dot{Y}_{\mathrm{mis}}$

Joint modelling - Software

R/S Plus	`norm`, `cat`, `mix`, `pan`, `Amelia`, `jointAI`
SAS	`proc MI`, `proc MIANALYZE`
STATA	`MI` command
Stand-alone	`Amelia`, `solas`, `norm`, `pan`

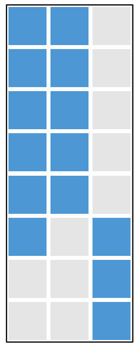
Joint modeling: Pro's

- Yield correct statistical inference under the assumed JM
- Efficient parametrization (if the model fits)
- Known theoretical properties
- Works very well for parameters close to the center
- Many applications

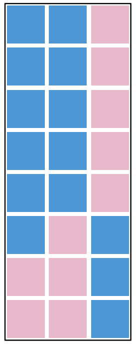
Joint Modeling: Con's

- Lack of flexibility
- May lead to large models
- Can assume more than the complete data problem
- Can impute impossible data

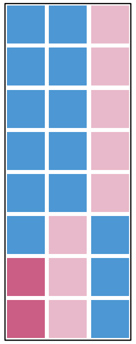
Fully conditional specification - 1



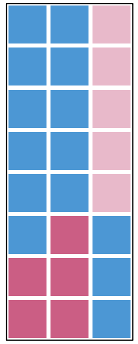
Fully conditional specification - 2



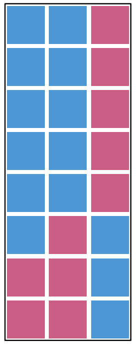
Fully conditional specification - 3



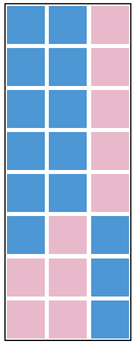
Fully conditional specification - 4



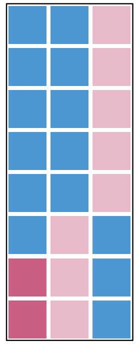
Fully conditional specification - 5



Fully conditional specification - next iteration - 6



Fully conditional specification - next iteration - 7



Overview E

- Multivariate missing data
- Three imputation approaches
- **MICE algorithm**
- Assessment of convergence
- Compatibility

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

Overview E

- Multivariate missing data
- Three imputation approaches
- MICE algorithm
- **Assessment of convergence**
- Compatibility

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

Imputations and Iterations

- Practitioners often confuse
  - number of imputations: m
  - number of iterations: M

Imputations and Iterations: m = number of imputations

- Y is our dataset with p incomplete variables
- Y_j is the jth incomplete variable with $j = 1, \ldots, p$
- Generate m complete versions of Y (and thus for each Y_j)
- We replace each missing value in every Y_j by m imputations
- Why: **to reflect the uncertainty of each missing value**
- In `mice`, the number of imputations is the `m` argument

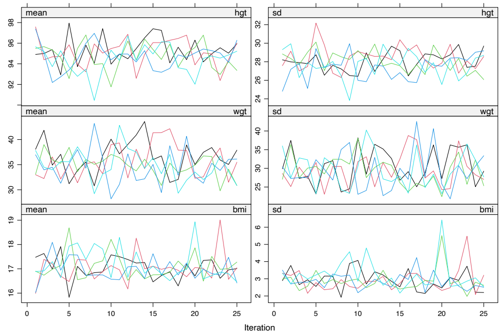
Imputations and Iterations: M = number of iterations

- Number of iterations M is the number of passes through the data matrix
- MICE overwrites the imputations from the previous iteration $t - 1$
- Why: **to reach convergence of imputation model**
- On a serial machine `mice` nests the m loop within the M loop
- In `mice`, the number of iterations is the `maxit` argument

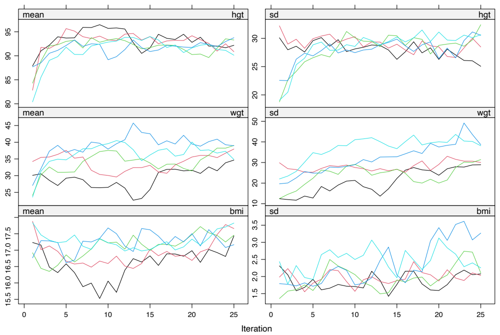
Convergence

- MICE is an iterative algorithm to solve the missingness problem
- MICE does **not** optimize a particular value. There is not a single quantity that we can monitor
- Rather: MICE convergences in distribution
- With simulation, we may stop after each iteration, and study statistical properties
- For many problems, properties do not change anymore after 10 iterations
- In practice, monitor the trace plot for deviant patterns

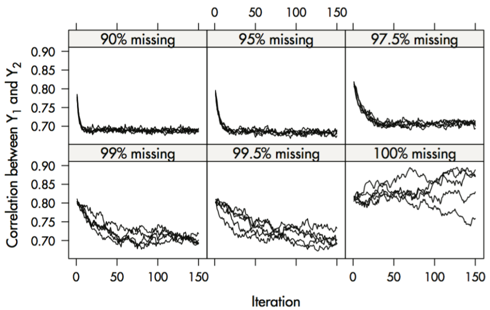
Convergence is usually fast



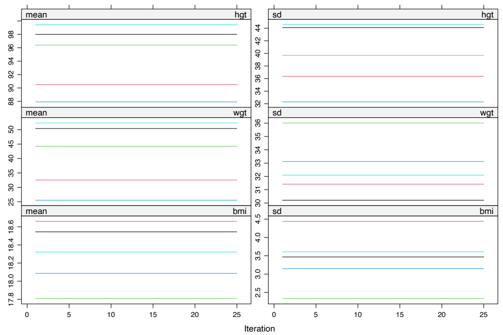
Convergence can be problematic



Convergence can be slow



Convergence can be pathological



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.

