Statistical Analysis of Incomplete Data

Course Notes: Prof. Dr. Susanne Rässler & Dr. Florian Meinfelder

Presenter: Florian Meinfelder

Department of Statistics and Econometrics
Otto-Friedrich-Universität Bamberg

EMOS Webinar - 23/05/2018

Outline

Outline

Roadmap to this webinar

- The first two sections focus on general information, relevant assumptions, and notation
- Section Missing-Data Patterns introduces pre-impuation analysis
- Section Single Imputation is a short prelude of inadequate ways to handle missing data
- Section MI in a Nutshell shows how to *analyze* multiply imputed data
- Section A Short Note on MI Algorithms shows how to actually *perform* MI (a look under the hood)
- Section Problems in Empirical Application Settings confronts us with the cold and bitter real world...

Introduction

Missing data can occur in many situations

- Unit-nonresponse
- Item-nonresponse
- Drop-out in panel studies
- Data fusion
- Split questionnaire survey design
- Synthetic data
- Rubin's Causal Model

Introduction

Item	Sex	Age	Education	Health	Personal net income	...
1	female	40-45	high	good	?	...
2	male	30-35	middle	bad	4500-5000	...
3	female	>60	?	middle	4000-4500	...
4	male	20-25	high	?	?	...
5	male	20-25	low	?	1500-2000	...
6	female	30-35	low	good	1500-2000	...
...	...	...	...	...	...	...

Erase of the cases

Lfd. Nr.	Sex	Age	Education	Health	Personal net income	...
2	male	30-35	middle	bad	4500-5000	...
6	female	30-35	low	good	1500-2000	...
...	...	...	...	...	...	...

$\Rightarrow$ **In multivariate analysis a considerable amount of the data might be lost due to case deletion**

The goal of the statistical analysis of missing data is...

To make valid and efficient inference about population parameters from an incomplete dataset!

The goal of statistical analysis with missing data is *NOT*...

- ...to estimate, predict, or recover missing values
- ...to obtain the same answers that would have been seen without missing data

$\Rightarrow$ Attempts to reconstruct or recover the true missing values can result in bias and may actually harm the inference.

Overview: Statistical analysis of incomplete data

- Methods which use only the available (AC = available cases) or the complete cases (CC = complete cases)
- Weighting, in general to adjust Unit-Nonresponse or "Oversampling" etc.
- Likelihood-based parameter estimations, e.g. via Expectation Maximization-algorithm(=EM-algorithm) of Dempster, Laird & Rubin (1977) or pattern mixture models or Full Information Maximum Likelihood (FIML)/Structural Equation Modelling (SEM)
- Single imputation and adjustment of the variance estimators
- Multiple imputation (= MI) according to Rubin (1978, 1987).

Outline

Missing Data Mechanisms

- View missingness as a probabilistic phenomenon (Rubin 1976, Little & Rubin 2002)

  Y = complete data = (Y_{obs}, Y_{mis})

  R = response indicators

- Missing completely at random (MCAR): Cause of missingness is completely random process (like coin flip)

$$f(R|Y) = f(R) \text{ does not depend on } Y_{mis} \text{ or } Y_{obs}$$

- Missing at random (MAR): Missingness may be related to observed variables but no residual relationship with missing variables

$$f(R|Y) = f(R|Y_{obs}) \text{ does not depend on } Y_{mis}$$

- Not missing at random (NMAR): Missingness is still related to missing variables

$$f(R|Y) \text{ depends on } Y_{mis} \text{ and presumably on } Y_{obs}$$

- MAR $\Leftrightarrow$ ignorable NMAR $\Leftrightarrow$ nonignorable

Example: (Enders 2010) Job Performance Ratings

	Job performance ratings II			
IQ	Complete	MCAR	MAR	MNAR
78	9	–	–	9
84	13	13	–	13
84	10	–	–	10
85	8	8	–	–
87	7	7	–	–
91	7	7	7	–
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	–	7	–
99	7	7	7	–
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	–	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	–	12	12

Hypothetical situation for complete cases and the 3 possible missing data mechanisms MCAR, MAR, and MNAR missing values.

Summary: Assumptions regarding the mechanism of the missing data

- MAR + Distinctness = Ignorability
- Need to incorporate all variables related to the missing-data mechanism and the missing data

Outline

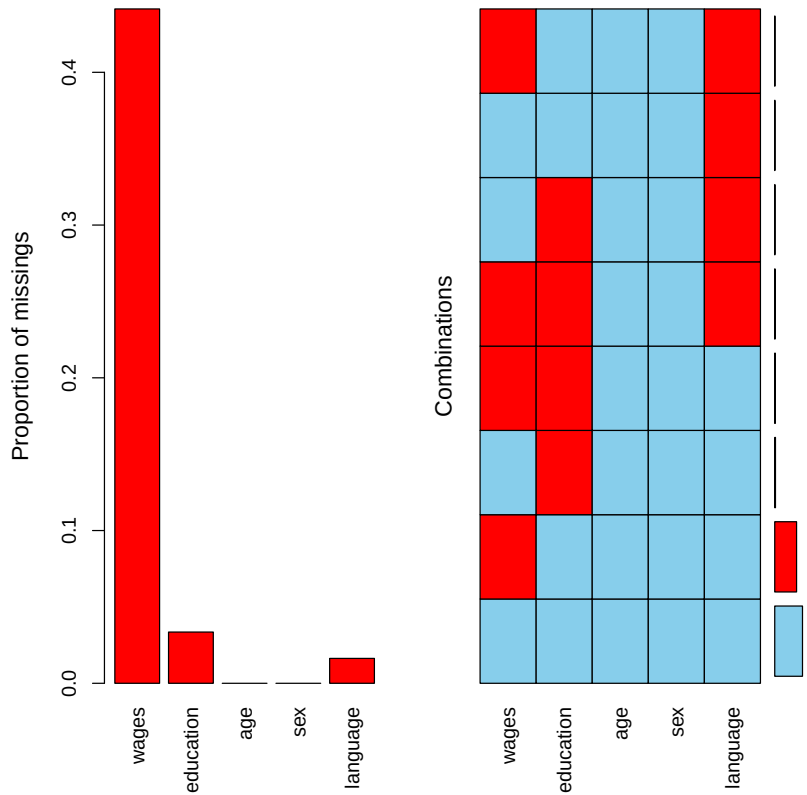
Overview of different missing-data patterns

Not only missing-data *mechanisms*, but also missing-data *patterns* can affect the way we handle missing-data.
Important aspect which is often confused: A pattern can be very non-random, but the mechanism might still be MCAR

