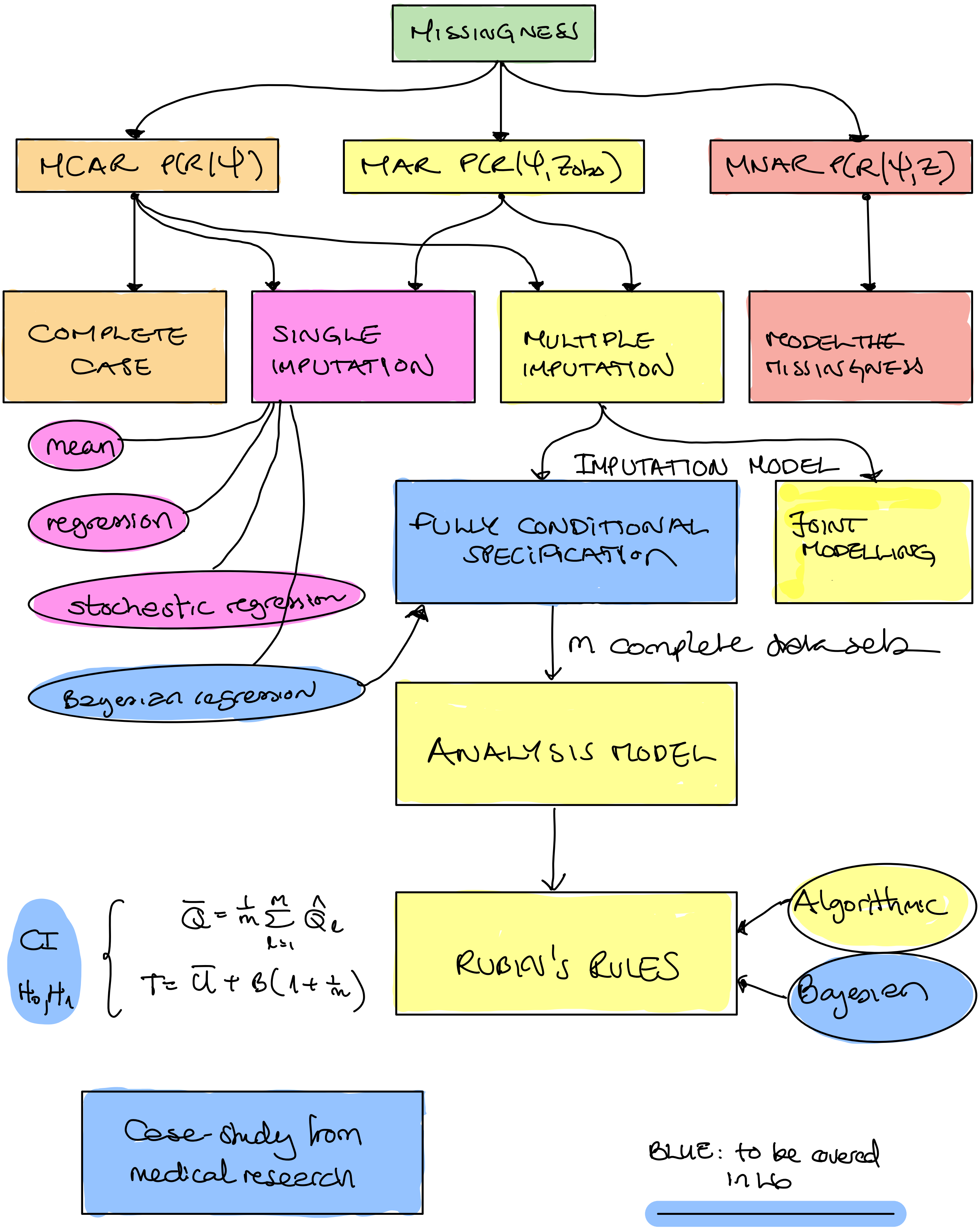
MA8701 Advanced methods in statistical inference and learning