Overview E

- Multivariate missing data
- Three imputation approaches
- MICE algorithm
- Assessment of convergence
- **Compatibility**

Compatibility

Incompatibility

Compatibility of conditionals is a theoretical requirement for the Gibbs sampler

What happens if the model is clearly incompatible?

Simulation setup

- **Generate bivariate normal data, correlation 0.6**
- **Scientific interest on β in $Y_1 = \alpha + \beta Y_2 + \varepsilon$**
- **Generate 50% missing per variable, 75% incomplete cases (MCAR): three mechanisms**

Incomplete data - Stef van Buuren April 8, 2010 66

Compatibility

Three imputation models

MI compatible bivariate linear

- $Y_1^* \sim N(\phi + \theta Y_2, \sigma_2^2)$
- $Y_2^* \sim N(\gamma + \delta Y_1, \sigma_1^2)$ $\quad E(\theta / \sigma_2^2) = E(\delta / \sigma_1^2)$

MI incompatible quadratic

- $Y_1^* \sim N(\phi + \theta Y_2, \sigma_2^2)$
- $Y_2^* \sim N(\gamma + \delta Y_1^2, \sigma_1^2)$ $\quad E(\theta / \sigma_2^2) \neq E(\delta / \sigma_1^2)$

MI incompatible log

- $Y_1^* \sim N(\phi + \theta Y_2, \sigma_2^2)$
- $Y_2^* \sim N(\gamma + \delta \log(Y_1), \sigma_1^2)$ $\quad E(\theta / \sigma_2^2) \neq E(\delta / \sigma_1^2)$

Incomplete data - Stef van Buuren April 8, 2010 67

Compatibility

Simulation setup

- **Bivariate normal data, correlation 0.6**
- **Generate 50% missing per variable, 75% incomplete cases (MCAR)**
- **500 replications**
- **Scientific interest on β in $Y = \alpha + \beta X + \varepsilon$**

- **MICE implementation with a derived variable**
- $Y^* \sim N(\phi + \theta X, \sigma_X^2)$
- $Z = \log(Y)$ passive imputation
- $X^* \sim N(\gamma + \delta Z, \sigma_Y^2)$

Incomplete data - Stef van Buuren April 8, 2010 68

Compatibility

Mechanism	Missing data method	FMI	E(*b*)	Cov
	Theoretical values		0.600	95
MARRIGHT	Complete case analysis		0.597	93
	MI compatible linear	0.63	0.595	95
	MI incompatible quadratic	0.63	0.589	95
	MI incompatible log	0.64	0.582	95
MARMID	Complete case analysis		0.678	79
	MI compatible linear	0.75	0.613	94
	MI incompatible quadratic	0.75	0.601	94
	MI incompatible log	0.75	0.579	94
MARTAIL	Complete case analysis		0.556	78
	MI compatible linear	0.50	0.596	94
	MI incompatible quadratic	0.50	0.590	94
	MI incompatible log	0.50	0.590	95

Incomplete data - Stef van Buuren April 8, 2010 69

Compatibility

Incompatibility - conclusion

Compatibility of conditionals is a theoretical requirement for the Gibbs sampler

Unclear what exactly happens when the conditions is not met

- **Gibbs sampler may not converge**
- **results could depend on sequence of variables**

- **MICE appears to be robust against incompatibility, at least in the cases studied**
- **The incompatible model was superior to CCA for MARMID and MARTAIL mechanisms**
- **More work is needed**

Incomplete data - Stef van Buuren April 8, 2010 70

Recent developments: Compatibility

- Incompatible conditional models cannot provide imputations from any joint model
- However, multiple imputation using incompatible models is consistent as long as each conditional model was correctly specified (Liu 2013)
- Imputation models should closely model the data (Zhu 2015)

Compatibility and congeniality

- Compatibility: About relations among conditional distribution in the imputation model
- Congeniality: About relation between the imputation model and complete-data model
- https://stefvanbuuren.name/fimd/sec-FCS.html#sec:congeniality

Congeniality

- Imputation model should be more general than complete-data model (Meng, 1994)
- If not, imputer introduces restrictions to the later complete-data estimates

Recent development: Model-based imputation

- First choose complete-data model, then determine imputation model (Wu 2010, Bartlett 2015, Erler 2016)
- Create joint model for both complete-data model and imputation model
- Optimize imputations to reflect complete-data model relations
- Software: `smcfcs`, `mdmb`, `Blimp`
- Useful for strong, pre-specified complete-data models
- https://stefvanbuuren.name/fimd/sec-FCS.html#sec:modelbased

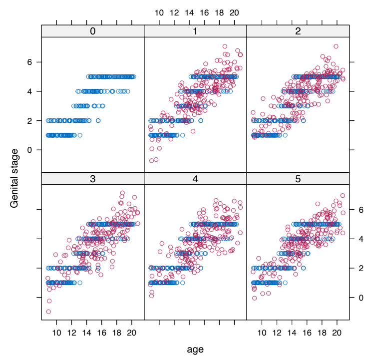
Joint model vs Fully conditional specification

- Fourth Dutch Growth Study 1997
- 22000 children between ages 0 and 21
- Tanner maturation stages
- Boys 8–21 years
- Genital development (5 stages)
- 42% missing data
- How does the probability per stage change with age?

