Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Dealing with missing values – part 2

Applied Multivariate Statistics – Spring 2012

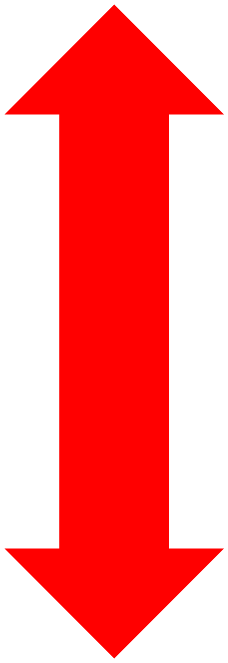
Overview

- More on Single Imputation: Shortcomings
- Multiple Imputation: Accounting for uncertainty

Single Imputation

- Unconditional Mean
- Unconditional Distribution
- Conditional Mean
- Conditional Distribution

Easy / Inaccurate



Hard / Accurate

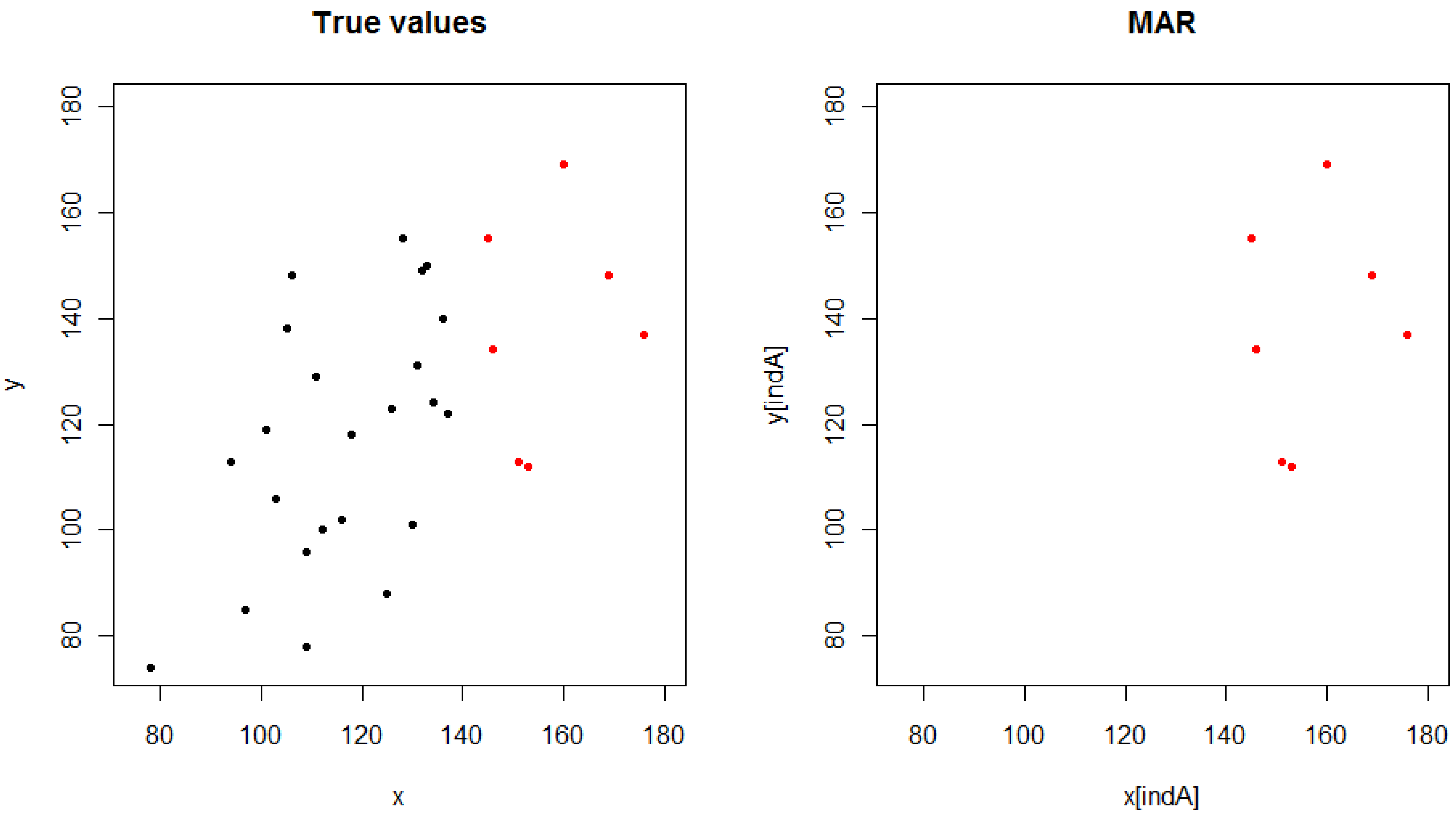
Example: Blood Pressure - Revisited

- 30 participants in January (X) and February (Y)
- MCAR: Delete 23 Y values randomly
- MAR: Keep Y only where X > 140 (follow-up)
- MNAR: Record Y only where Y > 140 (test everybody again but only keep values of critical participants)

	Y			
X	Complete	MCAR	MAR	MNAR
Data for individual participants				
169	148	148	148	148
126	123	—	—	—
132	149	—	—	149
160	169	—	169	169
105	138	—	—	—
116	102	—	—	—
125	88	—	—	—
112	100	—	—	—
133	150	—	—	150
94	113	—	—	—
109	96	—	—	—
109	78	—	—	—
106	148	—	—	148
176	137	—	137	—
128	155	—	—	155
131	131	—	—	—
130	101	101	—	—
145	155	—	155	155
136	140	—	—	—
146	134	—	134	—
111	129	—	—	—
97	85	85	—	—
134	124	124	—	—
153	112	—	112	—
118	118	—	—	—
137	122	122	—	—
101	119	—	—	—
103	106	106	—	—
78	74	74	—	—
151	113	—	113	—

Black points are missing (MAR)

