Multiple Imputation in Practice (S28)

Stef van Buuren[1,2]
Gerko Vink[2]

[1]Netherlands Organization for Applied Scientific Research TNO, Leiden

[2]Methodology and Statistics, FSBS, Utrecht University

Utrecht, July 22-25, 2019

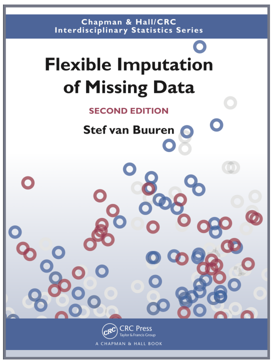
Why this course

- 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 of the course: get comfortable with a modern and powerful way of solving missing data problems

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.
- Van Buuren, S. (2018). Flexible Imputation of Missing Data. Second Edition. Chapman & Hall/CRC, Boca Raton, FL.

Flexible Imputation of Missing Data. Second Edition



Flexible Imputation of Missing Data. Second Edition.

- Published: July 16, 2018
- ISBN 9781138588318
- Full text: `https://stefvanbuuren.name/fimd`
- Book ordering: `https://www.crcpress.com/Flexible-Imputation-of-Missing-Data-Second-Edition/Buuren/p/book/9781138588318`

R

- Why R?

R software and examples

- R Install from `http://cran.r-project.org/package=mice`
- R package: `mice 3.6.0`
- Development version: https://github.com/stefvanbuuren/mice
- Documentation: `https://stefvanbuuren.github.io/mice/`
- Example code: `https://github.com/stefvanbuuren/fimdbook/blob/master/R/fimd.R`
- This course: `https://www.gerkovink.com/mimp/`

Time slots

Date	Lecture 9:00–10:30	Practical 10:45–12:15	Lecture 13:15–14:30	Practical 14.45–16.00
Mon Jul 22	**A**	**B**	**C**	**D**
Tue Jul 23	**E**	**F**	**G**	**H**
Wed Jul 24	**I**	**J**	**K**	**L**
Thu Jul 25	**M**	**N**	**O**	**P**

Room: Koningsberger Cosmos (1.26)

Schedule for Monday, Jul 22

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

Schedule for Tuesday, Jul 23

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

Schedule for Wednesday, Jul 24

Slot	Type	Description	FIMD2
I	L	Combining inferences	
J	P	Analysis in R	
K	L	Sensitivity analysis	3.8, 9.2, 10.2
L	P	Approach to sensitivity analysis	`leiden85`

Schedule for Thursday, Jul 25

Slot	Type	Description	FIMD2
M	L	RMultilevel data	12.2, 7.2-7.10
N	P	Multilevel data sets in `mice`	`popmis`
O	L	Capita selecta	
P	P	Get advice/support	

PART A

Why are missing data interesting?

- "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 problems are the heart of statistics

Causes of missing data

- Respondent skipped the item
- Data transmission/coding error
- Drop out in longitudinal research
- Refusal to cooperate
- Sample from population
- Question not asked, different forms
- Branching, routing
- Censoring

Consequences of missing data

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

In general, missing data can severely complicate interpretation and analysis.

Some notation

- Y random variable with missing data
- Y_{obs} observed values of Y
- Y_{mis} missing values of Y
- R response indicator
- $R = 1$ if Y is observed
- $R = 0$ if Y is missing
- X and R are complete covariates

MCAR, MAR, MNAR

- MCAR: Missing Completely At Random
- MAR: Missing At Random
- MNAR: Missing Not At Random

MCAR, MAR, MNAR

- MCAR: The probability to be missing is constant for all units
- MAR: The probability to be missing depends on observed data
- MNAR: The probability to be missing depends on unobserved data

MCAR: Missing Completely at Random

- Probability to be missing is not related to any data
- $P(R|Y_{\text{obs}}, Y_{\text{mis}}, X, \psi) = P(R|Y_{\text{obs}}, \psi)$
- 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
- $P(R|Y_{\text{obs}}, Y_{\text{mis}}, X, \psi)$ does not simplify
- Examples
  - Income, without covariates related to income
  - Body weight report

When may we 'ignore' the missing data?

- If the data are MAR
- and if the parameters of the missing data mechanism and the complete data model are a priori independent
- then the likelihood factors into two independent parts, so we need to study only $f(Y_{\text{obs}}|X, \theta)$.

- 'ignorable' means: The observed data are sufficient to account for differences in the missing data probability.

Strategies to deal with missing data

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

Listwise deletion

- Analyze only the complete records
- Also known as Complete Case Analysis (CCA)
- Advantages
  - Simple (default in most software)
  - Unbiased under MCAR
  - Correct standard errors, significance levels Two special properties in regression

Listwise deletion

- Disadvantages
  - Wasteful
  - Large standard errors
  - Biased under MAR, even for simple statistics like the mean
  - Inconsistencies in reporting

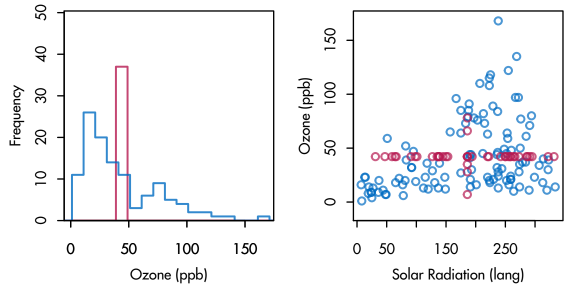
Listwise deletion: Special properties

- For any regression with missing in X, estimates under CCA are unbiased as long as the missingness does not depend on Y. Even some MNAR cases (Glynn Laird, 1986; Little 1992).
- In logistic regression: With missing in Y or X (but not both), estimates under CCA 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.
- FIMD2 2.7