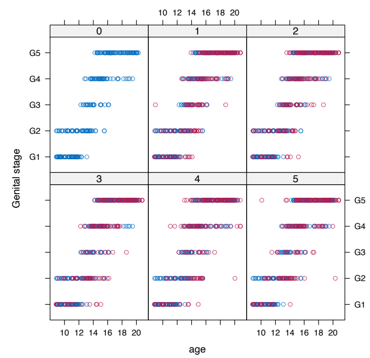
Imputation methods

- JM: multivariate normal
- JM: rounded
- FCS: predictive mean matching
- FCS: proportional odds model

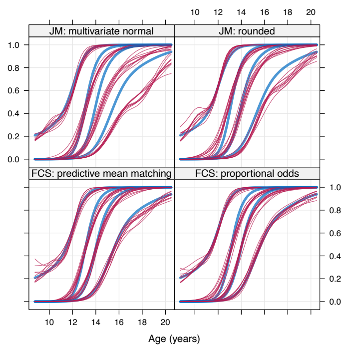
JM: Multivariate normal model



FCS: Proportional Odds model



JM vs FCS



Modelling choices, derived variables - MIMP G

Overview G

- Modelling choices
- Derived variables
- Diagnostics

Overview G

- **Modelling choices**
- Derived variables
- Diagnostics

How to set up the imputation model

1. MAR or MNAR
2. Form of the imputation model
3. Which predictors
4. Derived variables
5. What is m?
6. Order of imputation
7. Diagnostics, convergence

How to set up the imputation model

1. **MAR or MNAR**
2. Form of the imputation model
3. Which predictors
4. Derived variables
5. What is m?
6. Order of imputation
7. Diagnostics, convergence

When is the ignorability assumption suspect?

- If important variables that influence the probability to be missing are not available
- If there is reason to believe that responders differ from non-responders, even after accounting for the observed information
- If the data are censored, or below the detection limit

How to set up the imputation model

1. MAR or MNAR
2. Form of the imputation model
3. **Which predictors**
4. Derived variables
5. What is m?
6. Order of imputation
7. Diagnostics, convergence

Issues in thinking about the imputation model –> RECIPE

- We need to know about the **context of the problem**:

1. What will happen to the imputed data?
2. What do we know about the process that generated the missing data?
3. How well can be reconstruct the missing data from the observed data?

Issues in thinking about the imputation model - I

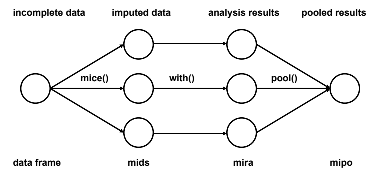
- We need to know about the **context of the problem**:

1. **What will happen to the imputed data?**
2. What do we know about the process that generated the missing data?
3. How well can be reconstruct the missing data from the observed data?

Imputation and Analysis model

- Statistical inference with with missing data involves two models:
  - Imputation model
  - Analysis model
- Address different aspect of the estimation problem

Multiple imputation in `mice`



Imputation and Analysis model: Imputation model

Imputation model

- The model we use to draw imputations
- Reflects our knowledge about the true (but unknown) values
- Technically: posterior predictive distribution of each missing entry

Imputation and Analysis model: Analysis model

Analysis model

- AKA: complete-data model, substantive model
- The model we use to estimate the parameters of scientific interest (Q)
- The model we would fit had the data been complete
- Technically: any model that estimates the thing we want to know

Imputation and Analysis model

- Are the imputation and analysis models entirely independent? **NO!**
- Imputation model should *more general* to the analysis model
- When this is true, Meng (1994) said that imputation and analysis models are *congenial*
- Take-home message: When creating imputed data,
  - imagine future analysis models applied to the imputed data sets
  - **extend imputation model to account for relations specified in the analysis model**

Imputation and Analysis model. Who's driving?

- Model-based imputation
  - First choose analysis model, then inform/derive the imputation model
  - When: If there is a strong scientific model
  - When: If you know that certain relations hold
- Data-based imputation
  - Use the observed data to impute the missing data, then do analyses
  - When: If there are multiple analysis models
  - When: If you are unsure about relation between variables
- Use both perspectives to improve imputation and analysis

Issues in thinking about the imputation model - II

- We need to know about the **context of the problem**:

1. What will happen to the imputed data?
2. **What do we know about the process that generated the missing data?**
3. How well can be reconstruct the missing data from the observed data?

Imputation model and Missing Data Model

- Missing Data Model = Missing Data Mechanism
  - Process that governs which Y's are observed and which Y's are unobserved (Rubin, 1976)
  - Sometimes we know this process (e.g.~experimental design, sampling)
- Default MICE assumes a Missing At Random (MAR) mechanism
  - Assumption: We can explain differences in response probability by the observed data
  - Implication (FIMD2, eq. 2.10): After conditioning on the observed data, the distribution of outcomes is the same for responders and non-responders
- Take-home message: When creating imputed data,
  - **extend imputation model with factors related to the missingness**

Issues in thinking about the imputation model - III

- We need to know about the **context of the problem**:

1. What will happen to the imputed data?
2. What do we know about the process that generated the missing data?
3. **How well can be reconstruct the missing data from the observed data?**

Predictability of the missing values

- Higher predictability means
  - more precise estimates
  - shorter confidence intervals
  - more powerful tests
  - fewer imputations (m) needed
- Higher predictability is beneficial, but "limited by nature"
- Social en medical data often do not predict well

Issues in thinking about the imputation model –> RECIPE

- We need to know about the **context of the problem**:

1. What will happen to the imputed data?
2. What do we know about the process that generated the missing data?
3. How well can be reconstruct the missing data from the observed data?