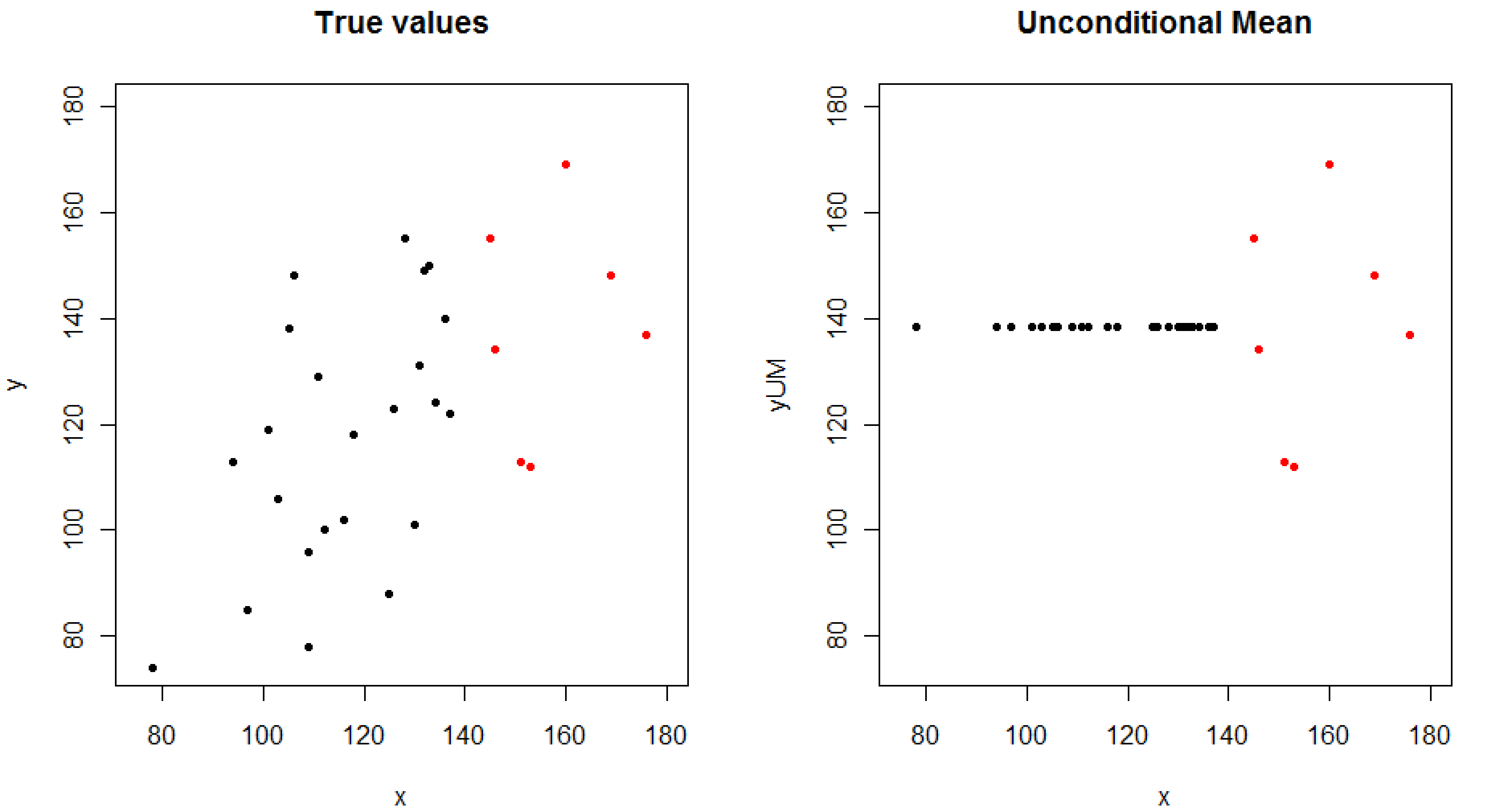
Example: Blood Pressure



+ Mean of Y ok

- Variance of Y wrong

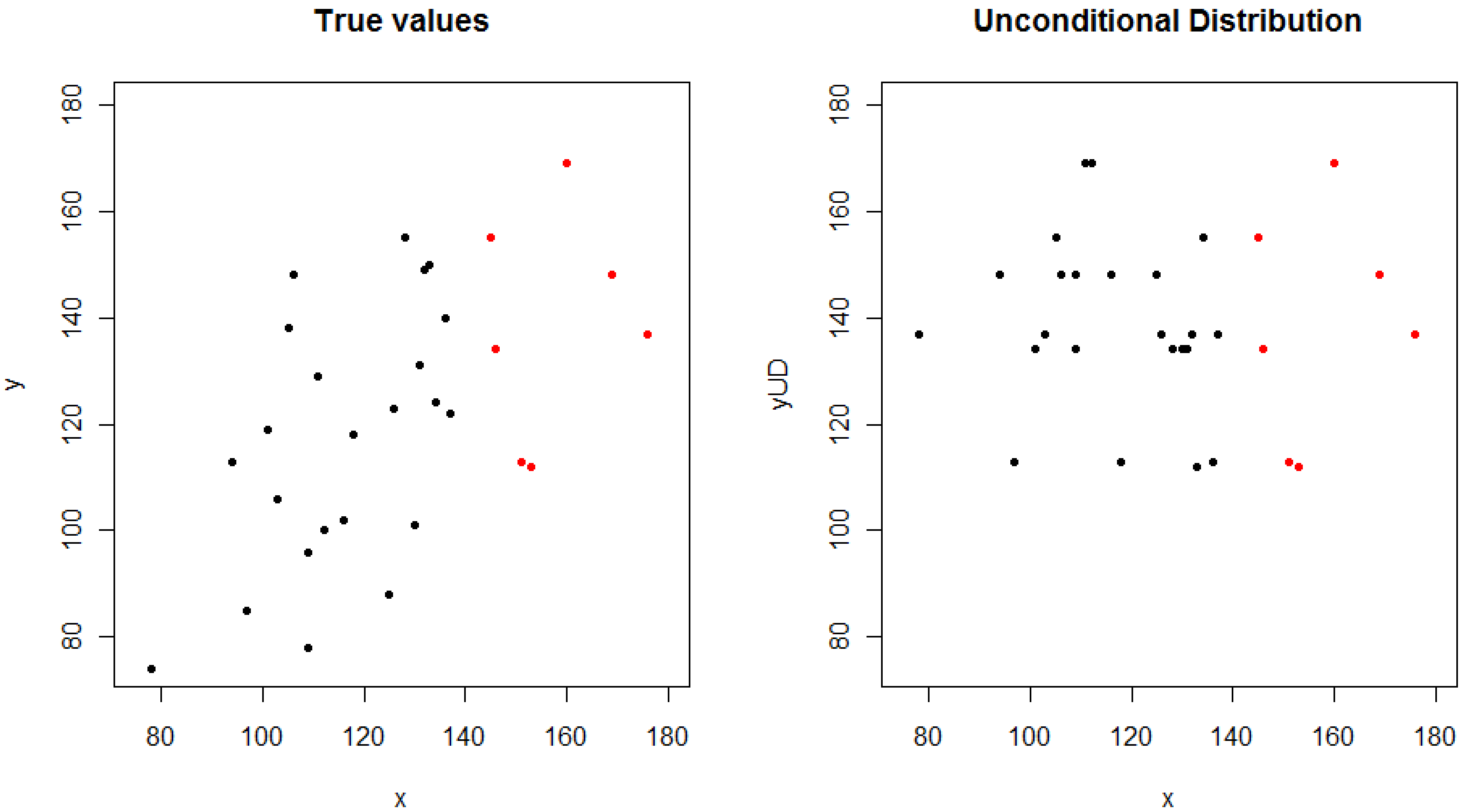
Unconditional Mean



+ Mean of Y ok, Variance better

- Correlation btw X and Y wrong