Graphical Diagnostics with VIM (Templ et al. 2016): Missing Data Patterns

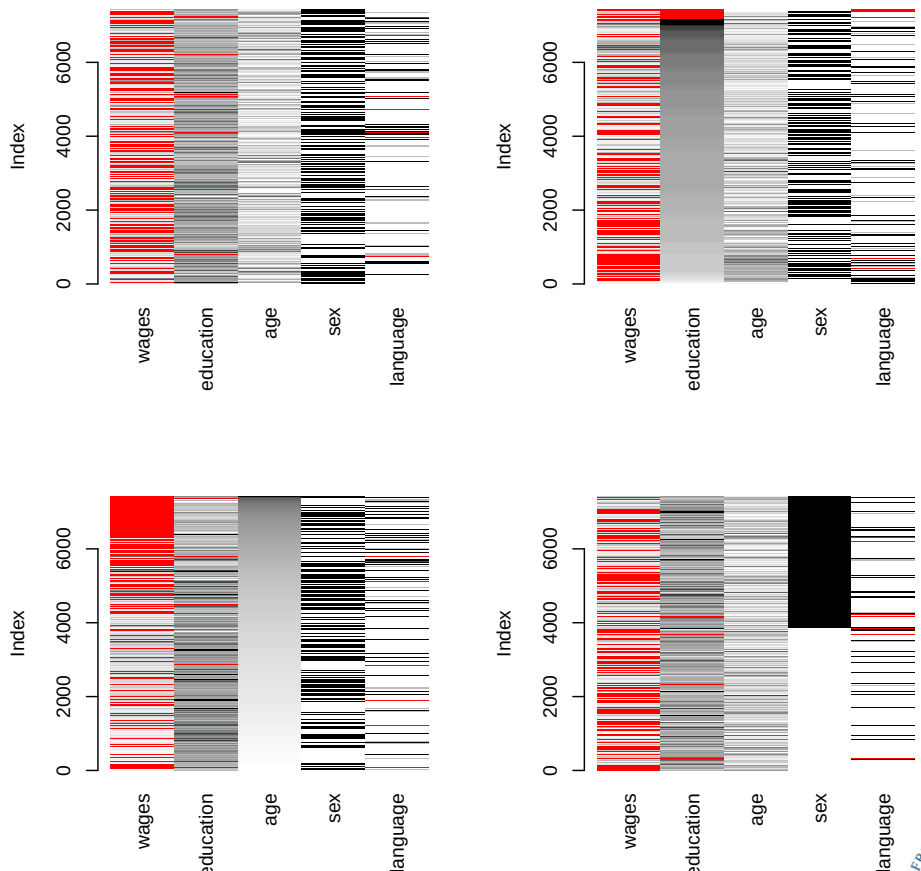
The R package `VIM` provides some nice graphical diagnostic tools:

The function `aggr()` gives an overview of the occurring combinations of missing data and their frequencies.



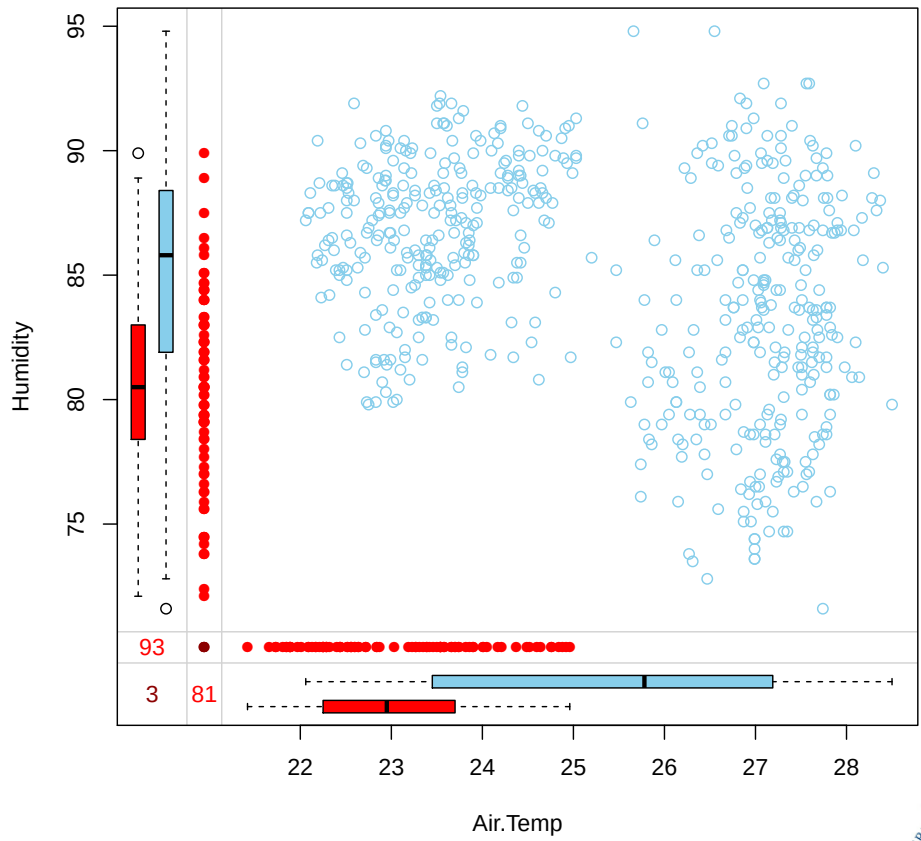
Graphical Diagnostics with VIM: Missing Data *Mechanisms*

The function `iimagMiss()` produces a matrix plot, where shades of grey (more than 50 if needed) are used for observed values, whereas 'red' displays missing values. Sorting by any variables indicates relationships between the values of this variable and the propensity to be missing.



Graphical Diagnostics with VIM: Missing Data Mechanisms (ctd.)

The function `marginplot()` produces a bivariate scatter plot that features boxplots for the marginal distributions conditioned on 'observed' or 'missing' for the other variable.