Mean imputation

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

Mean imputation



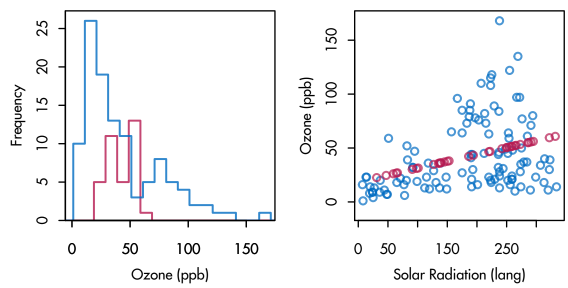
Mean imputation

- 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
  - Unbiased estimates of regression coefficients (under MAR)
  - Good approximation to the (unknown) true data if explained variance is high
- Prediction is the favorite among non-statisticians

Regression imputation



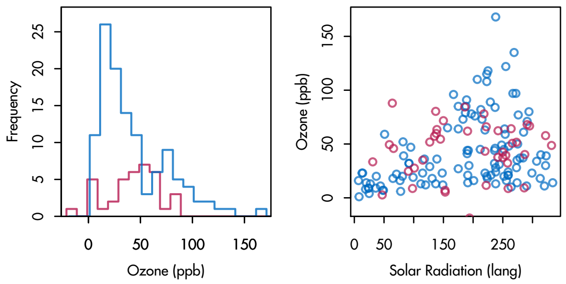
Regression imputation

- 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

Table: Overview of assumptions made by simple methods

	unbiasedness			standard error
	mean	reg weight	correlation	
listwise deletion	MCAR	MCAR	MCAR	too large
pairwise deletion	MCAR	MCAR	MCAR	complicated
mean imputation	MCAR	–	–	too small
regression imp	MAR	MAR	–	too small
stochastic imp	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 or complete cases
- Weighting methods
- Likelihood methods, EM-algorithm
- Multiple imputation

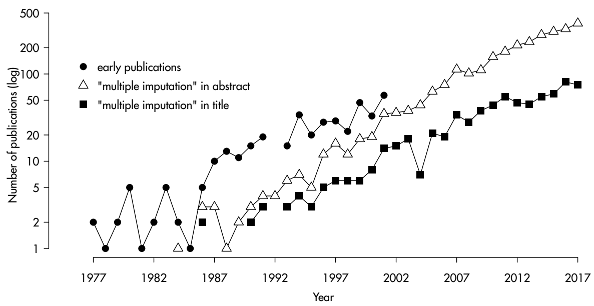
PART C

Slot C: multiple imputation

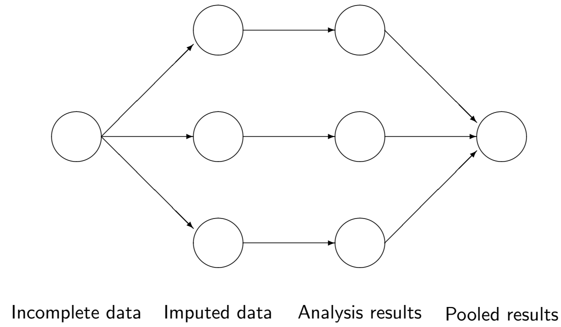
- Theory of multiple imputation
- Univariate imputation
  - General idea
  - Predictive mean matching
  - Binary outcomes
  - Ordered and unordered outcomes

FIMD Sections Ch2, 3.1–3.7

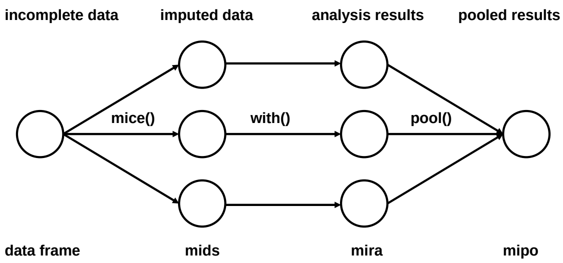
Rising popularity of multiple imputation



Main steps used in multiple imputation



Steps in `mice`



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 and is represented as a $k \times 1$ column vector

The pooled estimate $\bar{Q}$ is simply the average

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

Within-imputation variance

Average of the complete-data variances as

$$\bar{U} = \frac{1}{m}\sum_{\ell=1}^{m} \bar{U}_\ell, \tag{2}$$

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

The within-imputation variance is large if the sample is small

Between-imputation variance

Variance between the m complete-data estimates is given by

$$B = \frac{1}{m-1}\sum_{\ell=1}^{m}(\hat{Q}_\ell - \bar{Q})(\hat{Q}_\ell - \bar{Q})', \tag{3}$$

where $\bar{Q}$ is the pooled estimate (c.f. equation 1)
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

$$\begin{aligned} T &= \bar{U} + B + B/m \\ &= \bar{U} + \left(1 + \frac{1}{m}\right) B \end{aligned} \tag{4}$$

for the total variance of $\bar{Q}$, 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 $\bar{Q}$ itself is based on finite m.

Variance ratio's (1)

Proportion of the variation attributable to the missing data

$$\lambda = \frac{B + B/m}{T}, \tag{5}$$

Relative increase in variance due to nonresponse

$$r = \frac{B + B/m}{\bar{U}} \tag{6}$$

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

Variance ratio's (2)

Fraction of information about Q missing due to nonresponse

$$\gamma = \frac{r + 2/(\nu + 3)}{1 + r} \tag{7}$$

This measure needs an estimate of the degrees of freedom ν.

