Handling Missing Data in R with MICE

Stef van Buuren$^{1,2}$

$^1$Methodology and Statistics, FSBS, Utrecht University

$^2$Netherlands Organization for Applied Scientific Research TNO, Leiden

Winnipeg, June 11, 2017
Why this course?

- Missing data are everywhere
- Ad-hoc fixes often do not 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
Course materials

https://github.com/stefvanbuuren/winnipeg

Flexible Imputation of Missing Data (FIMD)
R software and examples

- R: Install from https://cran.r-project.org
- RStudio: Install from https://www.rstudio.com
- R package mice 2.30 or higher: from CRAN or from https://github.com/stefvanbuuren/mice
- More examples: http://www.multiple-imputation.com
## Time table (morning)

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>L/P</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>09.00 - 09.15</td>
<td>L</td>
<td></td>
<td>Overview</td>
</tr>
<tr>
<td>09.15 - 10.00</td>
<td>I</td>
<td>L</td>
<td>Introduction to missing data</td>
</tr>
<tr>
<td>10.00 - 10.30</td>
<td>I</td>
<td>P</td>
<td>Ad hoc methods + MICE</td>
</tr>
<tr>
<td>10.30 - 10.45</td>
<td></td>
<td></td>
<td>PAUSE</td>
</tr>
<tr>
<td>10.45 - 11.30</td>
<td>II</td>
<td>L</td>
<td>Multiple imputation</td>
</tr>
<tr>
<td>11.30 - 12.00</td>
<td>II</td>
<td>P</td>
<td>Boys data</td>
</tr>
<tr>
<td>12.00 - 13.15</td>
<td></td>
<td></td>
<td>PAUSE</td>
</tr>
</tbody>
</table>
## Time table (afternoon)

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>L/P</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.15 - 14.00</td>
<td>III</td>
<td>L</td>
<td>Generating plausible imputations</td>
</tr>
<tr>
<td>14.00 - 14.30</td>
<td>III</td>
<td>P</td>
<td>Algorithmic convergence and pooling</td>
</tr>
<tr>
<td>14.30 - 14.45</td>
<td></td>
<td></td>
<td>PAUSE</td>
</tr>
<tr>
<td>14.45 - 15.15</td>
<td>IV</td>
<td>L</td>
<td>Imputation in practice</td>
</tr>
<tr>
<td>15.15 - 15.45</td>
<td>IV</td>
<td>P</td>
<td>Post-processing and passive imputation</td>
</tr>
<tr>
<td>15.45 - 16.00</td>
<td>V</td>
<td>L</td>
<td>Guidelines for reporting</td>
</tr>
</tbody>
</table>
SESSION I
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
- 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.
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
Disadvantages

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

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

- Mean imputation
  - Frequency of ozone levels
  - Scatter plot of ozone vs. solar radiation

```
Ozone (ppb)
0 50 100 150
0 10 20 30 40 50

Solar Radiation (lang)
0 50 100 150 250
0 50 100 150
```
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

- Ad-hoc methods
  - 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
Handling Missing Data in R with MICE > Ad-hoc methods

Stochastic regression imputation

![Histogram of Ozone (ppb) frequencies and scatter plot of Ozone (ppb) vs. Solar Radiation (lang)]
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
Single imputation methods, wrapup

- Underestimate uncertainty caused by the missing data
- Unbiased only under restrictive assumptions
Alternatives

- Maximum Likelihood, Direct Likelihood
- Weighting
- Multiple Imputation

What is multiple imputation

Rising popularity of multiple imputation

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of publications (log)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1975</td>
<td>1</td>
</tr>
<tr>
<td>1980</td>
<td>2</td>
</tr>
<tr>
<td>1985</td>
<td>5</td>
</tr>
<tr>
<td>1990</td>
<td>10</td>
</tr>
<tr>
<td>1995</td>
<td>20</td>
</tr>
<tr>
<td>2000</td>
<td>50</td>
</tr>
<tr>
<td>2005</td>
<td>100</td>
</tr>
<tr>
<td>2010</td>
<td>200</td>
</tr>
</tbody>
</table>

- **early publications**
- 'multiple imputation' in abstract
- 'multiple imputation' in title
Main steps used in multiple imputation

Incomplete data → Imputed data → Analysis results → Pooled results
Steps in mice

- **incomplete data**
- **imputed data**
- **analysis results**
- **pooled results**

**Data flow diagram**:
- `mice()` -> `mids`
- `with()` -> `mira`
- `pool()` -> `mipo`
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 $\ell$-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

$$\bar{Q} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{Q}_\ell$$ (1)
Within-imputation variance

Average of the complete-data variances as

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

where $\bar{U}_\ell$ is the variance-covariance matrix of $\hat{Q}_\ell$ obtained for the $\ell$-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
Variance between the $m$ complete-data estimates is given by

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

(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

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

(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
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

\[ \lambda = \frac{B + B/m}{T} , \]  

(5)

Relative increase in variance due to nonresponse

\[ r = \frac{B + B/m}{\bar{U}} \]  

(6)

These are related by \( r = \lambda/(1 - \lambda) \).
Fraction of information about $Q$ missing due to nonresponse

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

This measure needs an estimate of the degrees of freedom $\nu$.