Unconditional Distribution

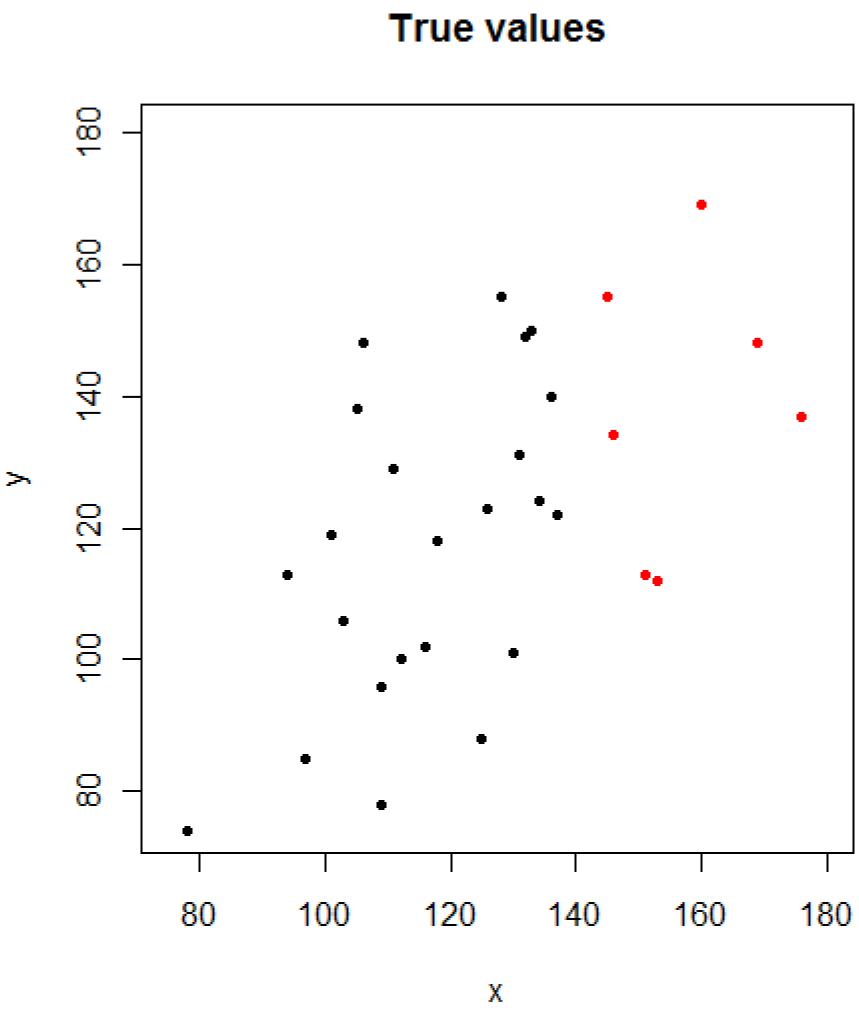


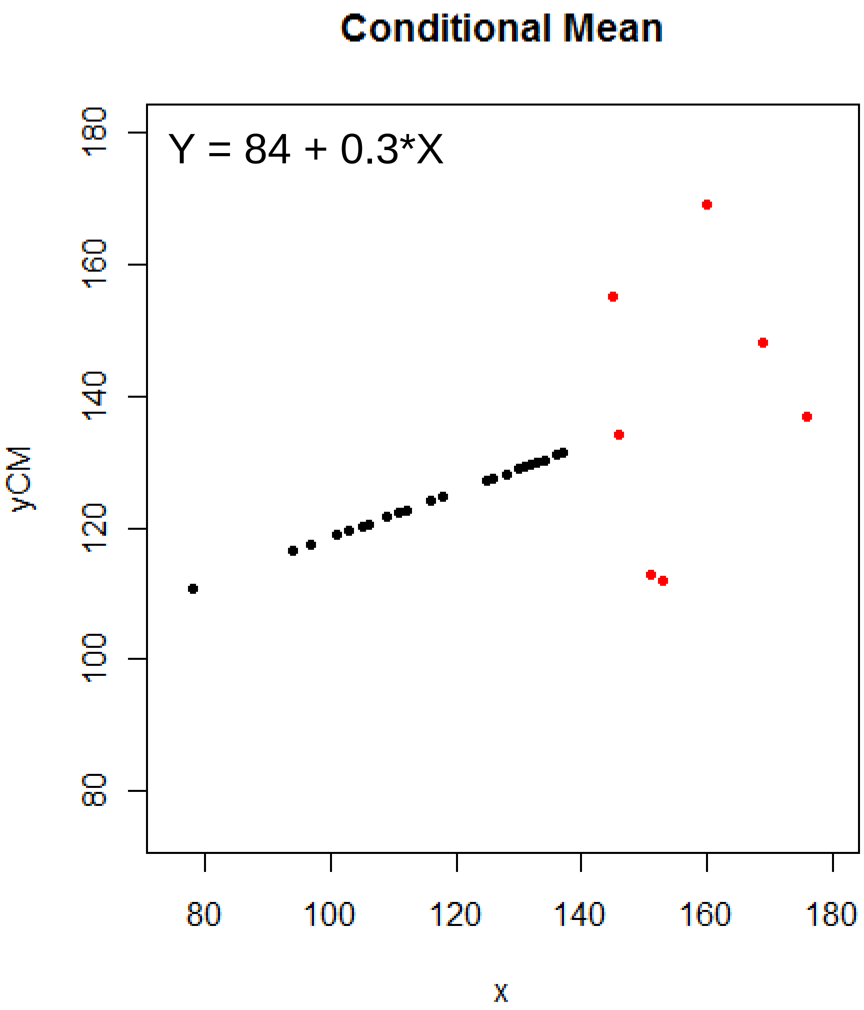
+ Conditional Mean of Y ok

+ Correlation ok

- (Conditional) Variance wrong

Conditional Mean



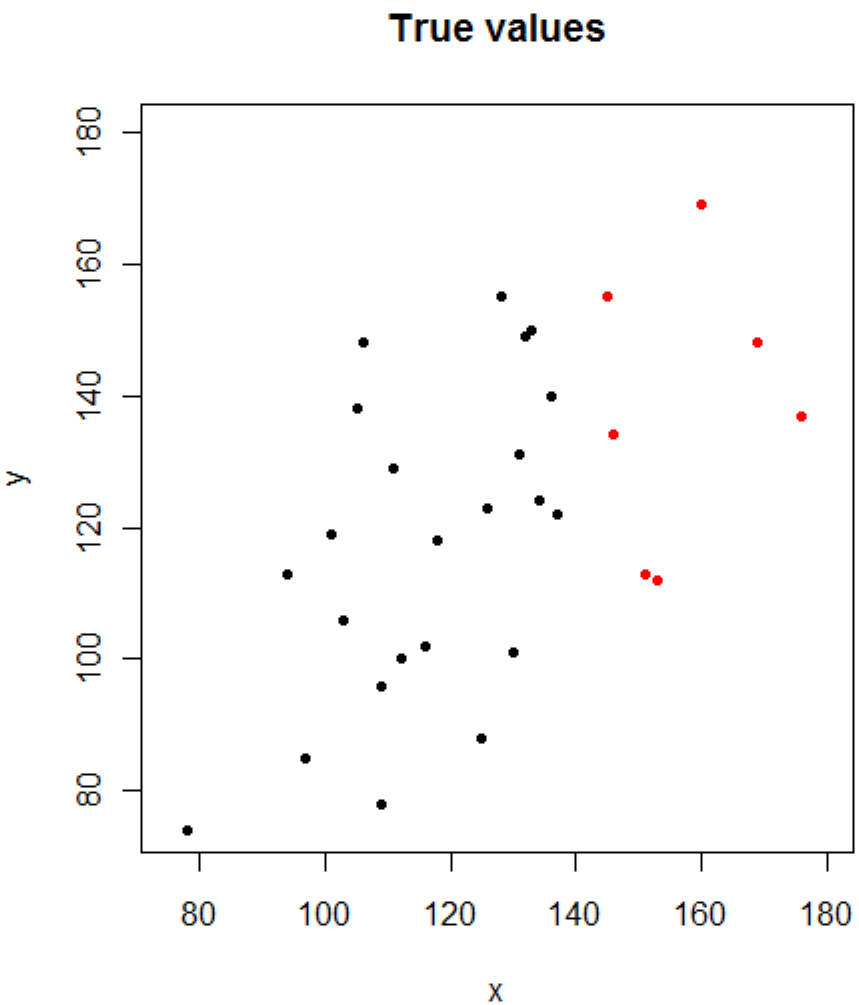