Relation between γ and λ

$$\gamma = \frac{\nu + 1}{\nu + 3}\lambda + \frac{2}{\nu + 3}. \tag{8}$$

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 $n = \infty$:

$$\begin{aligned} \nu_{\text{old}} &= (m-1)\left(1 + \frac{1}{r^2}\right) \\ &= \frac{m-1}{\lambda^2} \end{aligned} \tag{9}$$

Degrees of freedom (2)

The new formula is

$$\nu = \frac{\nu_{\text{old}}\nu_{\text{obs}}}{\nu_{\text{old}} + \nu_{\text{obs}}}. \tag{10}$$

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

$$\nu_{\text{obs}} = \frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3}\nu_{\text{com}}(1 - \lambda). \tag{11}$$

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

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

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

$$\bar{Q} \pm t_{(\nu,1-\alpha/2)}\sqrt{T}, \tag{12}$$

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

$$P_s = \Pr\left[F_{1,\nu} > \frac{(Q_0 - \bar{Q})^2}{T}\right] \tag{13}$$

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.
- Repeat the analysis with $m = 5$ with different seeds. If there are large differences for some parameters, this means that the data contain little information about them.

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
```

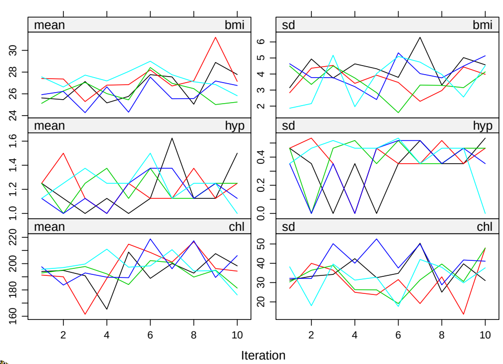

Inspect missing data pattern

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


Multply impute the data

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

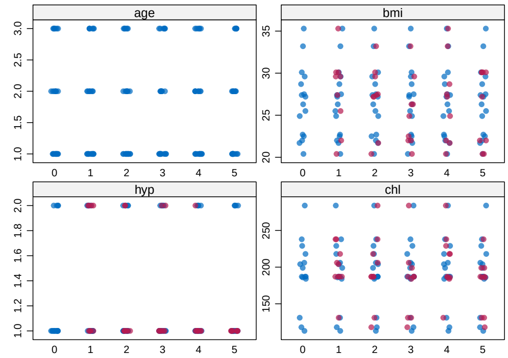

Inspect the trace lines for convergence



Stripplot of observed and imputed data

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

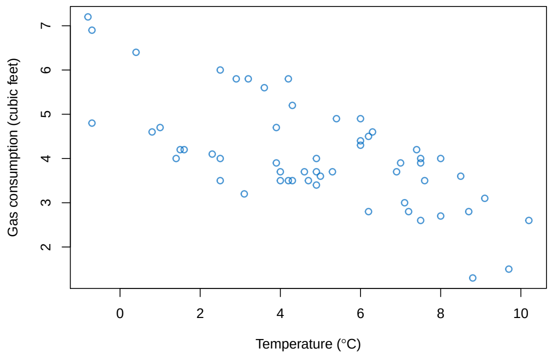
Stripplot of observed and imputed data



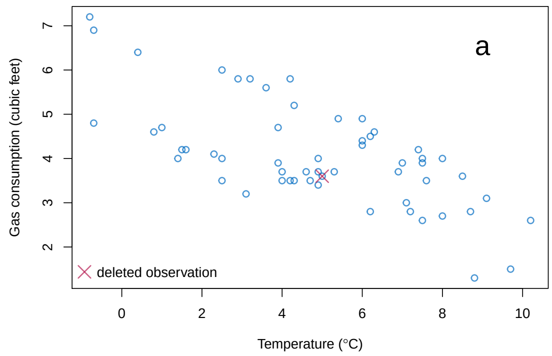
Fit the complete-data model