Week 3 ~~(L5+L6)~~: Missing data L6

Mette Langaas

~~1/26/23~~ 27.01.2023


MISSINGNESS
MCAR $P(R|Y)$
MAR $P(R|Y, Z_{obs})$
MNAR $P(R|Y, Z)$
COMPLETE CASE
SINGLE IMPUTATION
MULTIPLE IMPUTATION
MODEL THE MISSINGNESS
mean
regression
stochastic regression
Bayesian regression
IMPUTATION MODEL
FULLY CONDITIONAL SPECIFICATION
JOINT MODELLING
m complete data sets
ANALYSIS MODEL
RUBIN'S RULES
Algorithmic
Bayesian
CI
H_0, H_1
$\bar{Q} = \frac{1}{m}\sum_{l=1}^{M} \hat{Q}_l$
$T = \bar{U} + B(1 + \frac{1}{m})$
Case-study from medical research
BLUE: to be covered in L6

Multiple imputation as approximate Bayesian inference

Aim: inference about Q $\leftarrow \beta$

Bayesian inference is based on:

combining prior information $P(Q)$

and data likelihood $P(Y \mid Q, X)$ ↙ frequentists only use this

into the posterior $P(Q \mid Y, X) \propto P(Q) \cdot P(Y \mid Q, X)$

↗ Bayesians use this

INTEREST: $P(Q \mid X_{obs}, Y, R)$

using the law of total probability

→ condition on X_{mis}

$$= \boxed{\int_{X_{mis}}} P(Q|X_{obs}, Y, \cancel{R}, X_{mis}) \cdot \boxed{P(X_{mis}|X_{obs}, Y, R)}\, dX_{mis}$$

$$P(Q) = \int_{X_{mis}} P(Q|X_{mis}) \cdot P(X_{mis})\, dX_{mis}$$

also condition on X_{obs}, Y, R

If we can generate X_{mis} then we can make m full data data

$$P(Q|X_{obs}, Y, R) \approx \frac{1}{m} \sum_{l=1}^{m} P(Q|X_{obs}, X_{mis}, Y)$$

$$E(Y) = E(E(Y|X))$$

Often sufficient to work with the first two moments

$$E(Q|X_{obs}, Y, R) = E\left[\underbrace{E(Q|X_{obs}, Y, \cancel{R}, X_{mis})}_{\text{complete data estimate}} \,\middle|\, X_{obs}, Y, R\right]$$

$$\approx \frac{1}{m} \sum_{l=1}^{m} \hat{Q}_l = \bar{Q}$$ pooled Rubin's rule

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$$

↓

$$\mathrm{Var}(Q \mid X_{obs}, Y, R) = E\Big(\underbrace{\mathrm{Var}(Q \mid X_{obs}, Y, X_{mis})}_{\bar{U}_\ell} \Big| X_{obs}, Y, R \Big) + \mathrm{Var}\Big(\underbrace{E(Q \mid X_{obs}, Y, X_{mis})}_{\hat{Q}_\ell} \Big| X_{obs}, Y, R \Big)$$

$$\approx \underbrace{\frac{1}{m} \sum_{\ell=1}^{m} \bar{U}_\ell}_{\bar{U}} + \underbrace{\frac{1}{m-1} \sum_{\ell=1}^{m} (\hat{Q}_\ell - \bar{Q})(\hat{Q}_\ell - \bar{Q})^T}_{B}$$

If m is small we add $\frac{B}{m}$ to reflect the uncertainty in $\bar{Q}$ as an estimate of $E(Q \mid Y, X_{obs}, R)$

$$\Rightarrow \quad T = \bar{U} + B\left(1 + \frac{1}{m}\right)$$

Rubin's rules

What can Rubin's rules be used on?

For inference (see above for CI and p-value using t- and Fisher distribution) the assumption is that $\overline{Q}$ is approximately multivariate normal.