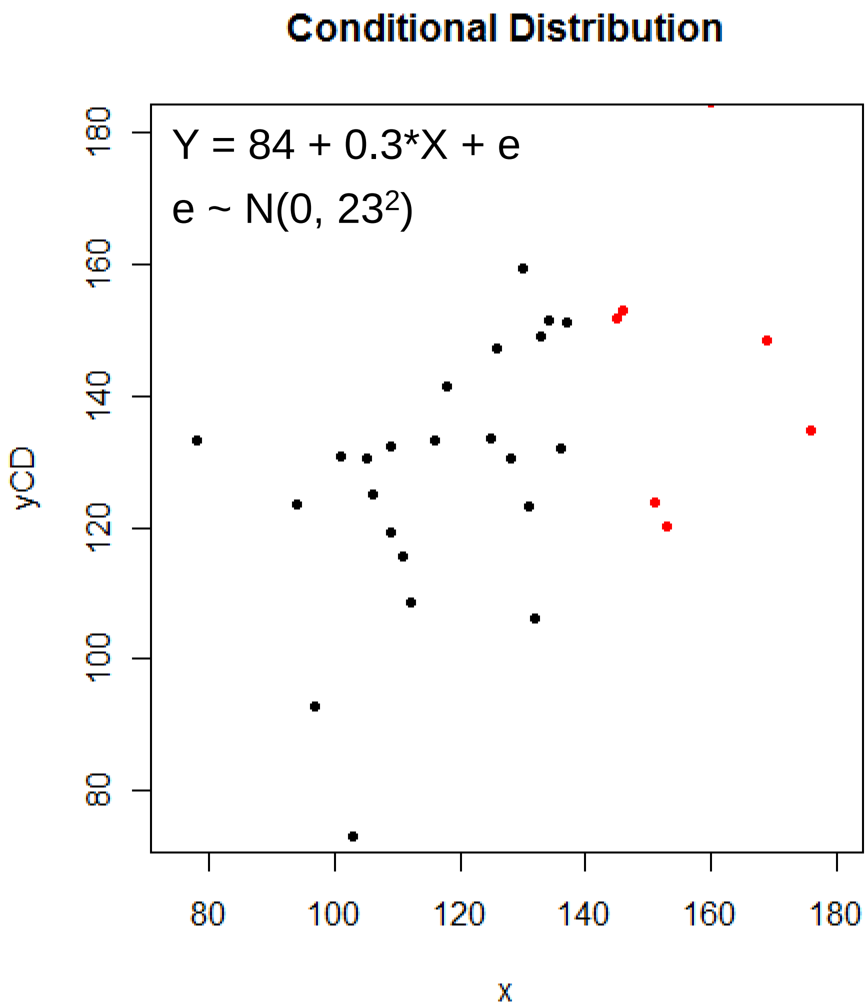
+ Conditional Mean of Y ok

+ Correlation ok

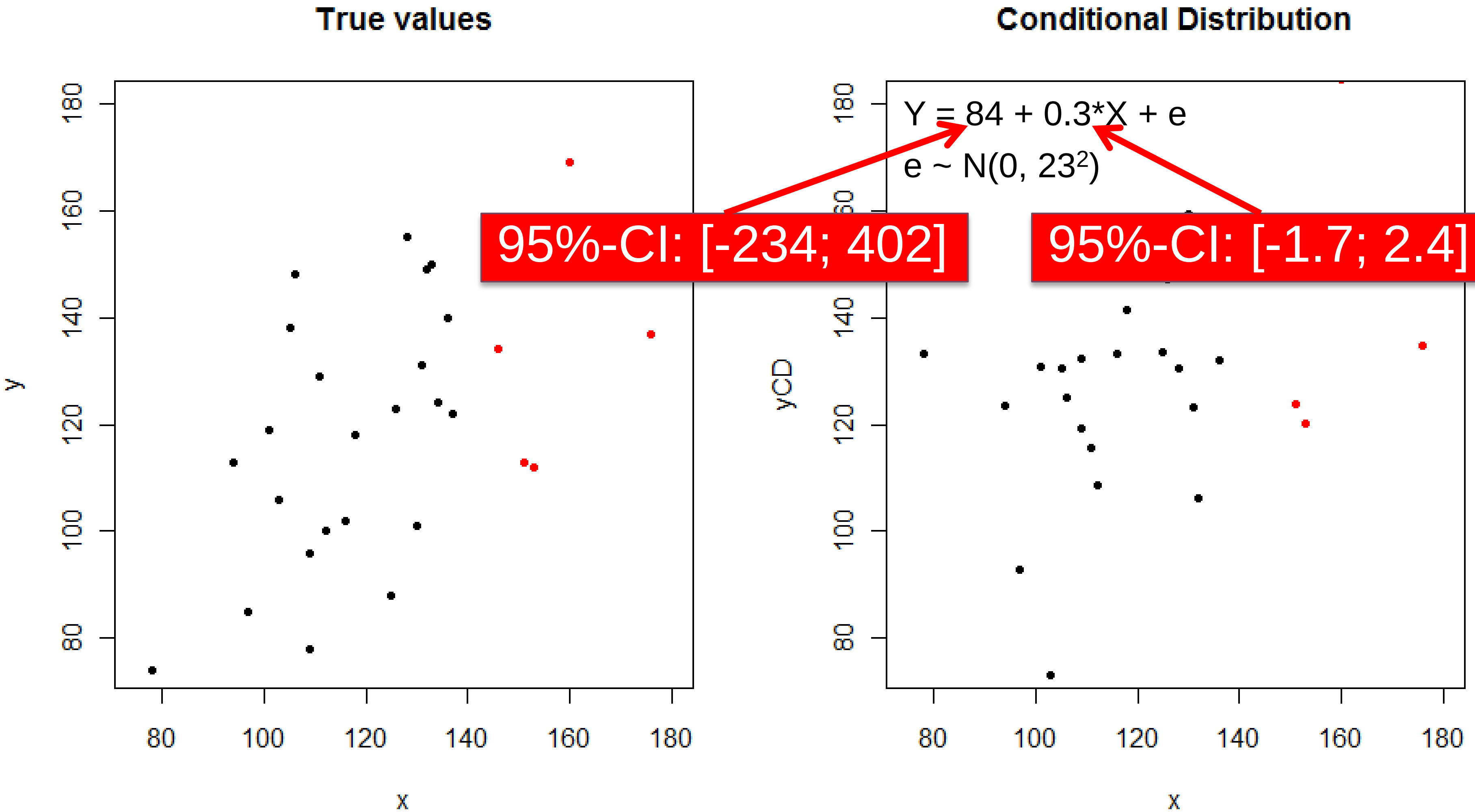
+ Conditional Variance of Y ok

Conditional Distribution





Problem: We ignore uncertainty



95%-CI: [-234; 402]