Which predictors to include? RECIPE

1. Include all variables that appear in the analysis model, including transformations and interactions
2. Include all variables that are related to the nonresponse
3. Include all variables that explain a considerable amount of variance
4. Remove variables that have too many missing values within the subgroup of incomplete cases

Functions `mice::quickpred()` and `mice::flux()`

https://stefvanbuuren.name/fimd/sec-modelform.html#sec:predictors

Overview G

- Modelling choices
- **Derived variables**
- Diagnostics

How to set up the imputation model

1. MAR or MNAR
2. Form of the imputation model
3. Which predictors
4. **Derived variables**
5. What is *m*?
6. Order of imputation
7. Diagnostics, convergence

Derived variables

- ratio of two variables
- sum score
- index variable
- quadratic relations
- interaction term
- conditional imputation
- compositions

Imputing a ratio

- Impute then transform (POST in FIMD1)
- Just another variable (JAV)
- Passive imputation
- Model-based imputation (new)

https://stefvanbuuren.name/fimd/sec-knowledge.html

Derived variables: summary

- Derived variables pose special challenges
- Plausible values should respect data dependencies
- If you can, create derived variables after imputation
- Best option: Probably model-based imputation
- More work needed to verify

Overview G

- Modelling choices
- Derived variables
- **Diagnostics**

Standard diagnostic plots in `mice`

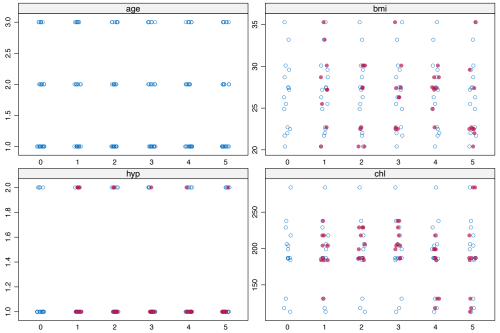
In general, inspect the overlap between red and blue points.

- One-dimensional scatter plot: `stripplot()`
- Box-and-whisker plot: `bwplot()`
- Densities: `densityplot()`
- Scattergram: `xyplot()`

Strip plot

```
library(mice)
imp <- mice(nhanes, seed = 29981, print = FALSE)
stripplot(imp, pch = c(1, 19))
```

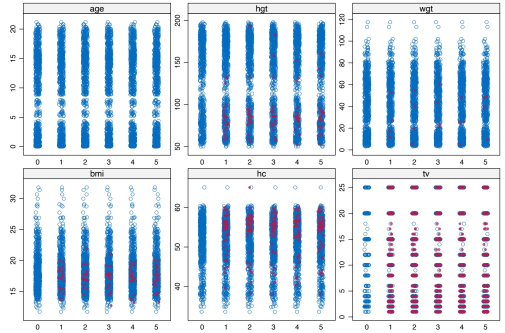
Strip plot



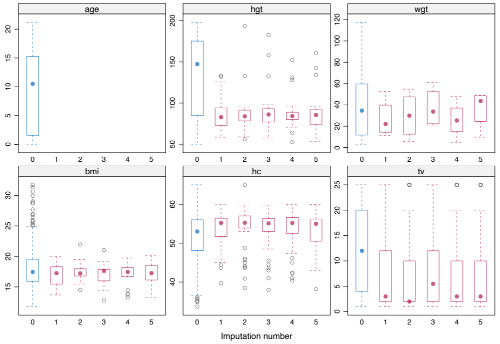
A larger dataset –> Use bwplot()

```r
imp <- mice(boys, seed = 24331, maxit = 1)
bwplot(imp)
```

A larger dataset –> Use bwplot()



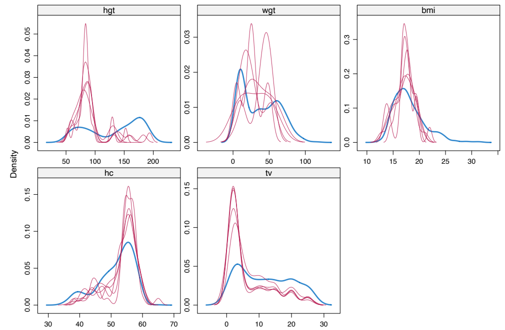
A larger dataset –> Use bwplot()



Density plot

```r
densityplot(imp)
```

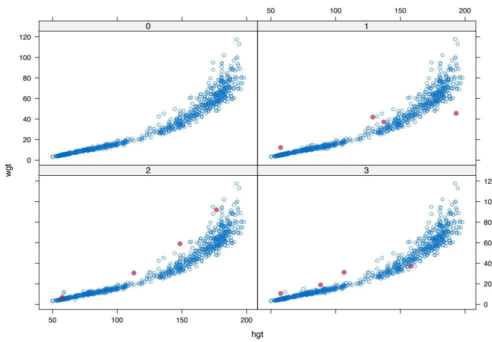
Density plot



Scatter plot

```r
imp <- mice(boys, seed = 24331, m = 3,
            maxit = 1, print = FALSE)
xyplot(imp, wgt ~ hgt | as.factor(.imp),
       pch = c(1, 20), cex = c(0.75, 1.5))
```