```
> fit <- with(imp, lm(bmi ~ age))
> est <- pool(fit)
> summary(est)
            estimate std.error statistic   df  p.value
(Intercept)    30.69      2.09     14.70 13.4 3.09e-11
age            -2.35      1.01     -2.33 17.4 3.23e-02
```

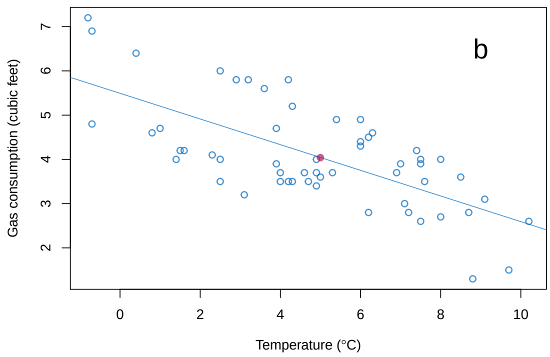

Relation between temperature and gas consumption



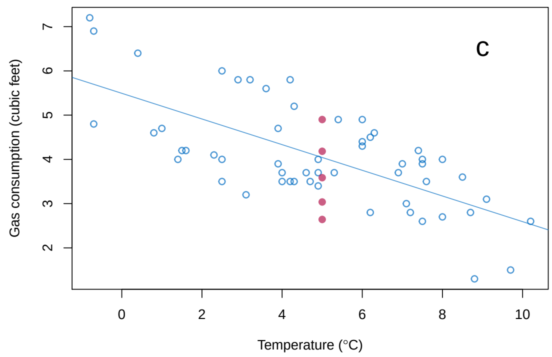
We delete gas consumption of observation 47



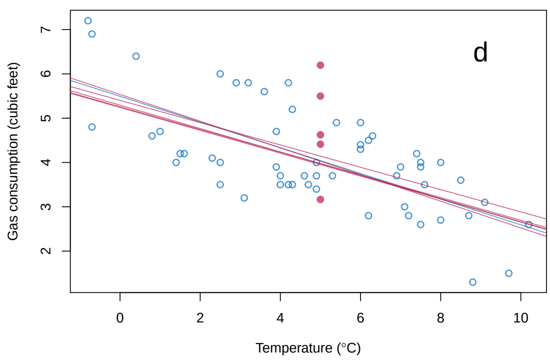
Predict imputed value from regression line



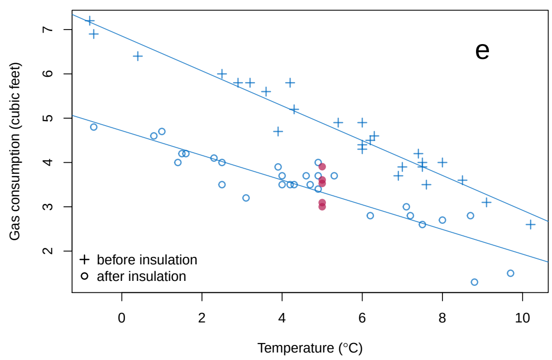
Predicted value + noise



Predicted value + noise + parameter uncertainty



Imputation based on two predictors



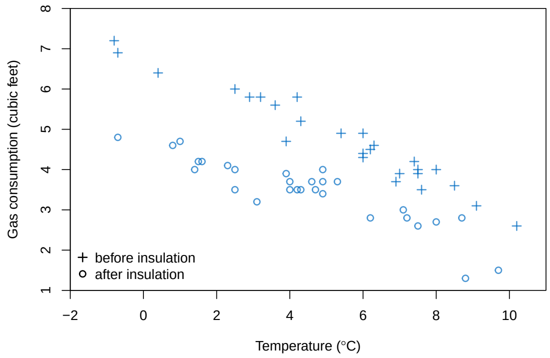
Four techniques

1. *Predict:* $\dot{y} = \hat{\beta}_0 + X_{\text{mis}}\hat{\beta}_1$ (`mice.impute.norm.predict()`)
2. *Predict + noise:* $\dot{y} = \hat{\beta}_0 + X_{\text{mis}}\hat{\beta}_1 + \dot{\epsilon}$ (`mice.impute.norm.nob()`)
3. *Bayesian multiple imputation:* $\dot{y} = \dot{\beta}_0 + X_{\text{mis}}\dot{\beta}_1 + \dot{\epsilon}$, where $\dot{\beta}_0$, $\dot{\beta}_1$ and $\dot{\sigma}$ are random draws from their posterior distribution (`mice.impute.norm()`)
4. *Bootstrap multiple imputation:* $\dot{y} = \dot{\beta}_0 + X_{\text{mis}}\dot{\beta}_1 + \dot{\epsilon}$, 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()`)

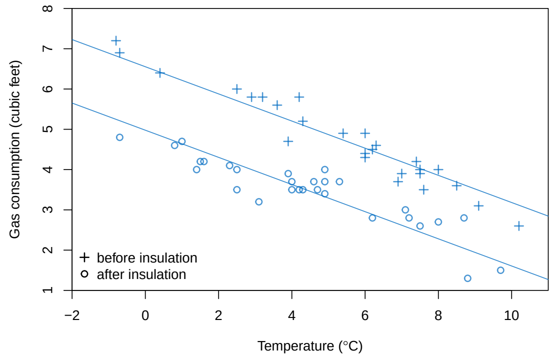
How to evaluate imputation methods

- `https://stefvanbuuren.name/fimd/sec-evaluation.html`
  - Four evaluation criteria
  - Example code
- `https://stefvanbuuren.name/fimd/sec-linearnormal.html#sec:perflin`
  - Five methods
  - Results + interpretation

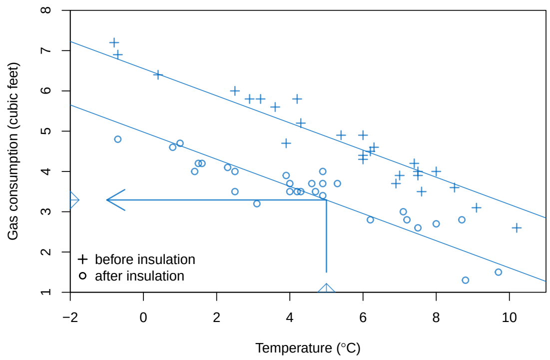
Predictive mean matching: Y given X



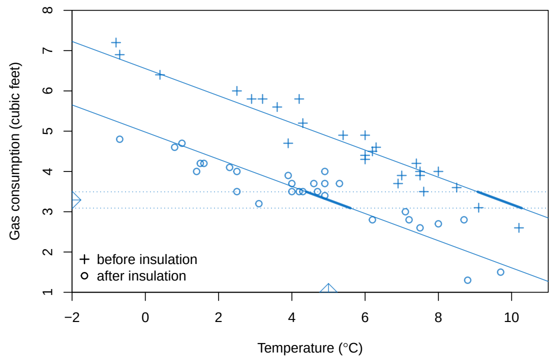
Add two regression lines



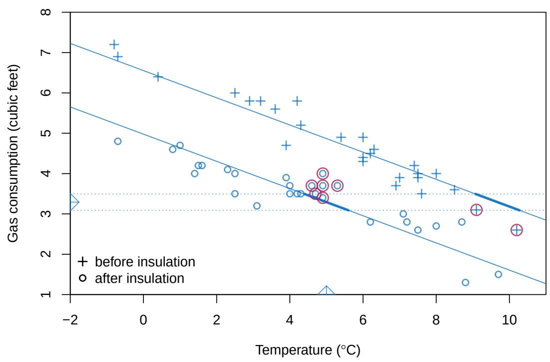
Predicted given 5° C, ‘after insulation’



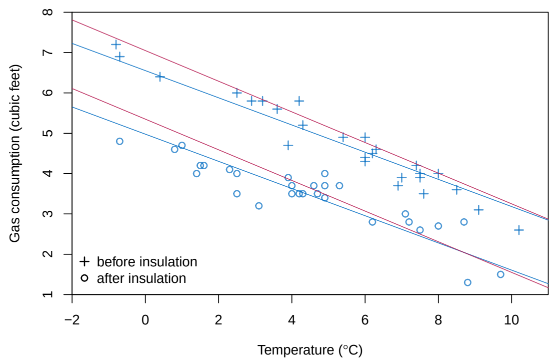
Define a matching range $\hat{y} \pm \delta$



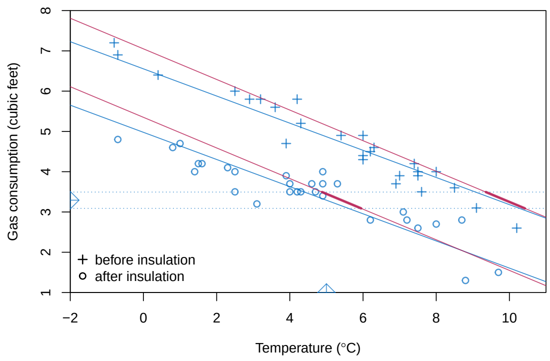
Select potential donors



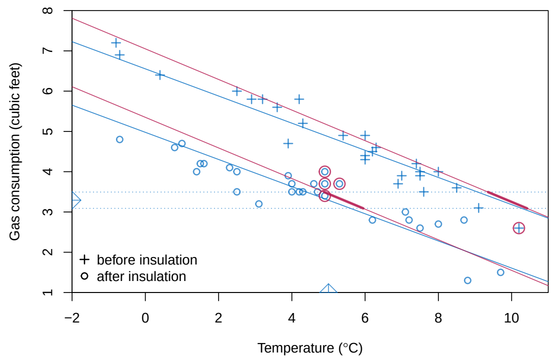
Bayesian PMM: Draw a line



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



Select potential donors

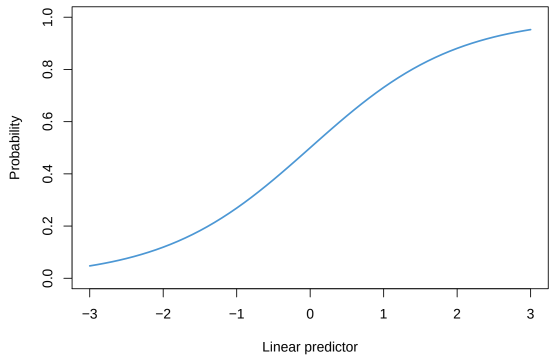


Imputation of a binary variable

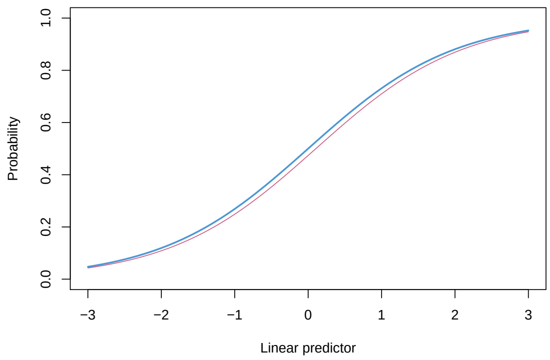
- *logistic regression*

$$\Pr(y_i = 1 | X_i, \beta) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}. \tag{14}$$

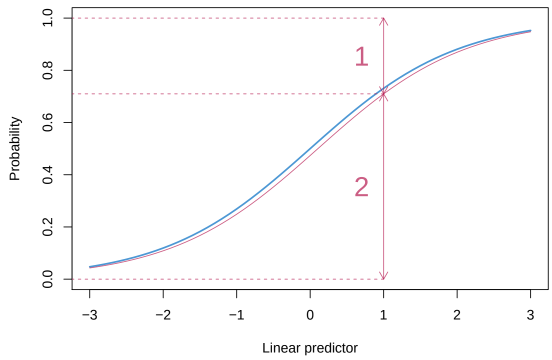
Fit logistic model



Draw parameter estimate



Read off the probability

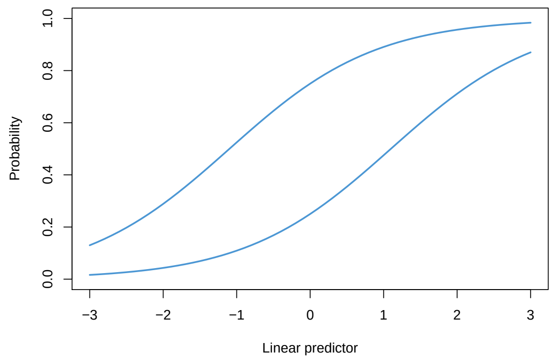