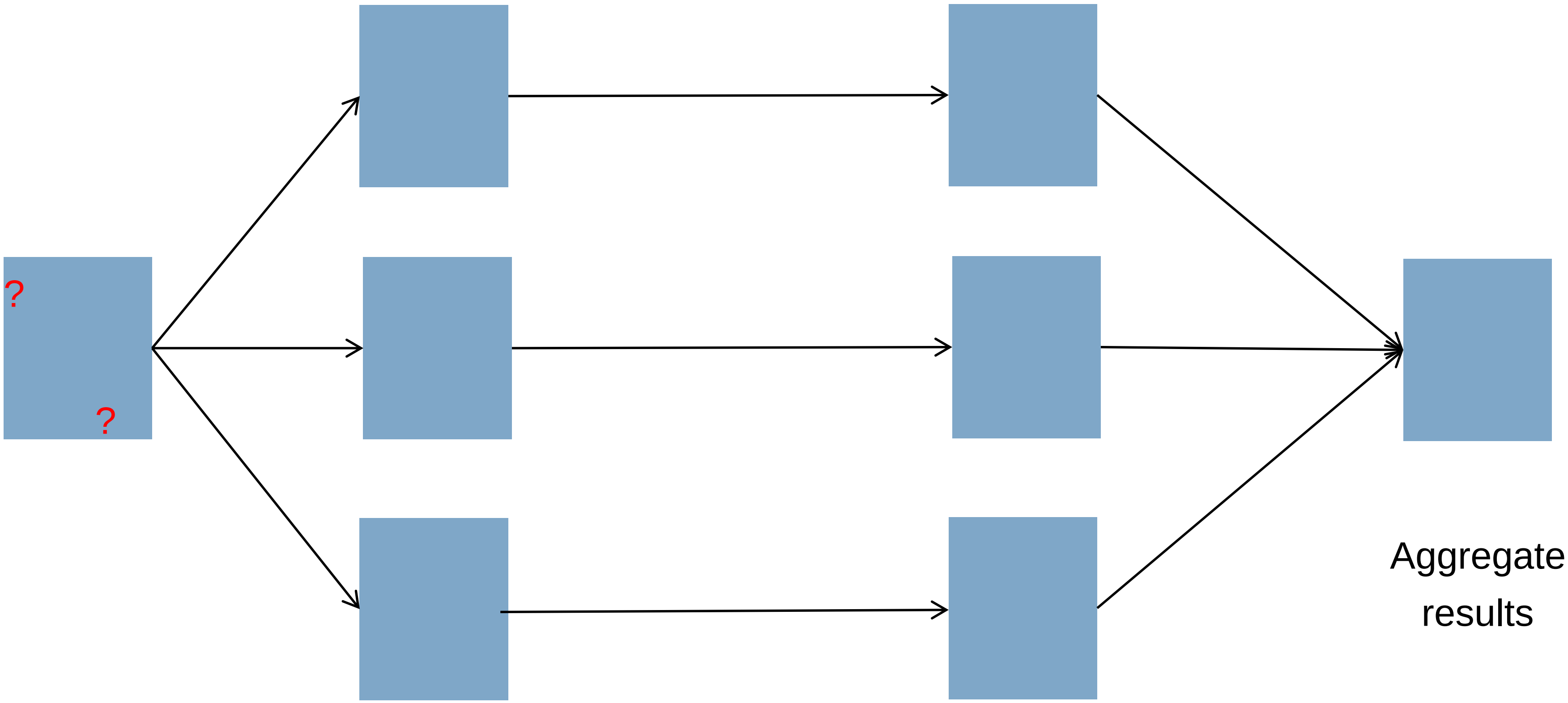
95%-CI: [-1.7; 2.4]

Problem of Single Imputation

- Too optimistic: Imputation model (e.g. in Y = a + bX) is just estimated, but not the true model
- Thus, imputed values have some uncertainty
- Single Imputation ignores this uncertainty
- Coverage probability of confidence intervals is wrong

- Solution: Multiple Imputation
  Incorporates both
  - residual error
  - model uncertainty (excluding model mis-specification)

Multiple Imputation: Idea

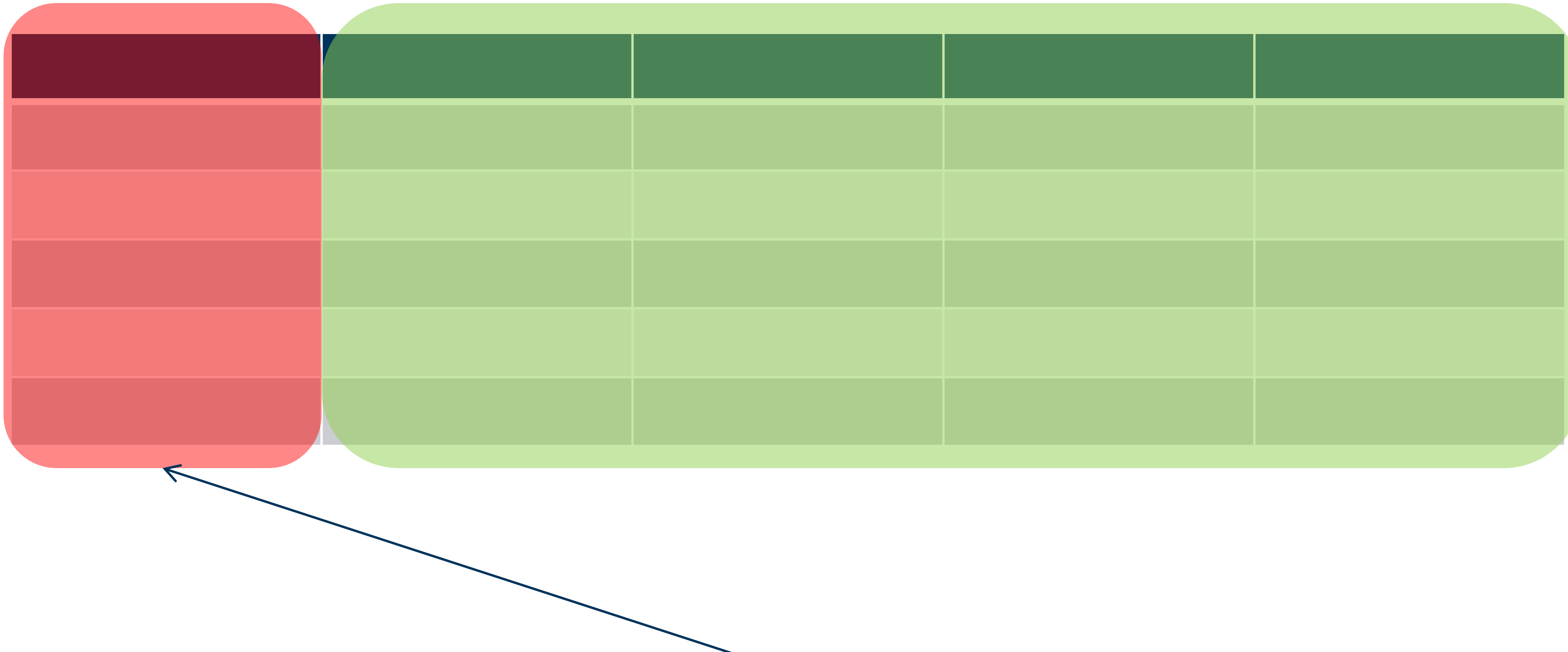


Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Multiple Imputation: Idea

- Need special imputation schemes that include both
  - uncertainty of residuals
  - uncertainty of model
  (e.g. values of intercept a and slope b)
- Rough idea:
  - Fill in random values
  - Iteratively predict values for each variable until some convergence is reached (as in missForest)
  - Sample values for residuals AND for (a,b)
- Gibbs sampler is used
- Excellent for intuition (by one of the big guys in the field): http://sites.stat.psu.edu/~jls/mifaq.html

