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

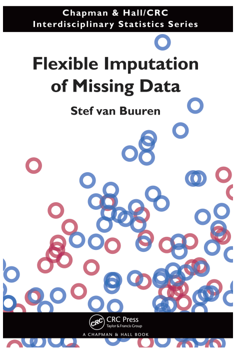
Course materials

- `https://github.com/stefvanbuuren/winnipeg`

Reading materials

1. 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. `https://www.jstatsoft.org/article/view/v045i03`
2. Van Buuren, S. (2012). Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton, FL. Chapters 1–6, 10. `http://www.crcpress.com/product/isbn/9781439868249`

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)

Time	Session	L/P	Description
09.00 - 09.15		L	Overview
09.15 - 10.00	I	L	Introduction to missing data
10.00 - 10.30	I	P	Ad hoc methods + MICE
10.30 - 10.45			PAUSE
10.45 - 11.30	II	L	Multiple imputation
11.30 - 12.00	II	P	Boys data
12.00 - 13.15			PAUSE

Time table (afternoon)

Time	Session	L/P	Description
13.15 - 14.00	III	L	Generating plausible imputations
14.00 - 14.30	III	P	Algorithmic convergence and pooling
14.30 - 14.45			PAUSE
14.45 - 15.15	IV	L	Imputation in practice
15.15 - 15.45	IV	P	Post-processing and passive imputation
15.45 - 16.00	V	L	Guidelines for reporting