Impute ordered categorical variable

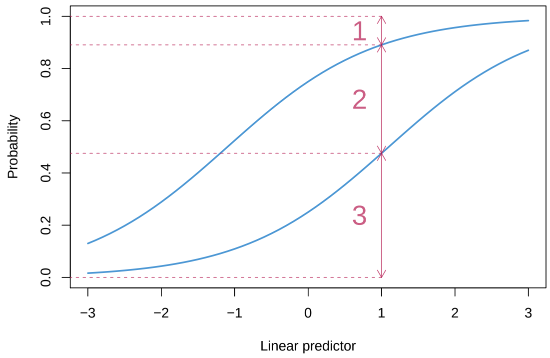
- K ordered categories $k = 1, \ldots, K$
- *ordered logit model*, or
- *proportional odds model*

$$\Pr(y_i = k | X_i, \beta) = \frac{\exp(\tau_k + X_i\beta)}{\sum_{k=1}^{K} \exp(\tau_k + X_i\beta)} \tag{15}$$

Fit ordered logit model



Read off the probability



Other types of variables

- Count data
- Semi-continuous data
- Censored data
- Truncated data
- Rounded data

Built-in imputation functions

- `http://stefvanbuuren.github.io/mice/reference/index.html`

PART E

Slot E: Multivariate imputation

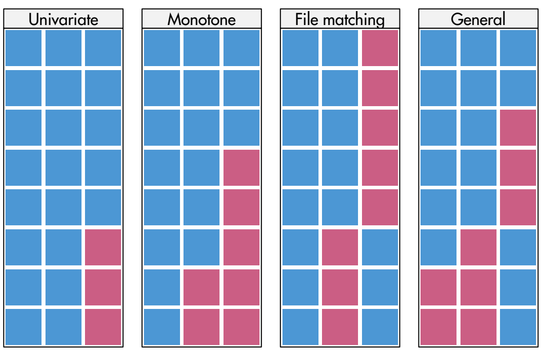
- Missing data patterns
- Multivariate multiple imputation
- Fully conditional specification
- Assessment of convergence
- Compatibility

FIMD Chapter 4

Problems in multivariate imputation

- Predictors themselves can be incomplete
- Mixed measurement levels
- Order of imputation can be meaningful
- Too many predictor variables
- Relations could be nonlinear
- Higher order interactions
- Impossible combinations

Missing data patterns



Influx and outflux

- *Influx* and *Outflux* are statistics of the missing data pattern
- *Influx coefficient* I_j:

$$I_j = \frac{\sum_j^p \sum_k^p \sum_i^n (1 - r_{ij}) r_{ik}}{\sum_k^p \sum_i^n r_{ik}} \tag{16}$$

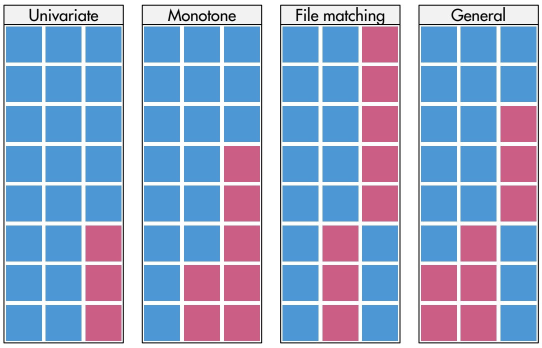
- *Outflux coefficient* O_j:

$$O_j = \frac{\sum_j^p \sum_k^p \sum_i^n r_{ij} (1 - r_{ik})}{\sum_k^p \sum_i^n 1 - r_{ij}} \tag{17}$$

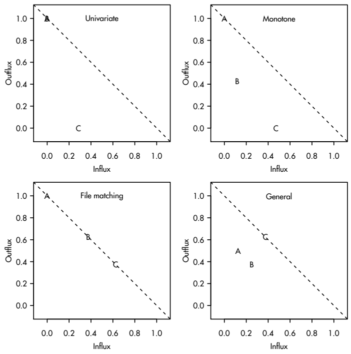
Influx and outflux

- *Influx* of a variable quantifies how well its missing data connect to observed data in other variables
- *Outflux* of a variable quantifies how well its observed data connect to the missing data in other variables

Missing data patterns



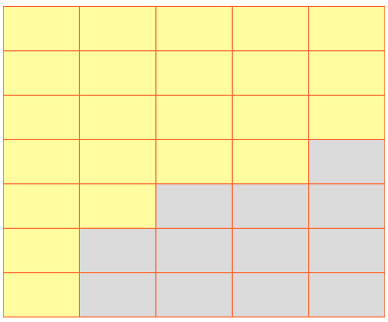
Fluxplot



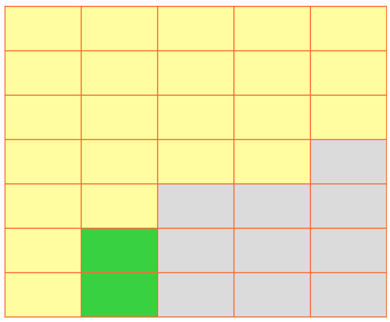
Three general strategies

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

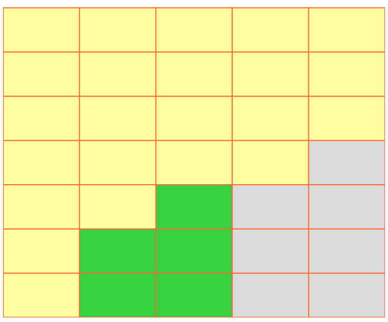
Imputation of monotone pattern



Imputation of monotone pattern



Imputation of monotone pattern



Joint Modeling (JM)

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 modeling: Software

R/S Plus	`norm`, `cat`, `mix`, `pan`, `Amelia`
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
- Many applications

Joint Modeling: Con's

- Lack of flexibility
- Leads to unrealistically large models
- Can assume more than the complete data problem
- Confounds the nonresponse and the complete data problems

Fully Conditional Specification (FCS)

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

Multivariate Imputation by Chained Equations (MICE)