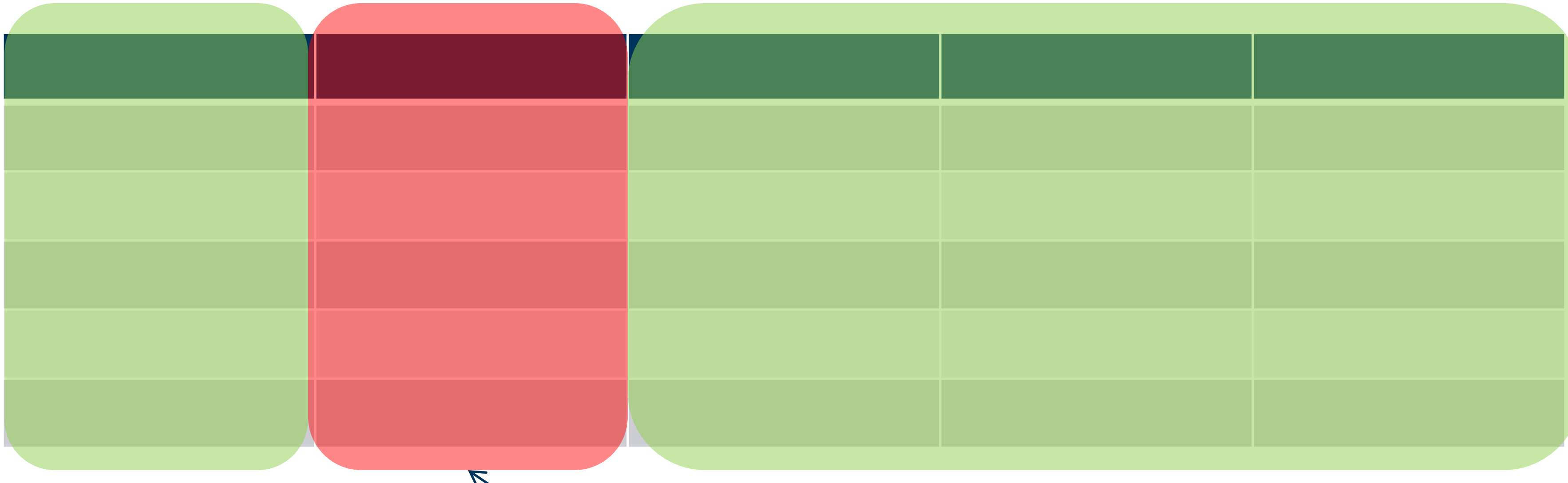
Multiple Imputation: Intuition



Predict missing values accounting for

- Uncertainty of residuals
- Uncertainty of parameter estimates

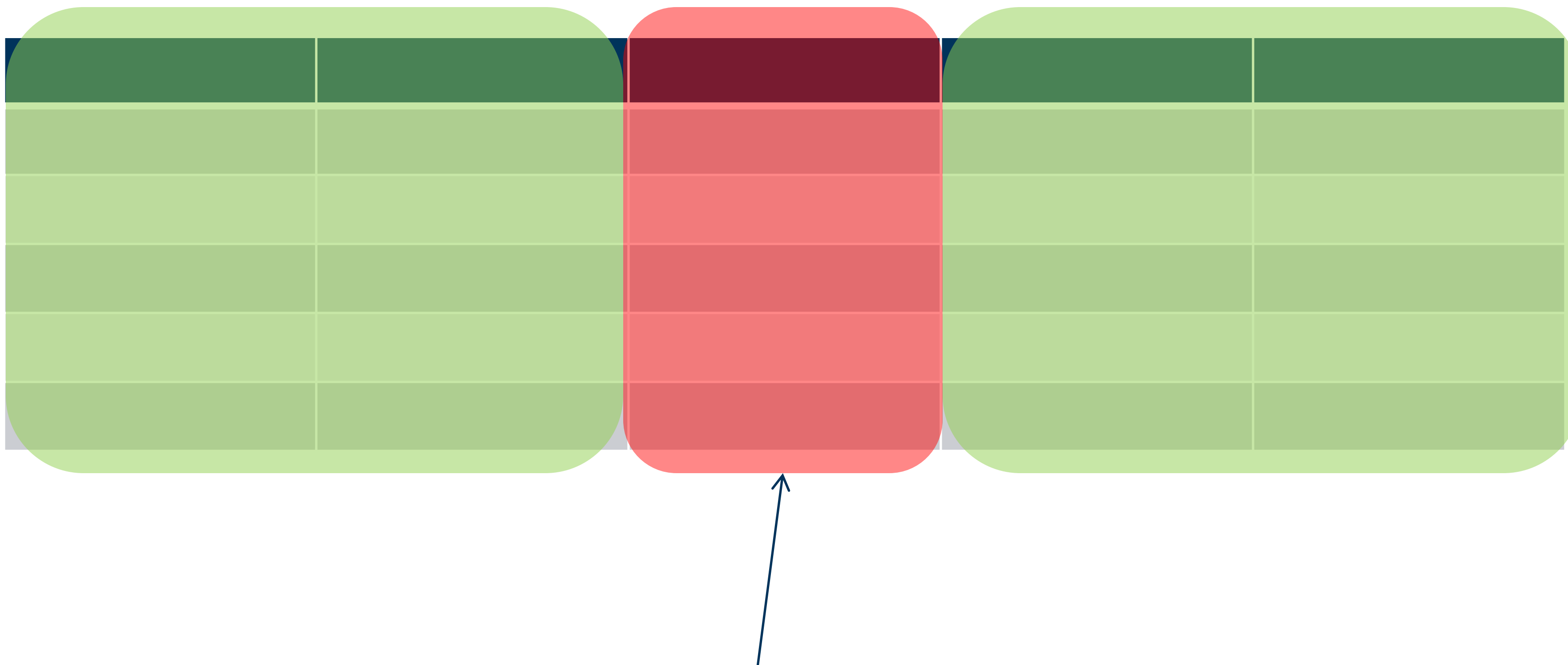
Multiple Imputation: Intuition



Predict missing values accounting for

- Uncertainty of residuals
- Uncertainty of parameter estimates

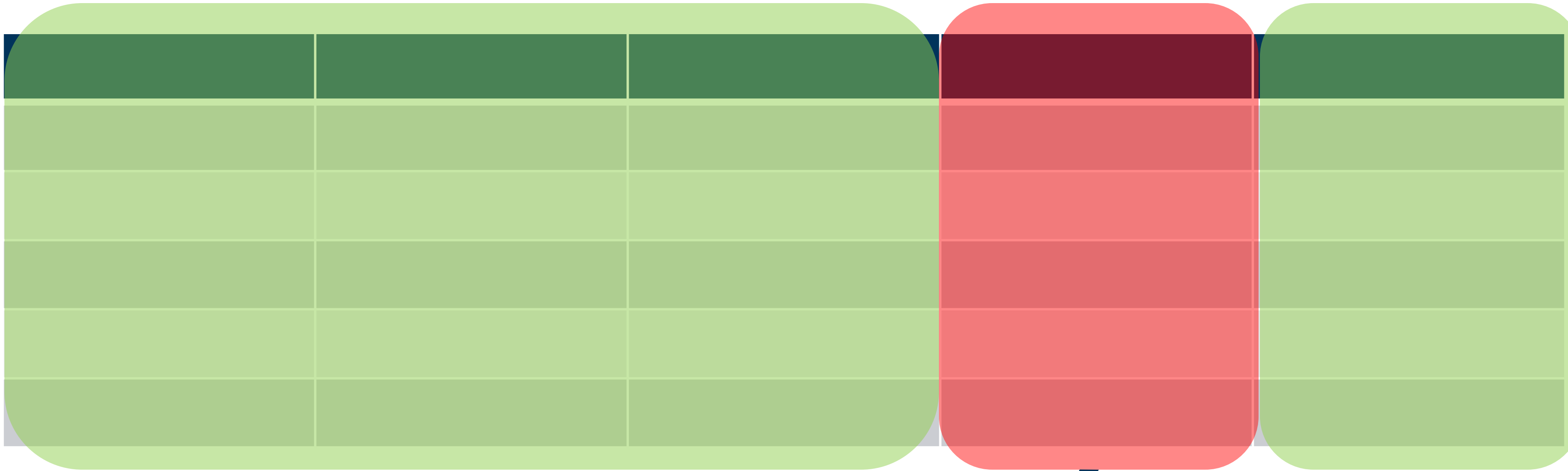
Multiple Imputation: Intuition



Predict missing values accounting for

- Uncertainty of residuals
- Uncertainty of parameter estimates