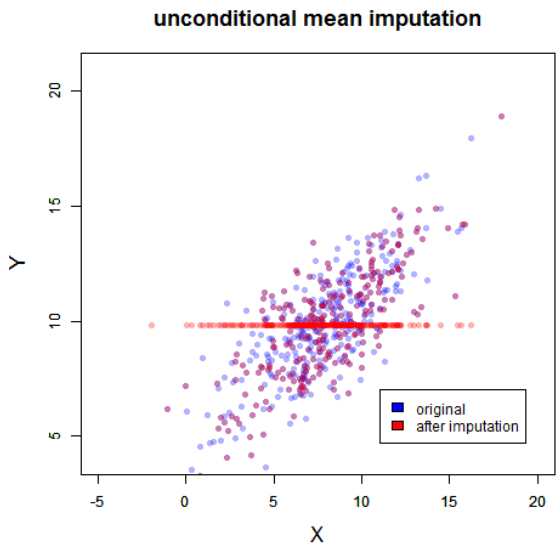


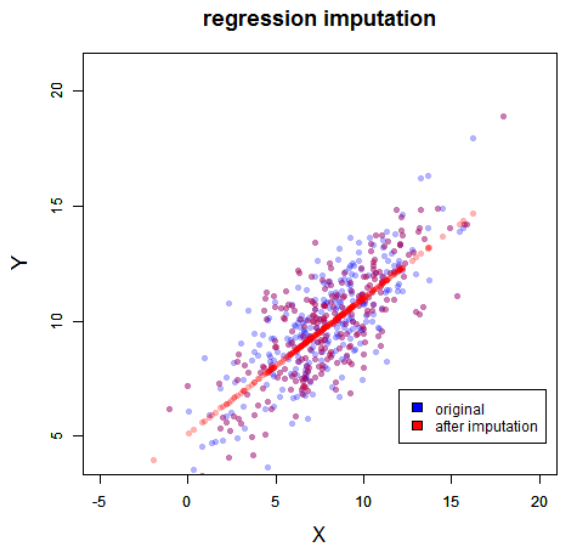
Outline

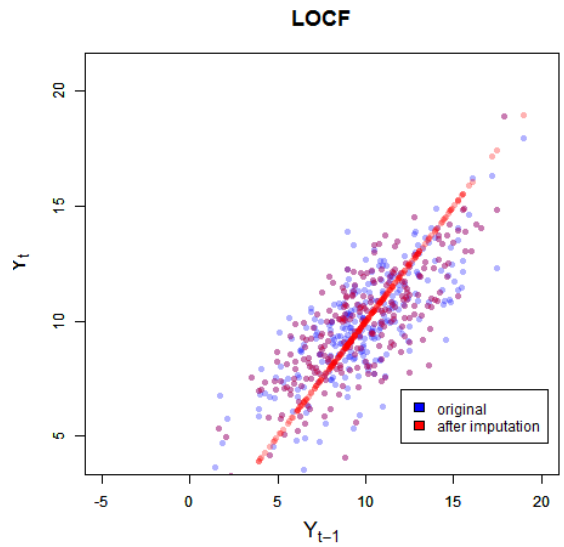
Overview of the shortcomings of naïve imputation methods

Replace missing values by 'estimates'

- means
- regression predictions
- 'last observation carried forward' (in panel studies)







All these methods tend to bias (variance) estimators!

A conceptual quantum leap: The introduction of randomness

Abandon the goal of predicting a missing value as well as possible

- (Simple) hot deck:
  - First introduced by the US Census Bureau in the late 1960s
  - Impute missing values of Y by generating n_{mis} draws with replacement from Y_{obs}
- Stochastic regression imputation:
  - Estimate $\hat{\beta}$ and $\hat{\sigma}^2$ from the observed subsample
  - Impute $y_{imp,i} = \mathbf{x_{mis,i}}\hat{\beta} + \tilde{u}_i$, where $\tilde{u}_i \sim N(0, \hat{\sigma}^2)$
  - $\rightarrow$ regression predictions plus residual noise

$\Rightarrow$ But even stochastic regression imputation underestimates the variance!

Outline

To MI or not to MI...

When *DO* we need MI?

- Confidence Intervals
- Hypothesis Testing
- Model uncertainty
- In short: Inferential Statistics

When *DO* we *NOT* need MI?

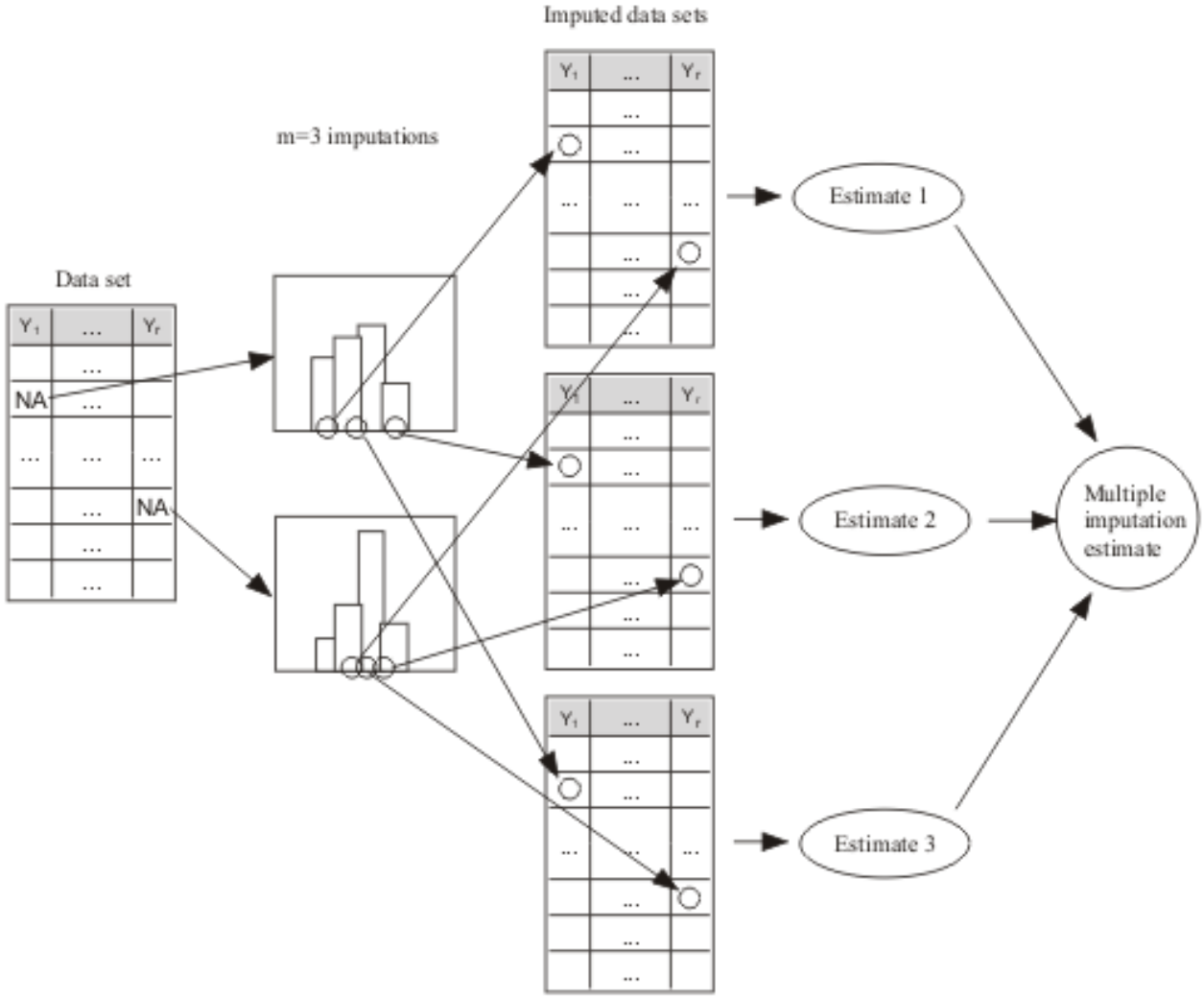
- Cross tabs
- Correlations (as descriptive measure)
- In short: Descriptive Statistics

Aside from yielding unbiased estimators under ignorability, MI only has one purpose...