- MICE algorithm
- Specify imputation model for each incomplete column
- Fill in starting imputations
- And iterate
- Model: Fully Conditional Specification (FCS)

Fully Conditional Specification: Con's

- Theoretical properties only known in special cases
- Cannot use computational shortcuts, like sweep-operator
- Joint distribution may not exist (incompatibility)

Fully Conditional Specification: Pro's

- Easy and flexible
- Imputes close to the data, prevents impossible data
- Subset selection of predictors
- Modular, can preserve valuable work
- Works well, both in simulations and practice

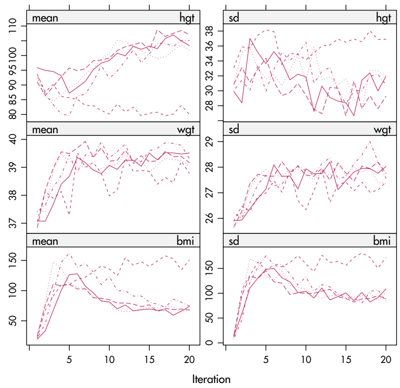
Fully Conditional Specification (FCS): Software

`R`	`mice`, `transcan`, `mi`, `VIM`, `baboon`
`SPSS V17`	`procedure multiple imputation`
`SAS`	`IVEware`, `SAS 9.3`
`STATA`	`ice command`, `multiple imputation command`
Stand-alone	`Solas`, `Mplus`

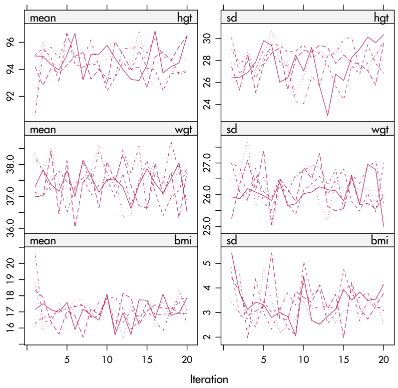
How many iterations?

- 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

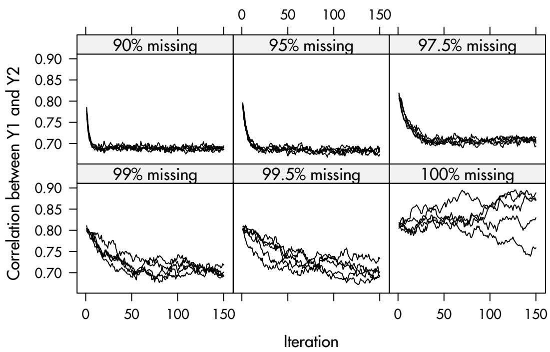


How many iterations does MICE need?

	X	Y1	Y2	
1000	1	1	1	0
4500	1	1	0	1
4500	1	0	1	1
	0	4500	4500	9000

- `https://stefvanbuuren.name/fimd/sec-FCS.html#sec:howlarget`

Convergence towards true correlation of 0.7



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.

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**

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

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

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

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

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`

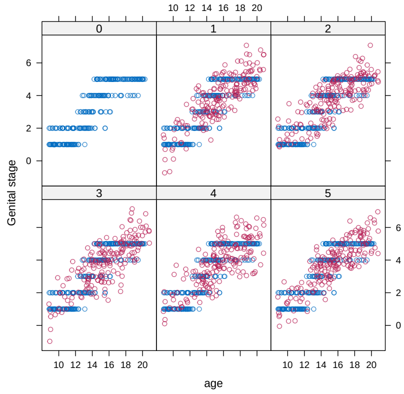
Joint model vs Fully conditional specification

- 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?

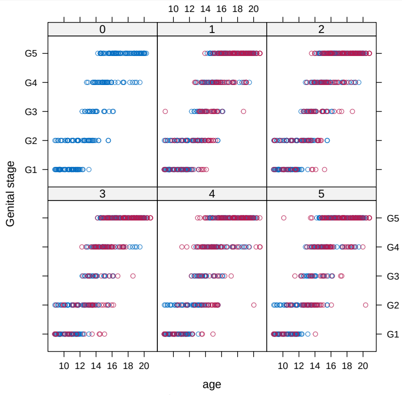
Imputation methods

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

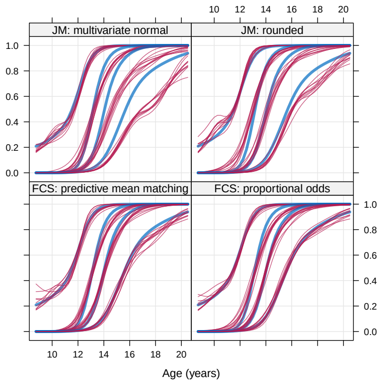
JM: Multivariate normal model



FCS: Proportional Odds model



JM vs FCS



PART G

Slot G: Modelling choices, derived variables

- Modeling choices
- Predictor selection
- Derived variables
- Diagnostics

FIMD2 Chapter 6
`https://stefvanbuuren.name/fimd/ch-practice.html`

Imputation model choices

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

Which predictors?

1. Include all variables that appear in the complete-data model, including transformations and interactions
2. In addition, include the variables that are related to the nonresponse
3. In addition, include variables that explain a considerable amount of variance