Scatter plot



Overview G

- Modelling choices
- Derived variables
- Diagnostics

Analysis of imputed data - MIMP I

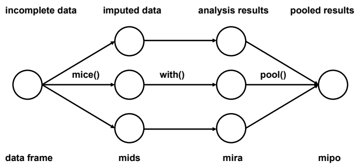
Overview I

- Workflows
- Pooling non-normal quantities
- Multi-parameter test
- Longitudinal data example

Overview I

- **Workflows**
- Pooling non-normal quantities
- Multi-parameter test
- Longitudinal data example

Multiple imputation in `mice`



Workflow 1: `mids` workflow using saved objects

```r
# mids workflow using saved objects
library(mice)
imp <- mice(nhanes, seed = 123, print = FALSE)
fit <- with(imp, lm(chl ~ age + bmi + hyp))
est1 <- pool(fit)
```

Workflow 2: `mids` workflow using pipes

```r
# mids workflow using pipes
library(magrittr)
est2 <- nhanes %>%
  mice(seed = 123, print = FALSE) %>%
  with(lm(chl ~ age + bmi + hyp)) %>%
  pool()
```

Workflow3: `mild` workflow using base::lapply

```r
# mild workflow using base::lapply
est3 <- nhanes %>%
  mice(seed = 123, print = FALSE) %>%
  mice::complete("all") %>%
  lapply(lm, formula = chl ~ age + bmi + hyp) %>%
  pool()
```

Workflow4: `mild` workflow using pipes and `base::Map`

```r
# mild workflow using pipes and base::Map
est4 <- nhanes %>%
  mice(seed = 123, print = FALSE) %>%
  mice::complete("all") %>%
  Map(f = lm, MoreArgs = list(f = chl ~ age + bmi + hyp)) %
  pool()
```

Workflow5: `mild` workflow using `purrr::map`

```r
# mild workflow using purrr::map
library(purrr)
est5 <- nhanes %>%
  mice(seed = 123, print = FALSE) %>%
  mice::complete("all") %>%
  map(lm, formula = chl ~ age + bmi + hyp) %>%
  pool()
```

Workflow6: long workflow using `base::by`

```
# long workflow using base::by
est6 <- nhanes %>%
  mice(seed = 123, print = FALSE) %>%
  mice::complete("long")  %>%
  by(as.factor(.$.imp), lm, formula = chl ~ age + bmi + hyp
  pool()
```

Workflow7: long workflow using a `dplyr` list-column

```
# long workflow using a dplyr list-column
library(dplyr)
est7 <- nhanes %>%
  mice(seed = 123, print = FALSE) %>%
  mice::complete("long") %>%
  group_by(.imp) %>%
  do(model = lm(formula = chl ~ age + bmi + hyp, data = .))
  as.list() %>%
  .[[-1]] %>%
  pool()
```

Not recommended: Average m imputed datasets

- Simple to do (for numeric data)
- One dataset to analyse
- Inherits all problems of single imputation
  - Ecological fallacy, e.g., overstates correlation
  - Biased parameter estimates
  - Wrong confidence intervals

Not recommended: Stack m imputed data sets

- Simple to do
- Weight each record by $1/m$
- One dataset to analyse
- Unbiased regression coefficients for linear models
- Inherits many problems of single imputation
  - Wrong confidence intervals, statistical test
  - Dubious for non-linear models

Overview I

- Workflows
- **Pooling non-normal quantities**
- Multi-parameter test
- Longitudinal data example

Pooling normal quantities

- Rubin (1987, p.~75) assumes normality of complete-data statistic
- Many statistics are approximately normally distributed, especially for large n
  - mean
  - standard deviation
  - regression coefficients
  - proportions
  - linear predictors
- Advice: Use Rubin's rules for such quantities

Pooling non-normal quantities

Table 3: Suggested transformations towards normality for various types of statistics. The transformed quantities can be pooled by Rubin's rules.

Statistic	Transformation	Source
Correlation	Fisher z	Schafer (1997)
Odds ratio	Logarithm	Agresti (1990)
Relative risk	Logarithm	Agresti (1990)
Hazard ratio	Logarithm	Marshall (2009)
Explained variance R^2	Fisher z on root	Harel (2009)
Survival probabilities	Complementary log-log	Marshall (2009)
Survival distribution	Logarithm	Marshall (2009)

Overview I

- Workflows
- Pooling non-normal quantities
- **Multi-parameter test**
- Longitudinal data example

Multi-parameter tests

- When?
  - Testing significance of set of variables
  - Testing significance of a categorical variable
  - If we only have test-statistics or P-values
- `D1` Multivariate Wald test
- `D2` Combined test statistics
- `D3` Likelihood ratio test
- https://stefvanbuuren.name/fimd/sec-multiparameter.html

Example: Test categorical variable age

```r
imp <- mice(nhanes2, m = 10, print = FALSE, seed = 71242)
m2 <- with(imp, lm(chl ~ age + bmi))
m1 <- with(imp, lm(chl ~ bmi))
summary(D1(m2, m1))
```

Example: Test categorical variable age

```
Models:
 model          formula
     1 chl ~ age + bmi
     2       chl ~ bmi

Comparisons:
   test statistic df1  df2 dfcom p.value   riv
 1 ~~ 2      5.02   2 11.9    21  0.0263 0.628

Number of imputations:  10   Method D1
```

D_1, D_2 or D_3?