Adjusting the variance of an estimator based on a sampling function, such that the uncertainty created by the missing information is accounted for correctly.

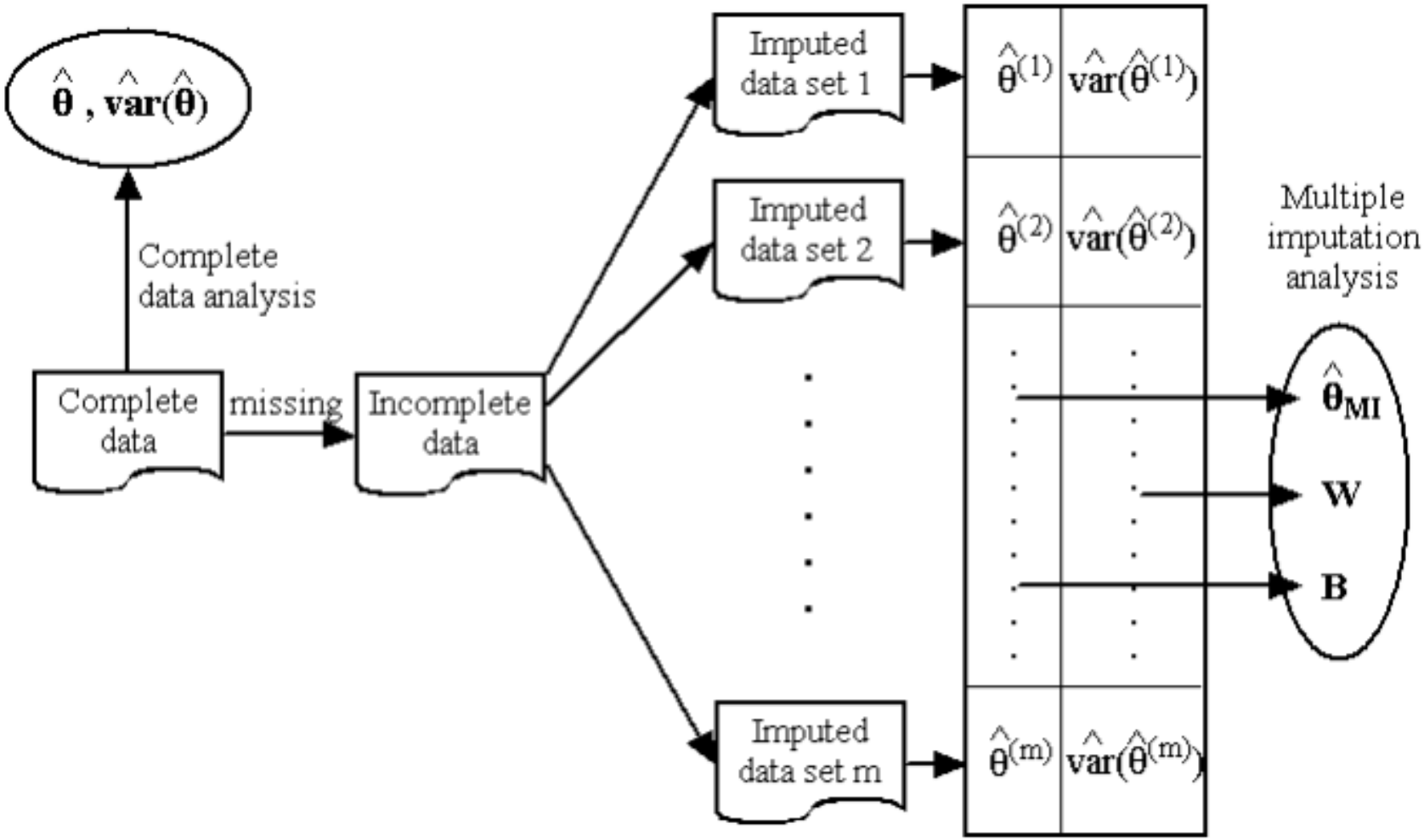
- We will (usually) get *wider* confidence intervals for MI estimators than we would have gotten for (hypothetical) complete-data estimators, but...
- We will (usually) get *narrower* confidence intervals for MI estimators than we would get for complete case-estimators (aka listwise deletion).

Multiple imputation: Principle I



$\Rightarrow$ MI reflects the uncertainty of the imputation **and** the model

Multiple imputation: Principle II



$\Rightarrow$ Results are combined according to the analysis of variance rules

Multiple imputation: Principle III

- **Estimation:** given complete data, estimators should be (approximately) normally distributed (Rubin & Schenker 1986) ,(Rubin 1987), i.e.

$$(\widehat{\theta} - \theta)/\sqrt{var(\widehat{\theta})} \sim N(0,1)$$

- Generate $m = 1, \ldots, M$ completed data sets and estimate $\widehat{\theta}^{(m)}$ and $\widehat{var}(\widehat{\theta}^{(m)})$
- This leads to the **Multiple Imputation estimator**

$$\widehat{\theta}_{MI} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}^{(m)}. \tag{1}$$

- The estimated variance is

$$T = W + (1 + \frac{1}{M})B \tag{2}$$

with "Within-Imputation" variance $W = \frac{1}{M}\sum_{m=1}^{M}\widehat{var}(\widehat{\theta}^{(m)})$ and "Between-Imputation" variance $B = \frac{1}{M-1}\sum_{m=1}^{M}(\widehat{\theta}^{(m)} - \widehat{\theta}_{MI})^2$

- The "Fraction of Missing Information" (FoMI) for a scalar θ is given by

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

where $r = \frac{(1+M^{-1})B}{W}$ describes the relative increase in variance due to missing information.

Multiple imputation: Principle III (ctd.)

$\Rightarrow$ **Using the MI estimator and the corrected variance we get** $(\widehat{\theta}_{MI} - \theta)/\sqrt{T} \sim t_\nu$, with

$$\nu = (M-1)\left(1 + \frac{W}{(1+M^{-1})B}\right)^2 \tag{4}$$

('classical' definition of the DoF) or

$$\nu_{br} = (\gamma^2/(M-1) + \hat{\nu}_{obs}^{-1})^{-1}, \tag{5}$$

revised DoF by Barnard & Rubin (1999) where $\gamma = \frac{(1+M^{-1})B}{T}$, and $\hat{\nu}_{obs} = (1-\gamma)\nu_{com}(\nu_{com}+1)/(\nu_{com}+3)$,
with ν_{com} being the degrees of freedom for the complete data.