- Regression parameters in multiple linear regression
- Regression parameters in logistic regression
- Correlations: but use Fishers z-transform to become more normally distributed
- ROC-AUC
- Recently also predictions from the analysis model

Confidence interval

Common assumption: $\overline{\mathbf{Q}}$ is multivariate normal with mean $\mathbf{Q}$ and estimated covariance matrix $\mathbf{T}$.

We look at one component of $\mathbf{Q}$, denoted Q (maybe regression parameter for a specific covariate), antd T is the appropriate component of the total variance estimate.

$(1-\alpha)100\%$ confidence interval for Q:

$$\overline{Q} \pm t_{\nu,1-\alpha/2}\sqrt{T}$$

where $t_{\nu,1-\alpha/2}$ is the value in the t-distribution with ν degrees of freedom with area $1-\alpha/2$ to the left.

What is ν?

Hypothesis test

We want to test $H_0 : Q = Q_0$ vs $H_1 : Q \neq Q_0$. The p-value of the test can be calculated as

$$P(F_{1,\nu} > \frac{(\overline{Q} - Q_0)}{T})$$

where $F_{1,\nu}$ is a random variable following a Fisher distribution with 1 and ν degrees of freedom.

Read for yourself

→ used for df in t and F-test !

Variance ratios

for scalar Q (for example one of the regression coefficients)

Proportion of variation “attributable” to the missing data

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

Relative increase in variance due to missingness

$$r = \frac{B + B/m}{\overline{U}}$$

Relation:

$$r = \frac{\lambda}{1 - \lambda}$$

Degrees of freedom

van Buuren (2018) Chapter 2.3.6 attributed this first solution to Rubind in 1987.

$$\nu_{\text{old}} = (m-1)(1+\frac{1}{r^2}) = \frac{m-1}{\lambda^2}$$

If $\lambda = 1$ then all variability is due to the missingness and then $\nu_{\text{old}} = m - 1$.
If $\lambda \to 0$ then $\nu_{\text{old}} \to \infty$ (normal distribution instead of t, chisq instead of F).