- If you can, use D_1()
- Use D_2() if you have only the test statistics/P values, and with $m > 20$
- D_3() or D_1() are about equally good for samples $n > 200$

Overview I

- Workflows
- Pooling non-normal quantities
- Multi-parameter test
- **Longitudinal data example**

Longitudinal data example

- **Long** and **Wide** data
- Wide matrix feels most natural to applied researchers
- Wide matrix is suitable if data are observed at (approximately) equal time points
- Long matrix is expected by software designed for time-varying data
- Convert wide –> long: `tidyr::pivot_longer()`
- Convert long –> wide: `tidyr::pivot_wider()`
- https://stefvanbuuren.name/fimd/sec-longandwide.html

Longitudinal data imputation

- If you can, impute the **Wide** data
- Preserves relations over time
- Independence of row (persons)
- If you cannot, use multilevel imputation

SE Fireworks Disaster

Saturday, May 13 2000, Enschede

SE Fireworks Disaster

- 23 killed
- 950 injured
- 500 houses destroyed
- 1250 homeless
- 10000 evacuated
- post-traumatic stress

Roombeek now

Embedded randomized controlled trial

- Mediant
- EMDR: Eye Movement Desensitization and Reprocessing
- CBT: Cognitive Behavioral Therapy
- 2×26 children
  - T1: pre-treatment
  - T2: post-treatment (4–8 weeks)
  - T3: follow-up (3 months)
  - Outcome: UCLA PTSD Reaction Index (PTSD-RI)

Research questions

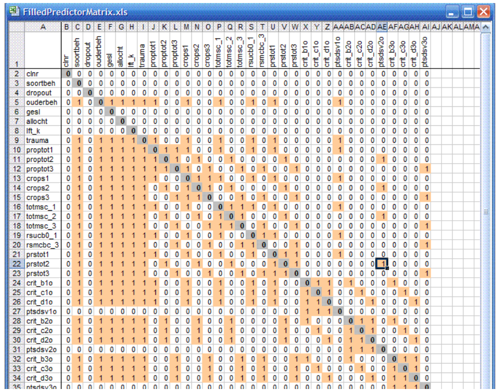
- Is one of these treatments more effective in reducing PTSD symptoms at T2 and T3?
- Does the number of sessions needed to produce the therapeutic effect differ between the treatments?

(Missing) Data

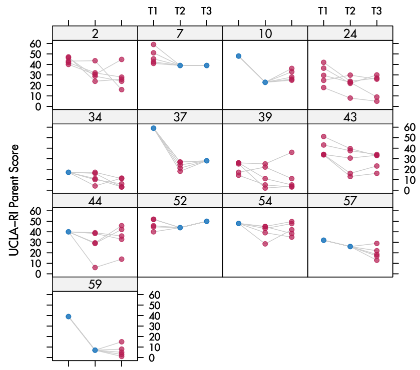
Table 9.1: SE Fireworks Disaster data. The UCLA PTSD Reaction Index of 52 subjects, children and parents, randomized to EMDR or CBT.

id	trt	pp	Y_1^c	Y_2^c	Y_3^c	Y_1^p	Y_2^p	Y_3^p	id	trt	pp	Y_1^c	Y_2^c	Y_3^c	Y_1^p	Y_2^p	Y_3^p
1	E	Y	–	–	–	36	35	38	32	E	N	28	17	8	40	42	33
2	C	N	45	–	–	–	–	–	33	E	N	–	–	–	38	22	25
3	E	N	–	–	–	13	19	13	34	E	N	–	–	–	17	–	–
4	C	Y	–	–	–	33	27	20	35	E	Y	50	20	–	19	1	5
5	E	Y	26	6	4	27	16	11	37	C	N	30	–	26	59	–	28
6	C	Y	8	1	2	32	15	13	38	C	Y	–	–	–	35	24	27
7	C	Y	41	26	31	–	39	39	39	E	N	–	–	–	–	–	–
8	C	N	–	–	–	24	13	35	40	E	Y	25	5	2	42	13	11
10	C	Y	35	27	14	48	23	–	41	E	Y	36	11	9	30	2	1
12	C	Y	28	15	13	45	33	36	43	E	N	17	–	–	–	–	–
13	E	Y	–	–	–	26	17	14	44	E	N	27	–	–	40	–	–
14	C	Y	33	8	9	37	7	3	45	C	Y	31	12	29	34	28	29
15	E	Y	43	–	7	25	27	1	46	C	Y	–	–	–	44	35	25
16	C	Y	50	8	35	39	21	34	47	C	Y	–	–	–	30	18	14
17	C	Y	31	21	10	32	21	19	48	E	Y	25	18	–	18	17	2
18	E	Y	30	17	16	47	28	34	49	C	N	24	23	16	44	29	34
19	E	Y	29	6	5	20	14	11	50	E	Y	31	13	9	34	18	13
20	E	Y	47	14	22	44	21	25	51	C	Y	–	–	–	52	13	13
21	C	Y	39	12	12	39	5	19	52	C	Y	30	35	28	–	44	50
23	C	Y	14	12	5	29	9	4	53	C	Y	19	33	21	36	21	21
24	E	N	27	–	–	–	–	–	54	C	N	43	–	–	48	–	–
25	E	Y	6	10	5	25	16	16	55	E	Y	64	42	35	44	31	16
28	C	Y	–	2	6	36	17	23	56	C	Y	–	–	–	37	6	9
29	E	Y	23	23	28	23	25	13	57	C	Y	31	12	–	32	26	–