Example: MI within the IAB establishment panel

- Estimating a production function

  productivity (Y) = f(labor (L), capital (K), sector ($SECT$), stage of technology ($TECH$), East/West (EW))

- Via a trans-log-production function (Greene 2000)

$$\begin{aligned} \ln Y &= \beta_0 + \beta_1 \ln L + \beta_2 \ln K + \frac{1}{2}\beta_3 \ln^2 L + \frac{1}{2}\beta_4 \ln^2 K + \beta_5 \ln L \ln K \\ &+ \beta_6 TECH + \sum_{j=7}^{25} \beta_j SECT_j + \beta_{26} EW + \ldots + U, \end{aligned}$$

- Using complete cases only, 40% of data get lost
- Impute with NORM according to Schafer (1999) under the assumption of a multivariate normal distribution model

Example: Results with the East/West-variable

- Maximum-Likelihood estimations via LIMDEP V7.0, Econometric Software, Inc. (Greene 1998)
- Results regarding β_{26}

data set	variable	n	coefficient	se	t-value	p-value
Available cases	EW	6489	0.358	0.115	3.110	0.002
Imputation 1	EW	10990	0.249	0.099	2.506	0.012
Imputation 2	EW	10990	0.346	0.098	3.550	0.000
Imputation 3	EW	10990	0.197	0.096	2.042	0.041
Imputation 4	EW	10990	0.308	0.097	3.164	0.002
Imputation 5	EW	10990	0.313	0.098	3.203	0.001
Imputation 6	EW	10990	0.296	0.099	3.000	0.003

- MI-estimator $\widehat{\beta}_{26}^{MI} = 0.285$
  ($\sqrt{T} = 0.114$, df ≈ 75, t-value $= 2.510$, p-value $= 0.014$)

$\Rightarrow$ marginal effects $e^{\widehat{\beta}} - 1$: 43% (CC) vs. 33% (MI)!

Outline

Two variants of *sequential regression* algorithms

1. M (parallel) chains
   - M burn-in iterations (discarded), final iteration T is stored for each of the M chains
   - starting solution often can be based on hot deck imputation (e.g. *mice*)
2. one single chain
   - only one burn-in cycle (discarded), then R-th iteration is stored as *imputation* (the others are discarded as thinning). Repeat M times.

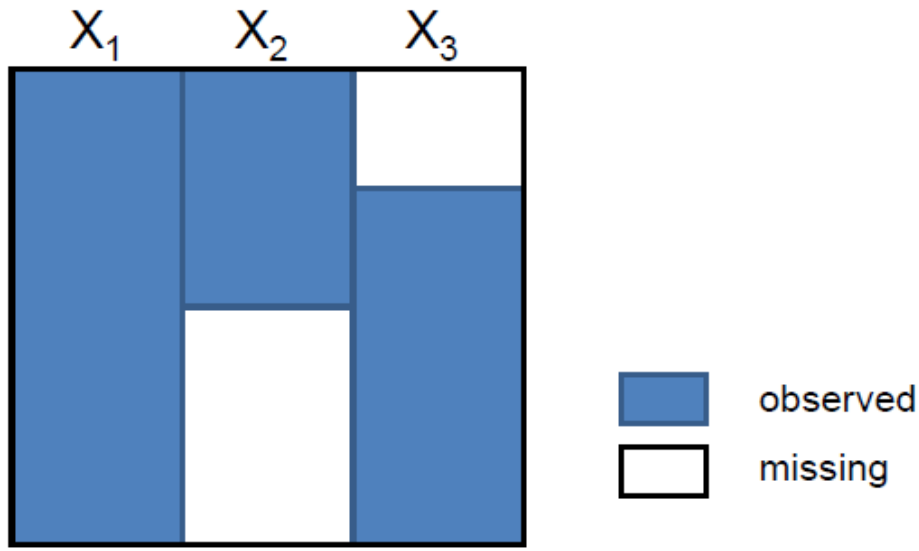
- $\Rightarrow$ Computational drawback of 'M chain variant' is only offset if multi-core threading is used
- better convergence diagnostics for 'M chain variant'

Pseudo Code for *mice* (description from van Buuren 2012)

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

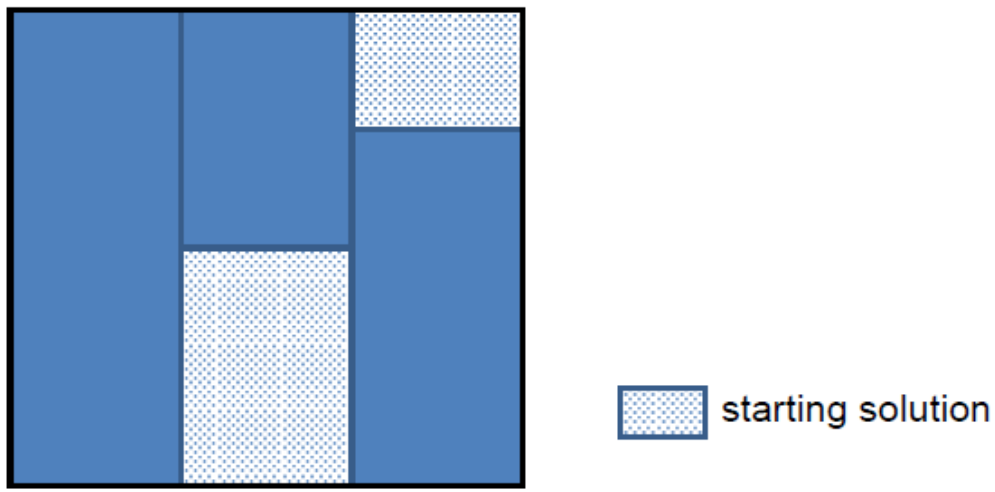
FCS: incomplete data set

We assume a non-monotone missing-data pattern.
(in the following only one iteration is considered!)