4. Remove from the variables selected in steps 2 and 3 those variables that have too many missing values within the subgroup of incomplete cases.

Function `quickpred()` and `flux()`

Derived variables

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

Imputing a ratio

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

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

Derived variables: summary

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

Standard diagnostic plots in `mice`

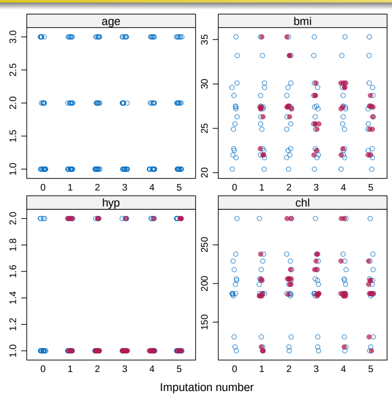
Since `mice` 2.5, plots for imputed data:

- one-dimensional scatter: `stripplot`
- box-and-whisker plot: `bwplot`
- densities: `densityplot`
- scattergram: `xyplot`

Stripplot

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

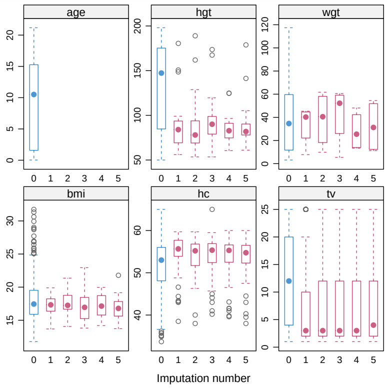

`stripplot(imp, pch=c(1,19))`



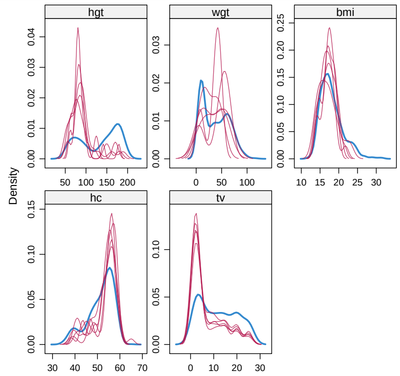
A larger data set

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

bwplot(imp)



densityplot(imp)



PART I

Slot I: Analysis of imputed data

- Workflows
- Pooling non-normal quantities
- Multiparameter test
- Longitudinal data example

Workflows

- Four different objects in `mice`
- Seven recommened workflows
- Two non-recommended workflows
- Custom calculations
- `https://stefvanbuuren.name/fimd/workflow.html`

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: 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)

Multiparameter tests

- `D1` Multivariate Wald test
- `D2` Combined test statistics
- `D3` Likelihood ratio test
- `https://stefvanbuuren.name/fimd/sec-multiparameter.html`

Saturday, May 13 2000, Enschede

SE Fireworks Disaster

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

An 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

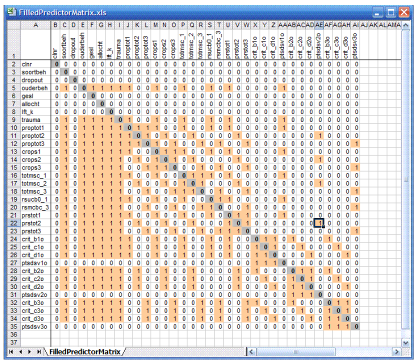
- Is one of these treatments more effective in reducing PTSD symptoms at T2 and T3?
- Does the number of sessions needed to produce the therapeutic effect differ between the treatments?

(Missing) Data

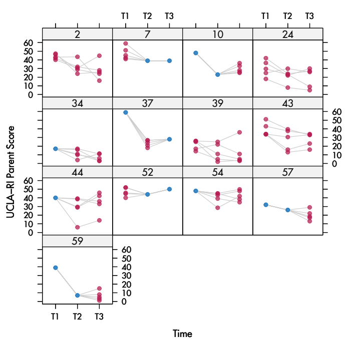
Table 9.1: SE Fireworks Disaster data. The UCLA PTSD Reaction Index of 52 subjects, children and parents, randomized to EMDR or CBT.

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	–
30	E	Y	–	–	–	20	23	12	58	E	Y	–	–	–	49	28	25
31	C	N	15	24	26	33	36	38	59	E	Y	39	7	–	39	7	–

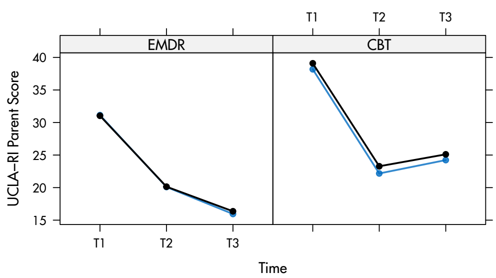
Predictor matrix for multiple imputation



Imputed Data



UCLA-RI Parent



Longitudinal data: Conclusions

- Imputation should preserve
  - Group compositions across time
  - Relations within time