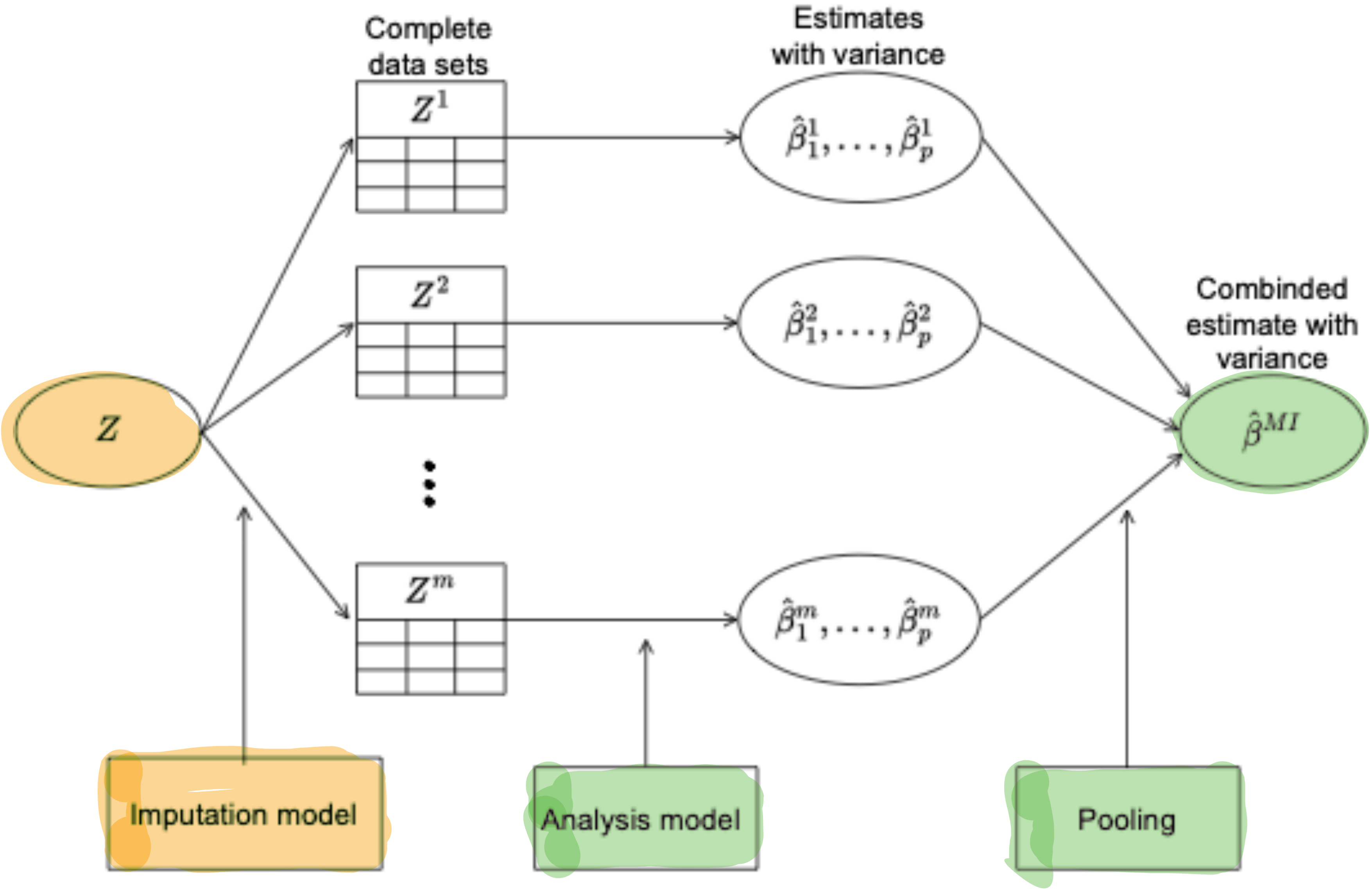
van Buuren (2018) Chapter 2.3.6: A newer solution is due to Barnard and Rubin in 1999.

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

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

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

Schematic for multiple imputation from Marthe Bøe Ludvigsen project thesis.


Complete data sets
Estimates with variance
Z^1
$\hat{\beta}_1^1, \ldots, \hat{\beta}_p^1$
Z^2
$\hat{\beta}_1^2, \ldots, \hat{\beta}_p^2$
Combinded estimate with variance
Z
$\hat{\beta}^{MI}$
Z^m
$\hat{\beta}_1^m, \ldots, \hat{\beta}_p^m$
Imputation model
Analysis model
Pooling

Single imputation

$$\underset{N\times 1}{Y} = \underset{N\times q}{X}\,\underset{q\times 1}{\phi} + \underset{N\times 1}{\varepsilon} \qquad \varepsilon \sim N(0,\sigma^2 I)$$

missing elements

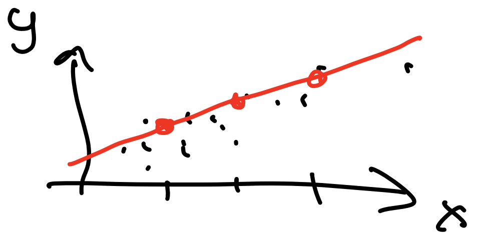
all observed

$$y = \begin{bmatrix} y_{obs} \\ y_{mis} \end{bmatrix} \begin{matrix} n_1 \\ n_0 \end{matrix} \qquad X = \begin{bmatrix} \overset{n_1\times q}{X_{obs}} \\ \underset{n_0\times q}{X_{mis}} \end{bmatrix}$$