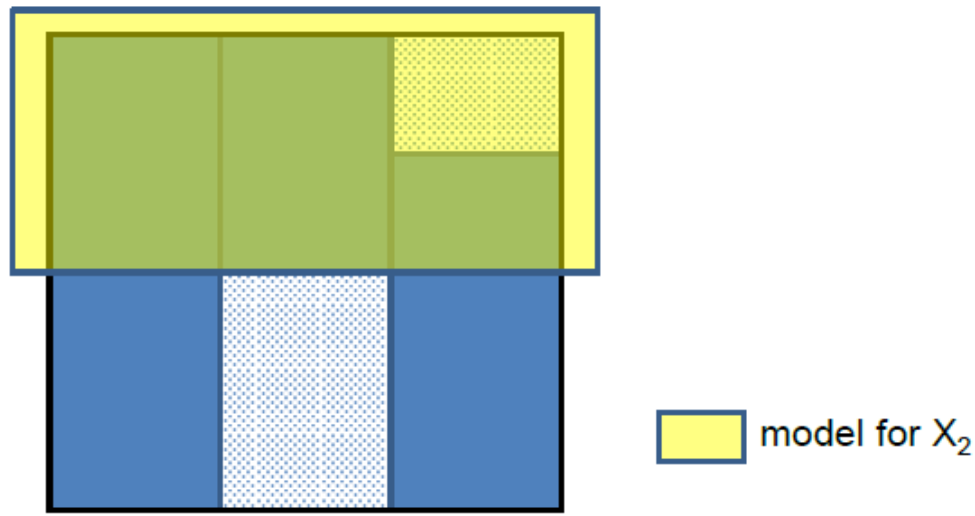
FCS: starting solution

Initial imputations (in `mice` generated via hot deck imputation):



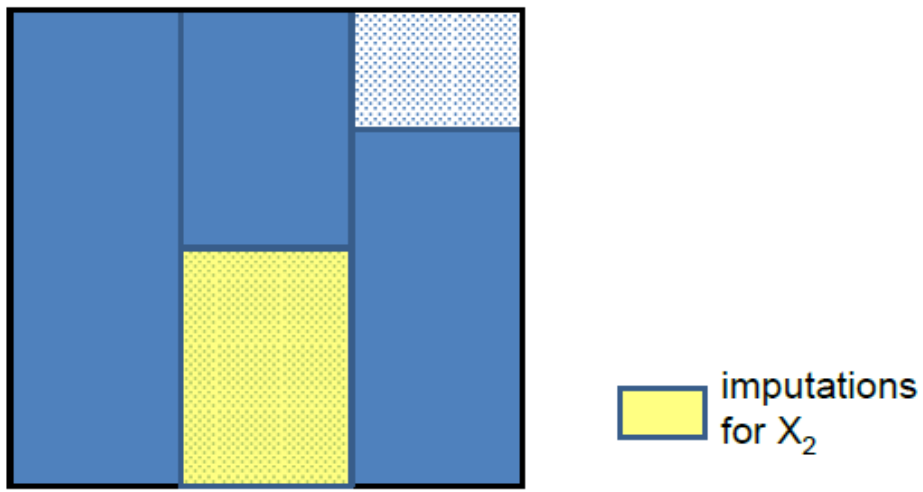
FCS: imputing X_2 conditoned on X_1 and X_3

The imputation model is based on the observed values of X_2.



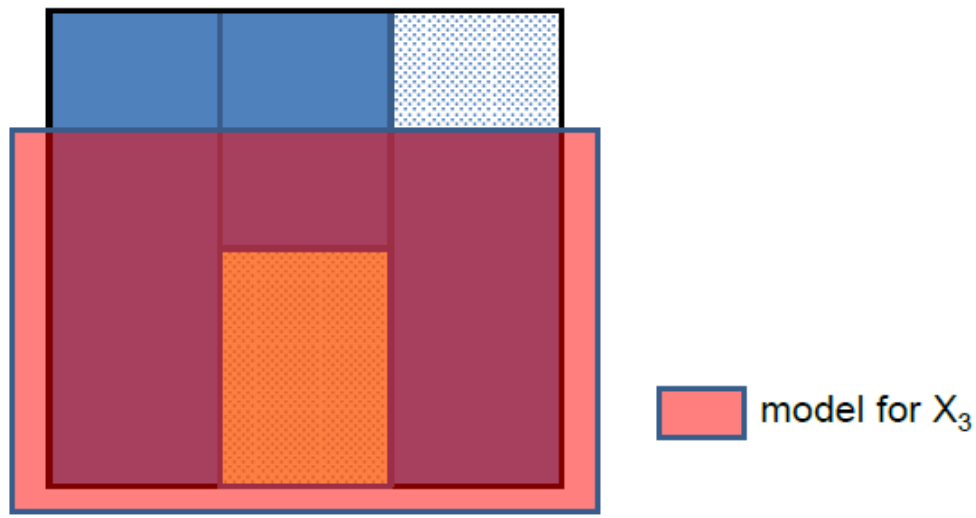
FCS: Imputation of X_2

Values are drawn for
$X_{2,mis}|X_{2,obs}, X_1, X_3, \psi_2$



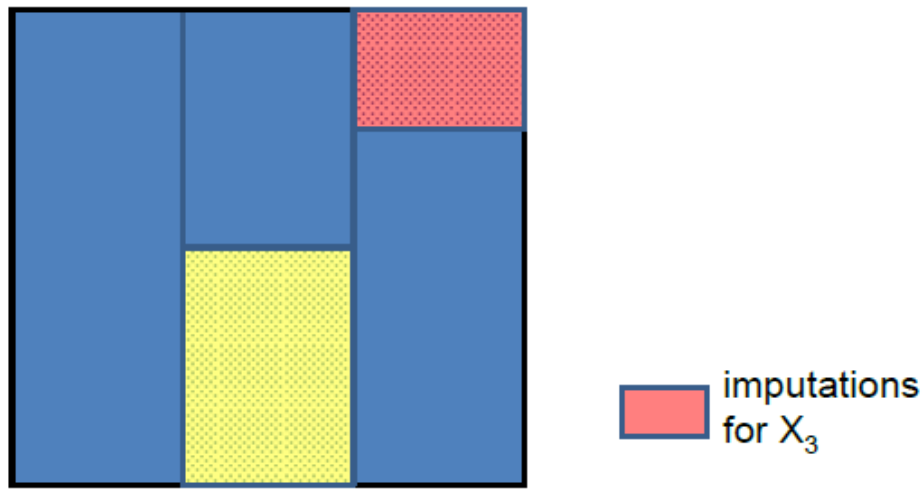
FCS: imputing X_3 conditoned on X_1 and X_2

The imputation model is based on the observed values of X_3.