Relation between $\gamma$ and $\lambda$

$$\gamma = \frac{\nu + 1}{\nu + 3} \lambda + \frac{2}{\nu + 3} \quad (8)$$

The literature often confuses $\gamma$ and $\lambda$. 
Statistical inference for $\bar{Q}$ (1)

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

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

where $t_{(\nu,1-\alpha/2)}$ is the quantile corresponding to probability $1 - \alpha/2$ of $t_\nu$.

For example, use $t(10, 0.975) = 2.23$ for the 95% confidence interval for $\nu = 10$. 

Statistical inference for $\bar{Q}$ (2)

Suppose we test the null hypothesis $Q = Q_0$ for some specified value $Q_0$. We can find the $p$-value of the test as the probability

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

where $F_{1,\nu}$ is an $F$ distribution with 1 and $\nu$ degrees of freedom.
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 $n = \infty$:

\[
\nu_{\text{old}} = (m - 1) \left(1 + \frac{1}{r^2}\right)
\]

\[
= \frac{m - 1}{\lambda^2}
\]  

(11)
Degrees of freedom (2)

The new formula is

$$\nu = \frac{\nu_{\text{old}} \nu_{\text{obs}}}{\nu_{\text{old}} + \nu_{\text{obs}}}.$$  \hspace{1cm} (12)

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

$$\nu_{\text{obs}} = \frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3 \nu_{\text{com}} (1 - \lambda)}.$$  \hspace{1cm} (13)

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

How large should $m$ be?

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

1. Use $m = 5$ or $m = 10$ if the fraction of missing information is low, $\gamma < 0.2$.

2. Develop your model with $m = 5$. Do final run with $m$ equal to percentage of incomplete cases.

3. 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.
The legacy
Introductions to multiple imputation


Relation between temperature and gas consumption

![Graph showing the relation between temperature and gas consumption. The graph displays a scatter plot with temperature on the x-axis (0 to 10 °C) and gas consumption (cubic feet) on the y-axis (2 to 7 cubic feet). The data points suggest a negative correlation, indicating that as temperature increases, gas consumption decreases.]
We delete gas consumption of observation 47.
Predict imputed value from regression line

Temperature (°C)

Gas consumption (cubic feet)
Handling Missing Data in R with MICE

III

Creating imputations, univariate

Predicted value + noise
Handling Missing Data in R with MICE > III > Creating imputations, univariate

Predicted value + noise + parameter uncertainty

Gas consumption (cubic feet) vs. Temperature (°C)
Imputation based on two predictors

Temperature (°C) vs. Gas consumption (cubic feet)

- + before insulation
- o after insulation
Handling Missing Data in R with MICE

III

Creating imputations, univariate

Predictive mean matching: $Y$ given $X$

Temperature (°C)

Gas consumption (cubic feet)

+ before insulation
○ after insulation
Add two regression lines

- Temperature (°C)
- Gas consumption (cubic feet)

- + before insulation
- ○ after insulation

Graph showing the relationship between temperature and gas consumption before and after insulation.
Predicted given 5° C, ‘after insulation’
Define a matching range $\hat{y} \pm \delta$
Select potential donors

Select potential donors

Gas consumption (cubic feet) vs. Temperature (°C)

- before insulation
- after insulation
Bayesian PMM: Draw a line

- Gas consumption (cubic feet)

- Temperature (°C)

+ before insulation

○ after insulation
Define a matching range $\hat{y} \pm \delta$
Select potential donors

Temperature (°C)

Gas consumption (cubic feet)

+ before insulation
○ after insulation

before insulation
after insulation

−2 0 2 4 6 8 10

Temperature (°C)
Imputation of a binary variable

- **logistic regression**

\[
\Pr(y_i = 1|X_i, \beta) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}.
\] (14)
Fit logistic model

-3 -2 -1 0 1 2 3

0.0 0.2 0.4 0.6 0.8 1.0

Probability

Linear predictor
Handling Missing Data in R with MICE > III > Creating imputations, univariate

Draw parameter estimate

![Graph showing the relationship between linear predictor and probability](image-url)
Read off the probability
Impute ordered categorical variable

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

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

Probability vs. Linear predictor chart.
Read off the probability

![Graph showing probability vs. linear predictor](image)

- **1**: Probability at linear predictor of 1
- **2**: Probability at linear predictor of 0
- **3**: Linear predictor at probability of 0.5
Other types of variables

- Count data
- Semi-continuous data
- Censored data
- Truncated data
- Rounded data
### Univariate imputation in *mice*