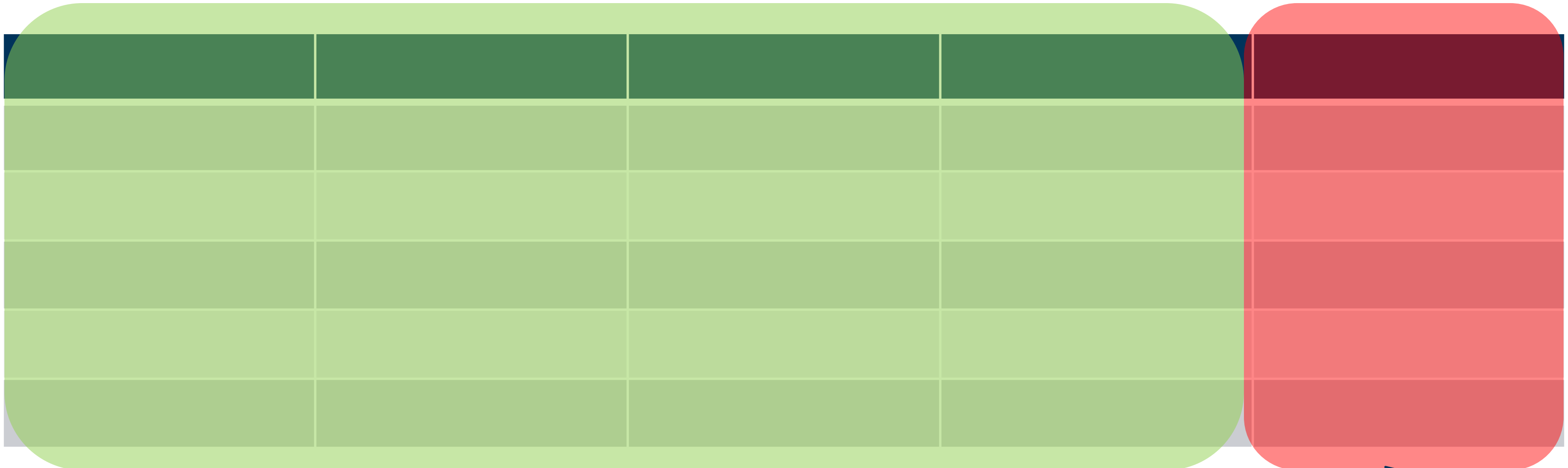
Multiple Imputation: Intuition



Predict missing values accounting for

- Uncertainty of residuals
- Uncertainty of parameter estimates

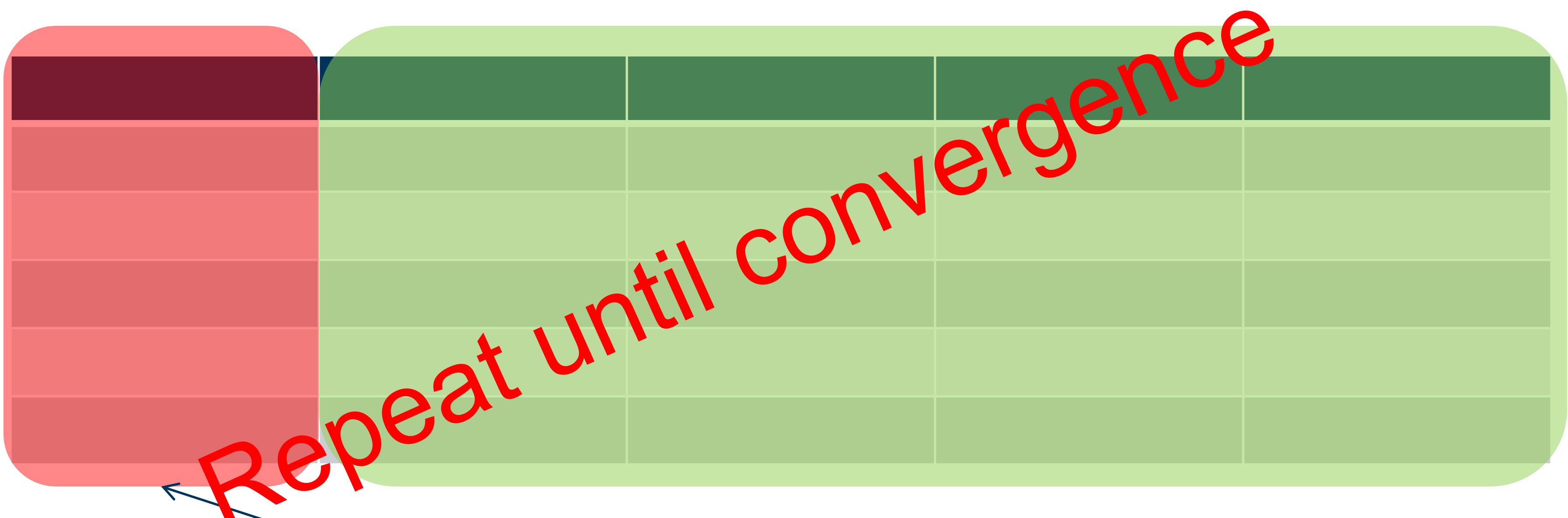
Multiple Imputation: Intuition



Predict missing values accounting for

- Uncertainty of residuals
- Uncertainty of parameter estimates

Multiple Imputation: Intuition



Predict missing values accounting for

- Uncertainty of residuals
- Uncertainty of parameter estimates

Multiple Imputation: Gibbs sampler (Not for exam)