AIM: construct imputed values for $\dot{y}_{mis}$

1) Regression imputation: $\underset{n_0\times 1}{\dot{y}_{mis}} = \underset{n_0\times q}{X_{mis}}\,\underset{q\times 1}{\hat{\phi}}$

$$\hat{\phi} = (X_{obs}^T X_{obs})^{-1} X_{obs}^T y_{obs}$$



2) Stochastic regression imp.

$$\dot{y}_{mis} = X_{mis}\hat{\phi} + \dot{\varepsilon} \quad \text{where} \quad \dot{\varepsilon} \sim N(0,\hat{\sigma}^2)$$

drawn from

$$\hat{\sigma}^2 = \frac{1}{n_1-q}(y_{obs}-X_{obs}\hat{\phi})^T(y_{obs}-X_{obs}\hat{\phi})$$

3) Bayesian method

Likelihood: $p(y \mid x, \phi, \sigma^2) \propto (\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\phi)^T(y - X\phi)\right)$

conjugate prior: $p(\phi, \sigma^2) = \underbrace{p(\phi \mid \sigma^2)}_{N(\mu_0,\, \sigma^2 \Lambda_0^{-1})} \cdot \underbrace{p(\sigma^2)}_{\text{inverse gamma}}$

Λ_0 = precision matrix

inverse of coverance matrix

Posterior: $p(\phi, \sigma^2 \mid y, x) \propto \underbrace{p(y \mid x, \phi, \sigma^2) \cdot p(\phi \mid \sigma^2) \cdot p(\sigma^2)}_{\text{normal-inv. gamma}}$

$E(\phi, \sigma^2 \mid y, x)$: $\begin{bmatrix} \mu_n \\ E(\sigma^2) \end{bmatrix}$

$$\mu_N = \overbrace{(X^TX + \Lambda_0)}^{\Lambda_n}{}^{-1}(X^TX\hat{\phi} + \Lambda_0\mu_0)$$

$\mathrm{Var}(\phi .. \mid ..) = \sigma^2 \Lambda_n^{-1}$

New method

$$\dot{y}_{mis} = X_{mis} \cdot \dot{\phi} + \dot{\varepsilon} \qquad \text{and} \qquad \dot{\varepsilon} \sim N(0, \dot{\sigma}^2)$$

where $\dot{\phi}$ and $\dot{\sigma}^2$ drawn from posterior above

Frequentist view: uncertainty in ϕ taken into account

Summary: if the imputation model is a multiple linear regression model, and all covariates in the model are known (only missing values in the covariate that we make our target response), we know how to draw new observations to impute the missing values - and we may construct many imputed data sets.