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Scale type</th>
</tr>
</thead>
<tbody>
<tr>
<td>pmf</td>
<td>Predictive mean matching</td>
<td>numeric*</td>
</tr>
<tr>
<td>norm</td>
<td>Bayesian linear regression</td>
<td>numeric</td>
</tr>
<tr>
<td>norm.nob</td>
<td>Linear regression, non-Bayesian</td>
<td>numeric</td>
</tr>
<tr>
<td>norm.boot</td>
<td>Linear regression with bootstrap</td>
<td>numeric</td>
</tr>
<tr>
<td>mean</td>
<td>Unconditional mean imputation</td>
<td>numeric</td>
</tr>
<tr>
<td>2L.norm</td>
<td>Two-level linear model</td>
<td>numeric</td>
</tr>
<tr>
<td>logreg</td>
<td>Logistic regression</td>
<td>factor, 2 levels*</td>
</tr>
<tr>
<td>logreg.boot</td>
<td>Logistic regression with bootstrap</td>
<td>factor, 2 levels</td>
</tr>
<tr>
<td>polyreg</td>
<td>Multinomial logit model</td>
<td>factor, &gt; 2 levels*</td>
</tr>
<tr>
<td>polr</td>
<td>Ordered logit model</td>
<td>ordered, &gt; 2 levels*</td>
</tr>
<tr>
<td>lda</td>
<td>Linear discriminant analysis</td>
<td>factor</td>
</tr>
<tr>
<td>sample</td>
<td>Simple random sample</td>
<td>any</td>
</tr>
</tbody>
</table>
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
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 $P(Y_{\text{mis}} | Y_{\text{obs}}, X, R)$
3. Use MCMC techniques to draw imputations $\hat{Y}_{\text{mis}}$
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
- Works very well for parameters close to the center
- Many applications
Joint Modeling: Con’s

- Lack of flexibility
- May lead to large models
- Can assume more than the complete data problem
- Can impute impossible data
Fully Conditional Specification (FCS)

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

- **mean hgt**: 80 85 90 95 100 110
- **sd hgt**: 28 30 34 36 38
- **mean wgt**: 26 27 28 29
- **sd wgt**: 37 38 39 40
- **mean bmi**: 0 50 100 150
- **sd bmi**: 5 10 15 20

Iteration: 5 10 15 20
Convergence

<table>
<thead>
<tr>
<th>mean</th>
<th>hgt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sd</th>
<th>hgt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>mean</th>
<th>wgt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sd</th>
<th>wgt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>mean</th>
<th>bmi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sd</th>
<th>bmi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SESSION IV
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
Which predictors?

1. Include all variables that appear in the complete-data model.
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
How to impute a ratio?

weight/height ratio: \( \text{whr} = \frac{\text{wgt}}{\text{hgt}} \text{ kg/m} \).

Easy if only one of wgt or hgt or whr is missing

Methods

- **POST**: Impute wgt and hgt, and calculate whr after imputation
- **JAV**: Impute whr as ‘just another variable’
- **PASSIVE1**: Impute wgt and hgt, and calculate whr during imputation
- **PASSIVE2**: As PASSIVE1 with adapted predictor matrix
> imp1 <- mice(boys)
> long <- complete(imp1, "long", inc = TRUE)
> long$whr <- with(long, wgt/(hgt/100))
> imp2 <- long2mids(long)
Method JAV: Just another variable

```r
> boys$whr <- boys$wgt/(boys$hgt/100)
> imp.jav <- mice(boys, m = 1, seed = 32093, maxit = 10)
```
### Method JAV

![Graph of Weight/Height vs Height (cm)](image)

<table>
<thead>
<tr>
<th>JAV</th>
<th>passive</th>
<th>passive 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>100</td>
<td>150</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>250</td>
</tr>
<tr>
<td>150</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>200</td>
<td>300</td>
<td>400</td>
</tr>
</tbody>
</table>

Derived variables:
Method PASSIVE

```r
> meth["whr"] <- "~I(wgt/(hgt/100))"
```
### Method PASSIVE, predictor matrix

<table>
<thead>
<tr>
<th>age</th>
<th>hgt</th>
<th>wgt</th>
<th>bmi</th>
<th>hc</th>
<th>gen</th>
<th>phb</th>
<th>tv</th>
<th>reg</th>
<th>whr</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hgt</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>wgt</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>bmi</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>hc</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gen</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>phb</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>tv</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>reg</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>whr</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Method PASSIVE

![Graph showing the relationship between Height (cm) and Weight/Height (kg/m) for JAV, passive, and passive 2 methods.](image-url)
Method PASSIVE2

> pred[c("wgt", "hgt", "hc", "reg"), "bmi"] <- 0
> pred[c("gen", "phb", "tv"), c("hgt", "wgt", "hc")]<- 0
> pred[, "whr"] <- 0
Method PASSIVE2, predictor matrix

<table>
<thead>
<tr>
<th>age</th>
<th>hgt</th>
<th>wgt</th>
<th>bmi</th>
<th>hc</th>
<th>gen</th>
<th>phb</th>
<th>tv</th>
<th>reg</th>
<th>whr</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hgt</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>wgt</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>bmi</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>hc</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gen</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>phb</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>tv</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>reg</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>whr</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Method PASSIVE2

<table>
<thead>
<tr>
<th>JAV</th>
<th>Passive</th>
<th>Passive 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weight/Height (kg/m)</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Derived variables: summary

- Derived variables pose special challenges
- Plausible values respect data dependencies
- If you can, create derived variables after imputation
- If you cannot, use passive imputation
- Break up direct feedback loops using the predictor matrix
Since `mice` 2.5, plots for imputed data:

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