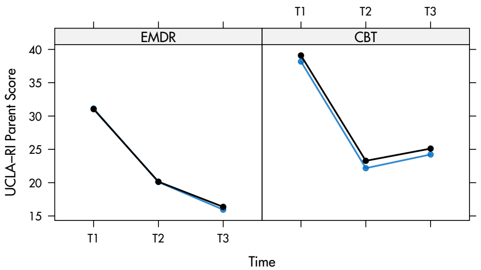
Predictor matrix for multiple imputation



Imputed Data



UCLA-RI Parent



Conclusion SE Firework Disaster

- More columns than rows –> careful `predictorMatrix` specification
- Preservation of longitudinal patterns
- Preservation of all children, as randomized
- Complete case analysis and multiple imputation led to same substantive result
- No significant difference between EMDR-CBT
- Potentially fewer sessions needed for EMDR

Longitudinal data: Conclusions

- Imputation should preserve
  - Group compositions across time
  - Relations within time
  - Relation across time
- If possible, code data in **Wide** form
- Codify predictor matrix to reflect data structure
- Use simple complete-data analysis: *t*-test, ANOVA, MANOVA

Sensitivity analysis, reporting - MIMP K

Overview K

- When to consider sensitivity analysis?
- Selection and pattern-mixture model
- Shift imputations
- Application to Leiden 85+ data
- Facilities in `mice`
- Reporting guidelines

Overview K

- **When to consider sensitivity analysis?**
- Selection and pattern-mixture model
- Shift imputations
- Application to Leiden 85+ data
- Facilities in `mice`
- Reporting guidelines

Relevance of ignorability assumption 1

Ignorability implies

$$P(Y|X, R = 0) = P(Y|X, R = 1) \tag{6}$$

so

$$P(Y_{\mathrm{obs}}|X) = P(Y_{\mathrm{mis}}|X) \tag{7}$$

In words: The way in which Y depends on X is the same for the observed and the missing data

Relevance of ignorability assumption 2

Consequence: We may use the relations in the observed data to create imputations for the missing data

Ignorability = the belief that the available data are sufficient to correct for the effects of the missing data

When is the ignorability assumption suspect?

- If important variables that govern the missing data process are not available
- If there is reason to believe that responders differ from non-responders, even after accounting for the observed information
- If the data are censored, or below the detection limit

Overview K

- When to consider sensitivity analysis?
- **Selection and pattern-mixture model**
- Shift imputations
- Application to Leiden 85+ data
- Facilities in `mice`
- Reporting guidelines

Models for nonignorable nonresponse

$P(Y, R)$ does not factorise into independent parts, and must be modelled jointly

Two approaches (there are some more):

- Selection model: $P(Y, R) = P(R|Y)P(Y)$
- Pattern mixture-model: $P(Y, R) = P(Y|R)P(R)$

Selection model

Selection model (Heckman, 1976) (Nobel prize Economics 2000)

$$P(Y, R|\psi, \theta) = P(R|Y, \psi)P(Y, \theta) \tag{8}$$

$P(R = 1|Y)$	response mechanism, selection function
$P(Y)$	(joint) distribution for the data

Assumption: $P(\psi, \theta) = P(\psi)P(\theta)$ distinct parameters

Selection model example

Table IV. Numerical example of an NMAR non-response mechanism, when there are more missing data for lower blood pressures

Y — Selection model

Class midpoint of Systolic BP (mmHg)	$p(R=0\|\mathrm{BP})$	$p(\mathrm{BP})$	$p(\mathrm{BP}\|R=1)$	$p(\mathrm{BP}\|R=0)$
100	0·35	0·02	0·01	0·06
110	0·30	0·03	0·02	0·07
120	0·25	0·05	0·04	0·10
130	0·20	0·10	0·09	0·16
140	0·15	0·15	0·15	0·19
150	0·10	0·30	0·31	0·25
160	0·08	0·15	0·16	0·10
170	0·06	0·10	0·11	0·05
180	0·04	0·05	0·05	0·02
190	0·02	0·03	0·03	0·00
200	0·00	0·02	0·02	0·00
Mean (mmHg)		150	151·6	138·6

Pattern mixture model

Pattern mixture-model (Rubin, 1977)

$$P(Y, R|\psi, \theta) = P(Y|R, \theta)P(R|\psi) \tag{9}$$

$P(Y|R = 1, \theta)$ (joint) distribution for the observed data
$P(Y|R = 0, \theta)$ (joint) distribution for the missing data
$P(R|\psi)$ response probability

Assumption: $P(\psi, \theta) = P(\psi)P(\theta)$ distinct parameters

Pattern mixture model example

Table IV. Numerical example of an NMAR non-response mechanism, when there are more missing data for lower blood pressures

Y — Mixture model

Class midpoint of Systolic BP (mmHg)	$p(R=0\|\mathrm{BP})$	$p(\mathrm{BP})$	$p(\mathrm{BP}\|R=1)$	$p(\mathrm{BP}\|R=0)$
100	0·35	0·02	0·01	0·06
110	0·30	0·03	0·02	0·07
120	0·25	0·05	0·04	0·10
130	0·20	0·10	0·09	0·16
140	0·15	0·15	0·15	0·19
150	0·10	0·30	0·31	0·25
160	0·08	0·15	0·16	0·10
170	0·06	0·10	0·11	0·05
180	0·04	0·05	0·05	0·02
190	0·02	0·03	0·03	0·00
200	0·00	0·02	0·02	0·00
Mean (mmHg)		150	151·6	138·6

Pattern mixture and selection models are related

- Selection to PM: $P(Y|R) = \frac{P(R|Y)P(Y)}{P(R)}$
- PM to selection: $P(R|Y) = \frac{P(Y|R)P(R)}{P(Y)}$

Sensitivity analysis as a substitute for ignorability

MAR $P(Y|X, R = 0) = P(Y|X, R = 1)$
MNAR $P(Y|X, R = 0) \neq P(Y|X, R = 1)$

The problem: The data contain no information about $P(Y|X, R = 0)$.