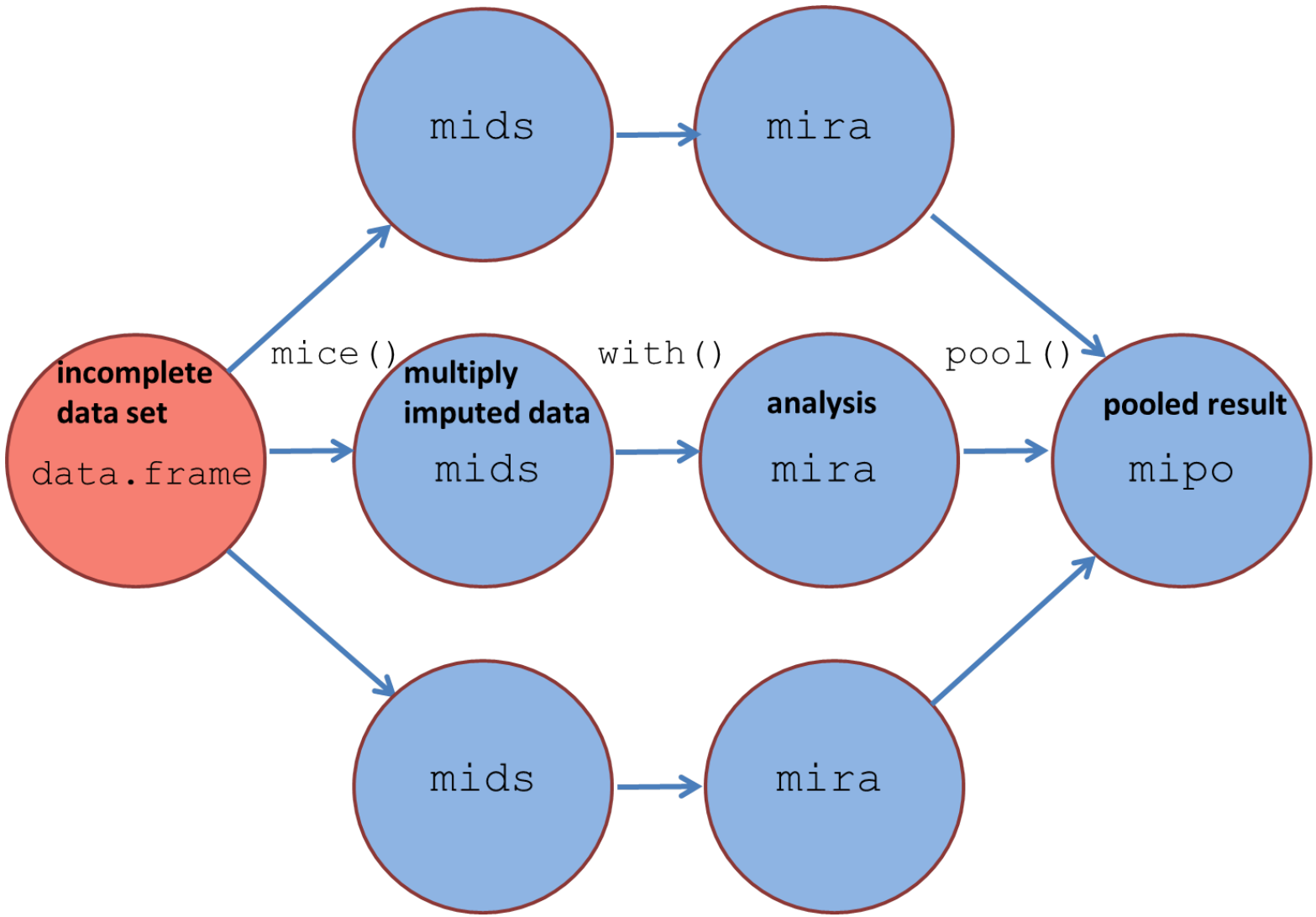


FCS: Imputation of X_3

Values are drawn for
$X_{3,mis}|X_{3,obs}, X_1, X_2, \psi_3$



Schematic overview of processing mice with R



source: own diagram based on van Buuren & Groothuis-Oudshoorn (2011)

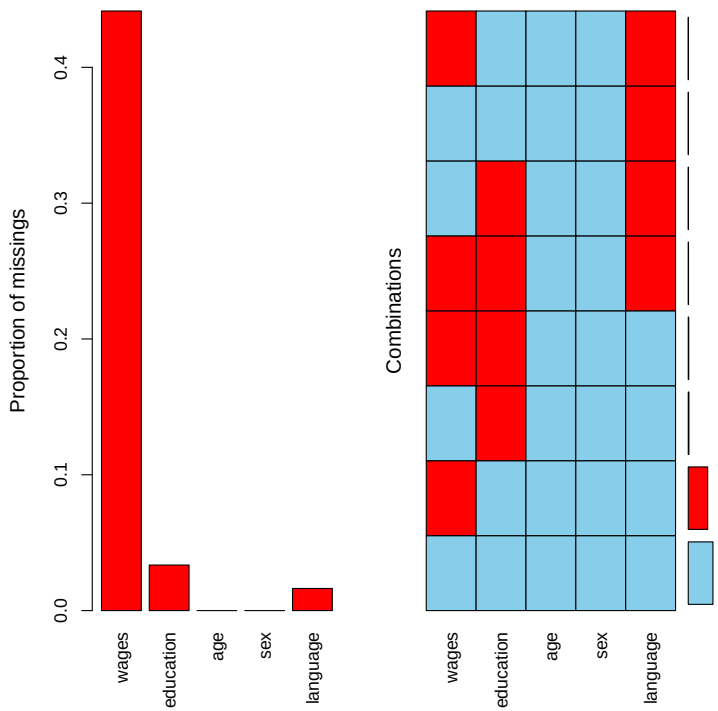
Getting started with mice in R

```
if (!require("VIM")) install.packages("VIM")
if (!require("mice")) install.packages("mice")
```

```
library(mice)
library(VIM)
data(SLID, package = "car")  # Survey of Labor and Income Dynamics
sapply(SLID, class)
```

```
##     wages education       age       sex  language
## "numeric" "numeric" "integer"  "factor"  "factor"
```

```
aggr(SLID)
```



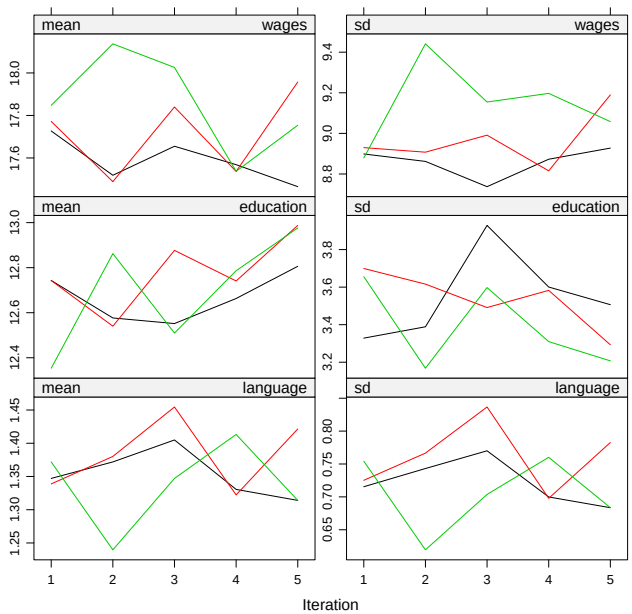
```
set.seed(237846)  # no seed -- no reproducibility!
imp <- mice(SLID, m = 3, printFlag = FALSE)  # only 3 imputations -- use bigger M in 'real life'
imp$method  # which imputation method did mice use?
```

```
##     wages education       age       sex  language
##     "pmm"     "pmm"        ""        "" "polyreg"
```

```
class(imp)
```

```
## [1] "mids"
```

```
plot(imp)  # convergence diagnostics
```




```
theta <- with(imp, lm(wages ~ education + age + sex))  # quantity of interest
class(theta)
```

```
## [1] "mira"   "matrix"
```

```
combRules <- pool(theta)  # application of Rubin's rules
# summary(combRules)
```