- Iteration t; repeat until convergence:
  For each variable i:

Intuition

$$\theta_i^{*(t)} \sim P(\theta_i | Y_i^{obs}, Y_{-i}^{(t)})$$

Sample (a,b)

$$Y_i^{*(t)} \sim P(Y_i | Y_i^{obs}, Y_{-i}^{(t)}, \theta_i^{*(t)})$$

Predict missings using
y = a + bx + e

"Chained Equations"

where $Y_i^{(t)} = (Y_i^{obs}, Y_j^{*(t)})$

R package: MICE
Multiple Imputation with Chained Equations

- MICE has good default settings; don't worry about the data type
- Defaults for data types of columns:
  - numeric: Predictive Mean Matching (pmm)
  (like fancy linear regression; faster alternative: linear regression)
  - factor, 2 lev: Logistic Regression (logreg)
  - factor, >2 lev: Multinomial logit model (polyreg)
  - ordered, >2 lev: Ordered logit model (polr)

Aggregation of estimates

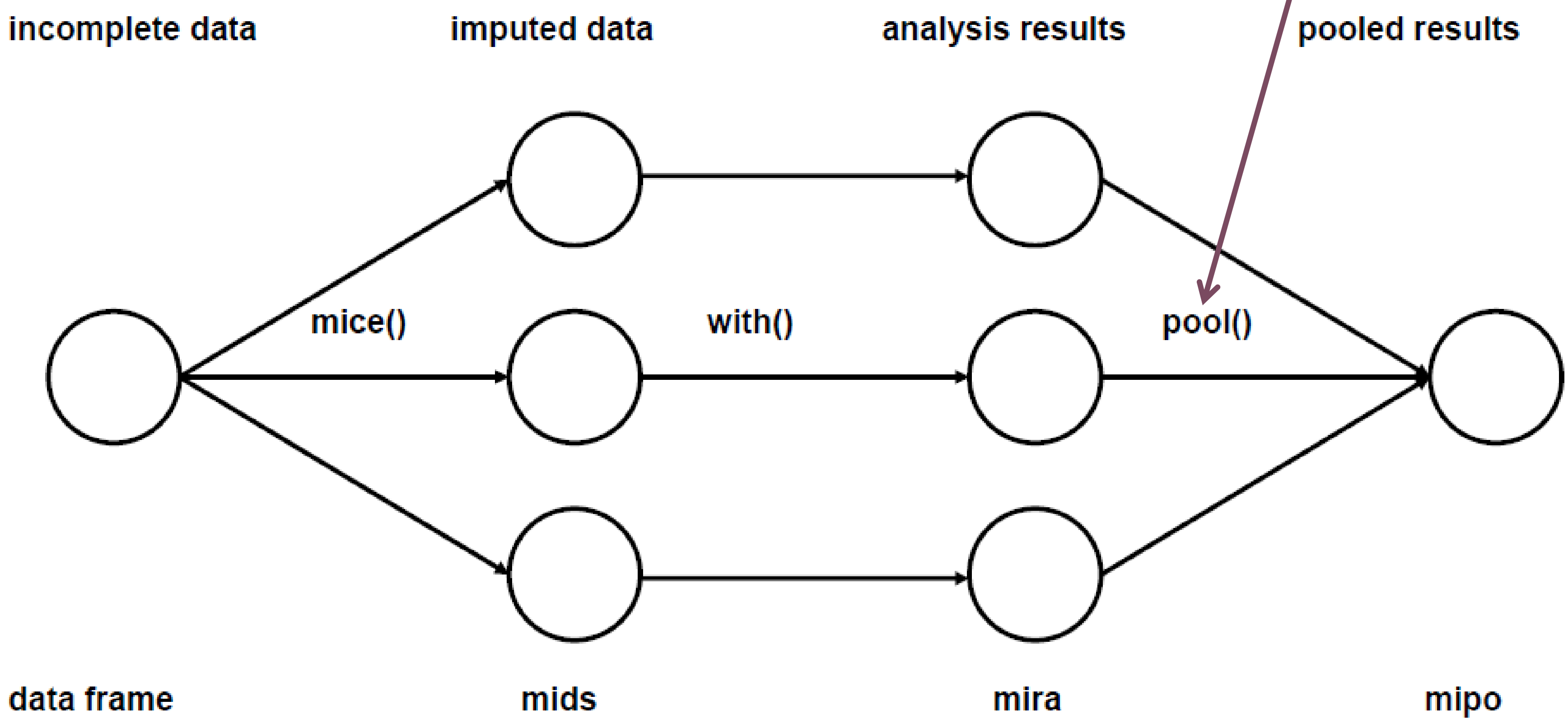
- $\hat{Q}_i$: Estimate of imputation i
  U_i : Variance of estimate (= square of std. error)
- Assume: $\frac{\hat{Q}-Q}{\sqrt{U}} \approx N(0,1)$
- Average estimate: $\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j$

- Within-imputation variance: $\bar{U} = \frac{1}{m}\sum_{j=1}^{m}\hat{U}_j$
- Between-imputation variance: $B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Q}_j - \bar{Q})^2$
- Total variance: $T = \bar{U} + \frac{1}{m-1}B$

- Approximately: $\frac{\bar{Q}-Q}{\sqrt{T}} \sim t_\nu$ with $\nu = (m-1)\left(1 + \frac{m\bar{U}}{(1+m)B}\right)^2$
- 95%-CI: $\bar{Q} \pm t_{\nu;0.975}\sqrt{T}$

Multiple Imputation with MICE

Do manually, if you have non standard analysis



How much uncertainty due to missings?

- Relative increase in variance due to nonrespose:

$$r = \frac{(1+\frac{1}{m})B}{U}$$

- Fraction (or rate) of missing information fmi:
(!! Not the same as fraction of missing OBSERVATIONS)

$$fmi = \frac{r+\frac{2}{\nu+3}}{r+1}$$

- Proportion of the total variance that is attributed to the missing data:

$$\lambda = \frac{B(1+\frac{1}{m})}{T}$$

Returned by mice

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

How many imputations?

Rule of thumb:

- Preliminary analysis: m = 5
- Paper: m = 20 or even m = 50

- Surprisingly few!
- Efficiency compared to $m = \infty$ depends on fmi:

$$eff = \left(1 + \frac{fmi}{m}\right)^{-1}$$

- Examples (eff in %):

Oftentimes OK (→ M = 5)

Perfect ! (→ M = 20)

M	fmi=0.1	fmi=0.3	fmi=0.5	fmi=0.7	fmi=0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Concepts to know

- Idea of mice
- How to aggregate results from imputed data sets?
- How many imputations?

R functions to know

- mice, with, pool

Next time

- Multidimensional Scaling
- Distance metrics