The solution: Specify a range of plausible imputation models, and study the influence on the outcomes

Models for $R = 0$ and $R = 1$ are different

Overview K

- When to consider sensitivity analysis?
- Selection and pattern-mixture model
- **Shift imputations**
- Application to Leiden 85+ data
- Facilities in `mice`
- Reporting guidelines

A simple model to shift imputations

Specify $P(Y|X, R)$

Model		
1	$Y = X\beta + \epsilon$	β is estimated from cases $R = 1$
2	$Y = X\beta + \delta + \epsilon$	imputations applied to $R = 0$

Combined formulation: $Y = X\beta + (1 - R)\delta + \epsilon$

δ cannot be estimated, and must be chosen by the user

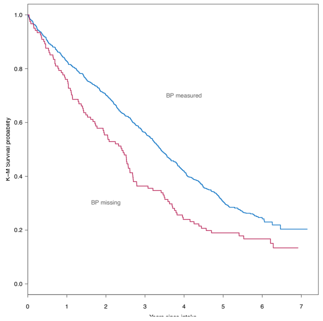
Overview K

- When to consider sensitivity analysis?
- Selection and pattern-mixture model
- Shift imputations
- **Application to Leiden 85+ data**
- Facilities in `mice`
- Reporting guidelines

Application

- Leiden 85+ cohort study
- N=1236, 85+ on Dec. 1, 1986
- N=956 were visited (1987-1989)
- BP is missing for 121 patients
- Do anti-hypertensive drugs shorten life in the oldest old?
- Scientific interest: Mortality risk as function of BP and age

Survival probability by response group



Why sensitivity analysis?

From the data we see - Those with no BP measured die earlier - Those that die early and that have no hypertension history have fewer BP measurements

Thus, imputations of BP under MAR could be too high values.

We need to lower the imputed values of BP, and study the influence on the outcome

How to specify δ?

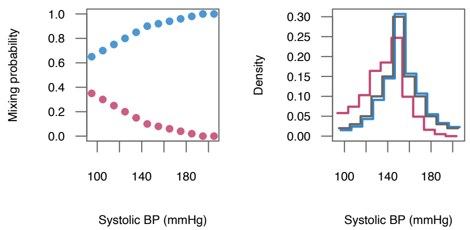
- Combined formulation: $Y = X\beta + (1 - R)\delta + \epsilon$
- δ cannot be estimated, and must be chosen by the user

Both models

Table IV. Numerical example of an NMAR non-response mechanism, when there are more missing data for lower blood pressures

Y	Selection model		Mixture model	
Class midpoint of Systolic BP (mmHg)	$p(R=0 \mid BP)$	$p(BP)$	$p(BP \mid R=1)$	$p(BP \mid R=0)$
100	0·35	0·02	0·01	0·06
110	0·30	0·03	0·02	0·07
120	0·25	0·05	0·04	0·10
130	0·20	0·10	0·09	0·16
140	0·15	0·15	0·15	0·19
150	0·10	0·30	0·31	0·25
160	0·08	0·15	0·16	0·10
170	0·06	0·10	0·11	0·05
180	0·04	0·05	0·05	0·02
190	0·02	0·03	0·03	0·00
200	0·00	0·02	0·02	0·00
Mean (mmHg)		150	151·6	138·6

Effect of response mechanism on BP



Overview K

- When to consider sensitivity analysis?
- Selection and pattern-mixture model
- Shift imputations
- Application to Leiden 85+ data
- **Facilities in `mice`**
- Reporting guidelines

How to impute under MNAR?

- Determine sensitivity parameters (delta)
- https://stefvanbuuren.name/fimd/sec-nonignorable.html

How to impute under MNAR?

- Post-process imputations (deduct delta)
- https://stefvanbuuren.name/fimd/sec-sensitivity.html

`mice` functions

- Estimating δ by the random indicator method (Jolani 2012): `mice.impute.ri()`
  - Iterative method that redraws the missing data indicator under a selection model
- Not-at-random fully conditional specification (NARFCS) to specify non-ignorable adjustments to imputation models
  - `mice.impute.mnar.norm()` for normal data
  - `mice.impute.mnar.logreg()` for binary data

General advice on MNAR

- Include as much data as possible in the imputation model
- State why the ignorability assumption is suspect
- Limit the possible non-ignorable alternatives

Overview K

- When to consider sensitivity analysis?
- Selection and pattern-mixture model
- Shift imputations
- Application to Leiden 85+ data
- Facilities in `mice`
- **Reporting guidelines**

Reporting guidelines

- Amount of missing data
- Reasons for missingness
- Differences between complete and incomplete data
- Method used to account for missing data
- Software
- Number of imputed datasets
- Imputation model
- Derived variables
- Diagnostics
- Pooling
- Listwise deletion
- Sensitivity analysis
- https://stefvanbuuren.name/fimd/sec-reporting.html#sec:guidelines