  - Relation across time
- If possible, code data in 'broad' form
- Codify predictor matrix to reflect data structure
- Use simple complete-data analysis: *t*-test, ANOVA, MANOVA

PART K

Relevance of ignorability assumption 1

Ignorability implies

$$P(Y|X, R = 0) = P(Y|X, R = 1) \tag{18}$$

so

$$P(Y_{\text{obs}}|X) = P(Y_{\text{mis}}|X) \tag{19}$$

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

Models for nonignorable nonresponse

$P(Y, R)$ does not factorise into independent parts, and must be modelled jointly
Two approaches (there are some more):

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

Selection model

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

$$P(Y, R|\psi, \theta) = P(R|Y, \psi)P(Y, \theta) \tag{20}$$

$P(R = 1|Y)$	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)	$p(R=0\|BP)$	$p(BP)$	$p(BP\|R=1)$	$p(BP\|R=0)$
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)

$$P(Y, R|\psi, \theta) = P(Y|R, \theta)P(R|\psi) \quad (21)$$

$P(Y|R = 1, \theta)$ (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)	$p(R=0 \mid BP)$	$p(BP)$	$p(BP \mid R=1)$	$p(BP \mid R=0)$
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

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

A simple model to shift imputations

Specify $P(Y|X, R)$

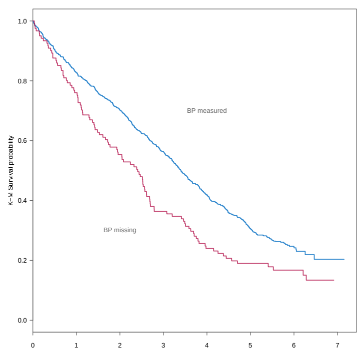
Model		
1	$Y = X\beta + \epsilon$	β is estimated from cases $R = 1$
2	$Y = X\beta + \delta + \epsilon$	imputations applied to $R = 0$

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

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

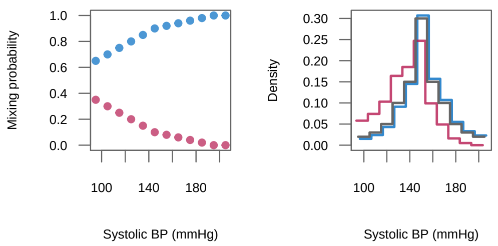
Both models

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)	$p(R=0\|BP)$	$p(BP)$	$p(BP\|R=1)$	$p(BP\|R=0)$
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

Effect of response mechanism on BP



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`

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

PART M

Slot M: Multilevel data

- Notation for multilevel models (7.2)
- Missing data, practical issues (7.3.1)
- Ad hoc methods: listwise, ignore, dummy (7.3.2)
- FCS imputation for multilevel data (7.5)
- Examples of specifications (7.10)
- Two recipes (7.10.8)

PART O

Slot O: Capita selecta

- Reporting guidelines
- ...any other business

Reporting guidelines

1. Amount of missing data
2. Reasons for missingness
3. Differences between complete and incomplete data
4. Method used to account for missing data
5. Software
6. Number of imputed datasets
7. Imputation model
8. Derived variables
9. Diagnostics
10. Pooling
11. Listwise deletion
12. Sensitivity analysis