Predictive mean matching PMM

Add a post processing step to the Bayesian method -

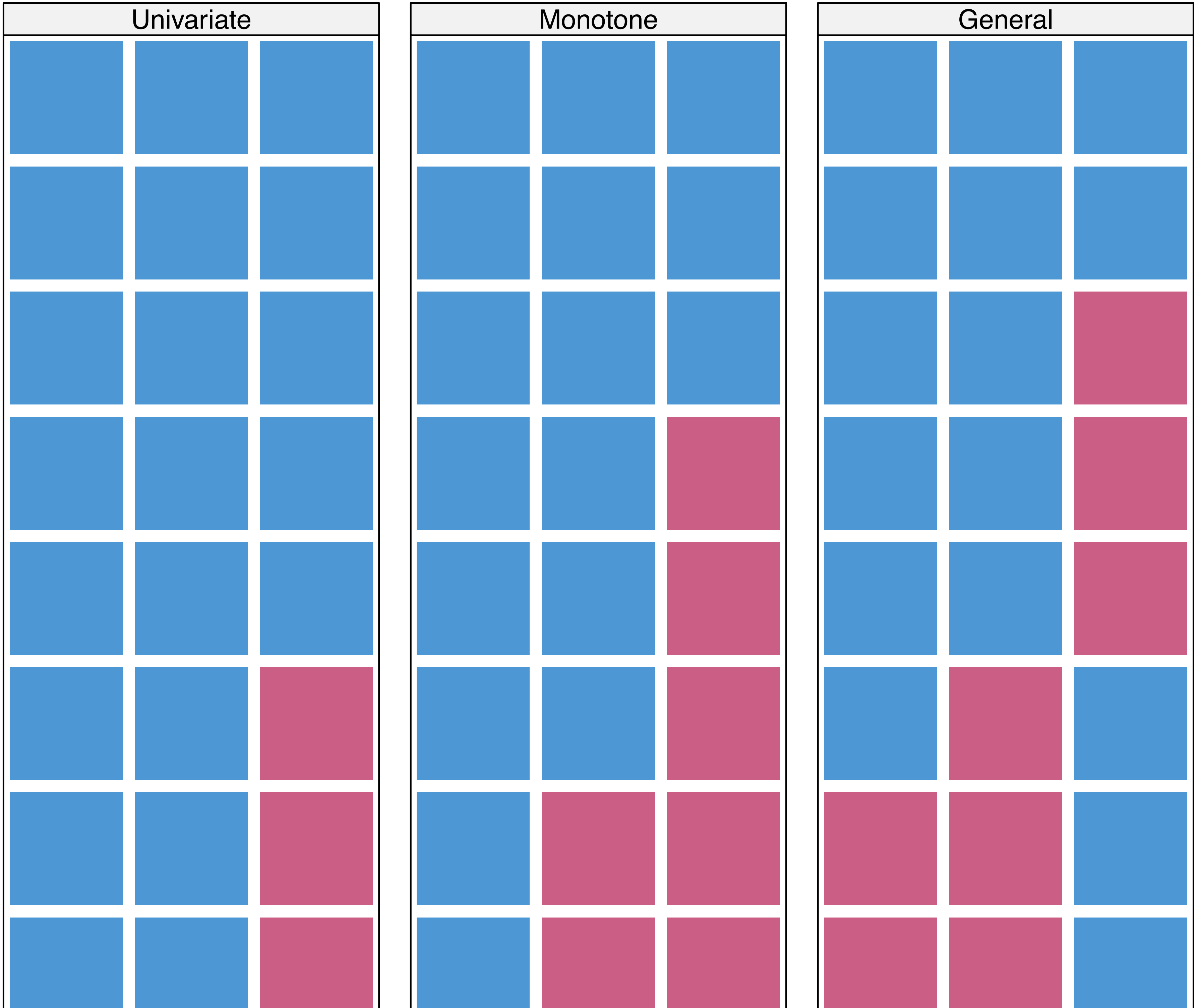
to draw $\mathring{y}_{mis}$ randomly from "the nearest neighbours"

Algorithm 3.3 in van Buuren (2018) Section 3.4.2, copied from github Rmd file:

1. Calculate $\dot\phi$ and $\hat\phi$ by Steps 1-8 of Algorithm 3.1.
2. Calculate $\dot\eta(i,j)=|X_i^{\rm obs}\hat\phi - X_j^{\rm mis}\dot\phi|$ with $i=1,\dots,n_1$ and $j=1,\dots,n_0$.
3. Construct n_0 sets Z_j, each containing d candidate donors, from $Y_{\rm obs}$ such that $\sum_d\dot\eta(i,j)$ is minimum for all $j=1,\dots,n_0$. Break ties randomly.
4. Draw one donor i_j from Z_j randomly for $j=1,\dots,n_0$.
5. Calculate imputations $\dot y_j = y_{i_j}$ for $j=1,\dots,n_0$.

This is the most popular imputation method!

Missing patterns



Fully conditional specification

(also called chained equations, sequ~~a~~ential regression multivariate imputation)
van Buuren (2018) Sections 4.5.1 and 4.5.2 and Molenberghs et al. (2014) Chapter 13.
Also this type of solution is for general missing patterns, and when missing data are MAR.