Results for the linear model (and how to calculate them by hand)

Variable	*est*	*se*	*t*	*df*	*Pr(>|t|)*	*lo 95*	*hi 95*	*nmis*	*fmi*	*lambda*
(Intercept)	-7.5578	0.4514	-16.7415	870.1986	0	-8.4439	-6.6718	NA	0.0471	0.0449
education	0.8995	0.027	33.2733	104.5642	0	0.8459	0.9531	249	0.1532	0.1372
age	0.2545	0.0084	30.2713	4.3484	0	0.2318	0.2771	0	0.7653	0.6776
sex2	3.4017	0.3062	11.1103	3.8299	5e-04	2.5365	4.2669	NA	0.8034	0.722

```
impDat <- complete(imp, action = "long")
mod <- vector(mode = "list", length = 3)
for (i in 1:3) mod[[i]] <- lm(wages ~ education + age + sex, data = impDat,
subset = impDat$.imp == i)
Theta.hat <- apply(sapply(mod, coef), 1, mean)
B <- apply(sapply(mod, coef), 1, var)
W <- apply(sapply(mod, function(x) summary(x)$coefficients[, 2])^2, 1, mean)
Tot <- W + (4/3) * B
round((se <- sqrt(Tot)), 4)  # MI standard errors

## (Intercept)   education         age     sexMale
##      0.4514      0.0270      0.0084      0.3062

round((tval <- Theta.hat/se), 4)  # t value

## (Intercept)   education         age     sexMale
##    -16.7415     33.2733     30.2713     11.1103

nu <- (3 - 1) * (1 + W/(4/3 * B))^2  # 'classical' df
round((lambda <- (4/3) * B/Tot), 4)  # lambda (appr. FoMI)

## (Intercept)   education         age     sexMale
##      0.0449      0.1372      0.6776      0.7220

nu.com <- summary(mod[[1]])$df[2]
cm <- (nu.com + 1)/(nu.com + 3) * nu.com * (1 - lambda)
round((df <- nu * cm/(nu + cm)), 4)  # Barnard & Rubin corrected df

## (Intercept)   education         age     sexMale
##    870.1986    104.5642      4.3484      3.8299

r <- (4/3) * B/W  # relative increase in variance
round((fmi <- (r + 2/(df + 3))/(r + 1)), 4)  # Fraction of Missing Information

## (Intercept)   education         age     sexMale
##      0.0471      0.1532      0.7653      0.8034
```

Problems in empirical application settings: Implausible values

- Example 'age': Values less than 0 or more than 120 not plausible or near impossible
- Solution:
  - Transformation (e.g. log(Alter))
  - Software that allows to specify bounds (e.g. IVEware)
  - Nearest-Neighbour techniques
- Example 'work experience': could be greater than 'age'
- Solution:
  - create artificial variables (e.g. 'age' - 'work experience') and impute 'age' OR 'work experience' + artificial variable
  - Nearest-Neighbour techniques which impute for the complete unit vector

Problems in empirical application settings: Imputation model misspecification

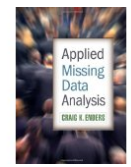
- functional form of the imputation model is incorrect
- wrong distributional assumption for error term
- Solution:
  - flexible (non- or semiparametric) models (Splines, GAM's, CART)
  - Nearest-Neighbour procedures

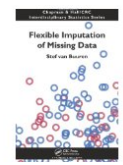
Problems in empirical application settings: Complex designs

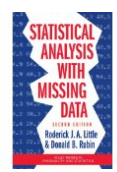
- Filter questions (e.g. 'Do you smoke?' if yes 'how many per day?')
- Logical dependencies (e.g. salary of welfare recipients)
- Multilevel data (e.g. schools, classes, students)
- Solution:
  - Software that allows to define filters (e.g. IVEware)
  - Stepwise imputation
  - 're-program' filter
  - segmentation of sample in combination with bounds
  - paneldata imputation models??

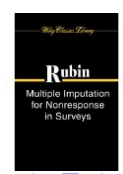
Background Reading: Missing Data Analysis

- Raghunathan, T.E. (2015) Missing Data Analysis in Practice. Chapman & Hall/CRC, Boca Raton.
- Enders, C.K. (2010) Applied Missing Data Analysis. Guilford, New York.
- van Buuren, S. (2012) Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton.
- Carpenter, J.R. (2013) Multiple Imputation and its Application. John Wiley and Sons, New York.
- Allison, P. (2001) Missing Data Analysis. Sage, Thousand Oaks.
- Little, R.J.A., Rubin, D.B. (1987, 2002) Statistical Analysis with Missing Data. John Wiley and Sons, New York.
- Rubin, D.B. (1987, 2004) Multiple Imputation for Nonresponse in Surveys. Wiley, New York.









THE END


UNIVERSITY of BAMBERG
OTTO-FRIEDRICH-UNIVERSITÄT BAMBERG

Barnard, J. & Rubin, D. B. (1999), 'Small-Sample Degrees of Freedom with Multiple Imputation', *Biometrika* **86**, 948–955.

Dempster, A. P., Laird, N. & Rubin, D. B. (1977), 'Maximum Likelihood from Incomplete Data via the EM Algorithm', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **39**, 1–38.

Enders, C. K. (2010), *Applied missing data analysis*, Methodology in the social sciences, Guilford Press, New York.

Little, R. J. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2 edn, Wiley, New York.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

Rubin, D. B. (1978), Multiple Imputation in Sample Surveys – A Phenomological Bayesian Approach to Nonresponse, *in* 'Proceedings of the Survey Research Method Section of the American Statistical Association', pp. 20–40.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Rubin, D. B. & Schenker, N. (1986), 'Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse', *Journal of the American Statistical Association* **81**, 366–374.

Schafer, J. L. (1999), 'NORM – Multiple Imputation under a Normal Model, version 2.03'.

Templ, M., Alfons, A., Kowarik, A. & Prantner, B. (2016), 'VIM: Visualization and Imputation of Missing Values'.
URL: *https://CRAN.R-project.org/package=VIM*

van Buuren, S. (2012), *Flexible Imputation of Missing Data*, Chapman & Hall/CRC interdisciplinary statistics series, CRC Press, Boca Raton, FL.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011), 'mice: Multivariate Imputation by Chained Equations in R', *Journal of Statistical Software* **45**(3), 1–67.
URL: *http://www.jstatsoft.org/v45/i03/*