IMPUTATION MODEL

yet another new notation)

ALL DATA → $\underset{N\times p}{Y}$

$$\begin{tikzcd}
 & Y_j^{obs} & \\
Y_j \text{ column} \arrow[ur, dash] \arrow[dr, dash] & & \\
 & Y_j^{mis} \arrow[r] & \dot{Y}_j \text{ imputed}
\end{tikzcd}$$

Y_{-j} all colums except j

R indicator missingness

ϕ_j parameter in model for Y_j as a function of Y_{-j}

current "response"

all covariates in the model

MICE algorithm

The following algorithm is presented in Molenberghs et al. (2014) Figure 13.3 and van Buuren (2018) Algorithm 4.3 (Section 4.5.2), and copied from github Rmd file.

1. Specify an imputation model $P(Y_j^{\overset{\text{is}}{\text{m}}}|Y_j^{\text{obs}}, Y_{-j}, R)$ 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}.

iterations, cycles

3. Repeat for $t = 1, \dots, M$.

columns with missing values

4. Repeat for $j = 1, \dots, p$.
5. Define $\dot{Y}_{-j}^{t} = (\dot{Y}_{1}^{t}, \dots, \dot{Y}_{j-1}^{t}, \dot{Y}_{j+1}^{t-1}, \dots, \dot{Y}_{p}^{t-1})$ as the currently complete data except Y_j.
6. Draw $\dot{\phi}_j^t \sim P(\phi_j^t | Y_j^{\text{obs}}, \dot{Y}_{-j}^{t}, R)$.
7. Draw imputations $\dot{Y}_j^t \sim P(Y_j^{\text{mis}} | Y_j^{\text{obs}}, \dot{Y}_{-j}^{t}, R, \dot{\phi}_j^t)$.
8. End repeat j.
9. End repeat t.

The number of iterations t is recommended to be 5 to 10. However, convergence of the algorithm can only be seen when it has been achieved, so more iterations may be needed.

Maybe this is not very clear from the algoritm, but this is a Markov Chain Monte Carlo method, and in particular it is a Gibbs sampler (if the conditional distribution together form a joint distribution). The algorithm must then be able to converge to a stationary distribution for us to use the results. Please refer to TMA4300 Computation statistics for detail on MCMC.

In the R mice package the m multiple imputation data sets (streams) are run in parallell - that is the MICE algorithm listed above is run m times simultaneously and convergence can be monitored for each and all streams together. In convergece plots then the m streams are plotted together for each of the t iterations.

The algorithm can run into problems if the variables in the imputation model are highly correlated, when the missing rate is high and when there are constraints on the imputation model.

See slides pages 85+86 of MICE course for difference between convergence and non-convergence of the MICE algorithm.

Predictors in imputation models

- Include “all” variables to be used in the main analysis (the analysis model)
- Better with too many predictors than too few (rich model is best)
- Include the data analysis model response as covariate in the imputation models, see for example Moons KG (2006)
- Include variables that are predictors of missingness, or associated with the varible to be inputed (none of these may be part of the analysis model)
- Limit the number of predictors for stability, and many MI methods does not handle correlated predictors very well
- Nonlinear effects and interactions - should that be included? Note: passive imputation!

Model selection and assessment when using multiple imputation

Will address this to some extent in Part 2 and 3. Here are some elements to consider.

- Model selection can be done in the statistical analysis in the MI-loop, and there is a count method combined with Wald test that can be used to make a consensus model from the potentially m different MI-models. This is of cause dependent on that we have a parametric model as an analysis model. Done in case study presented in class (not in notes).
- If the analysis model is not a parametric model, maybe a tree or neural net or ensemble, what do we then do with Rubin's rules? It is possible to use them on the predictions or on the ROC-AUC. But, is that useful to do on a "training set"?