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

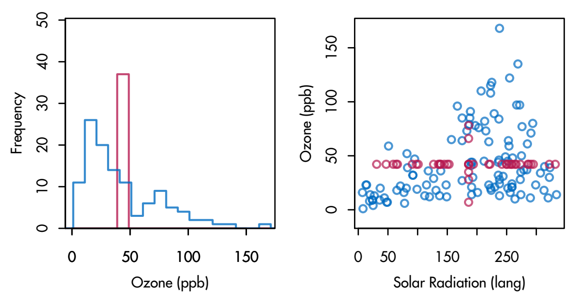
Listwise deletion

- 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



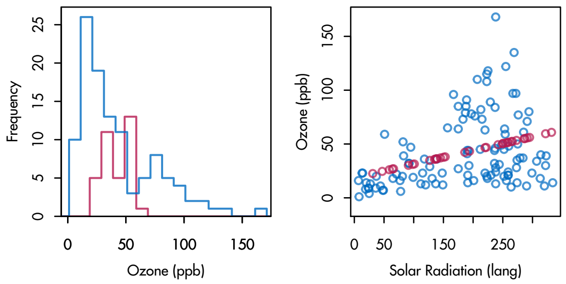
Mean imputation

- 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



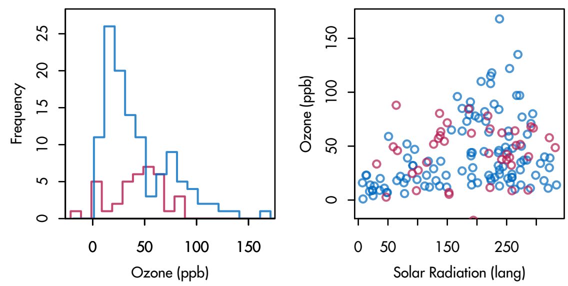
Regression imputation

- 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

Single imputation methods, wrapup

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

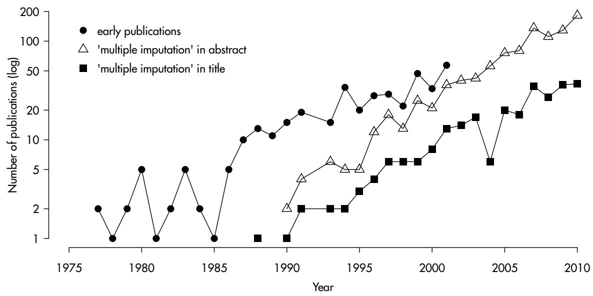
Alternatives

- Maximum Likelihood, Direct Likelihood
- Weighting
- Multiple Imputation

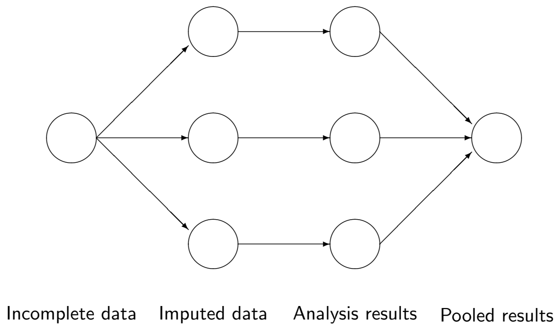
- Little, R.J.A. Rubin D.B. (2002) Statistical Analysis with Missing Data. Second Edition. John Wiley Sons, New York.

SESSION II

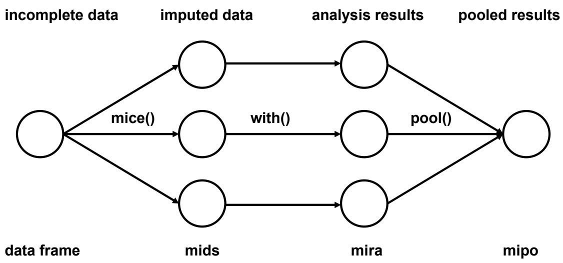
Rising popularity of multiple imputation



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

$$\bar{Q} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{Q}_\ell \qquad (1)$$

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.

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

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

$$\bar{Q} \pm t_{(\nu,1-\alpha/2)}\sqrt{T}, \tag{9}$$

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

$$P_s = \Pr\left[F_{1,\nu} > \frac{(Q_0 - \bar{Q})^2}{T}\right] \tag{10}$$

where $F_{1,\nu}$ is an F distribution with 1 and ν 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 $n = \infty$:

$$\begin{aligned} \nu_{\text{old}} &= (m-1)\left(1+\frac{1}{r^2}\right) \\ &= \frac{m-1}{\lambda^2} \end{aligned} \qquad (11)$$

Degrees of freedom (2)

The new formula is

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

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

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

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.

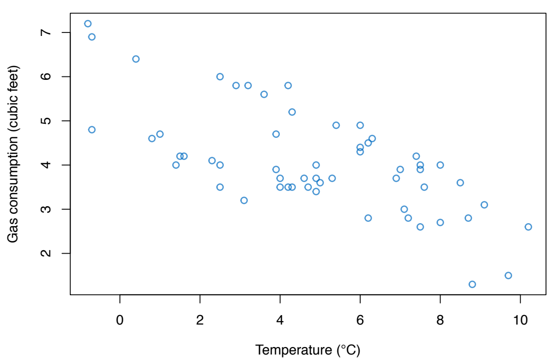
The legacy

Introductions to multiple imputation

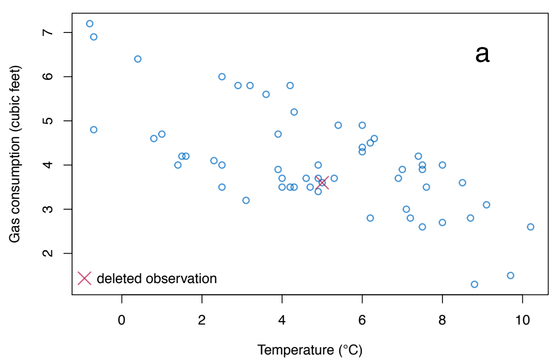
1. Schafer, J.L. (1999). Multiple imputation: A primer. Statistical Methods in Medical Research, 8(1), 3–15.
2. Sterne et al (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338, b2393.
3. Van Buuren, S. (2012). Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton, FL.

SESSION III

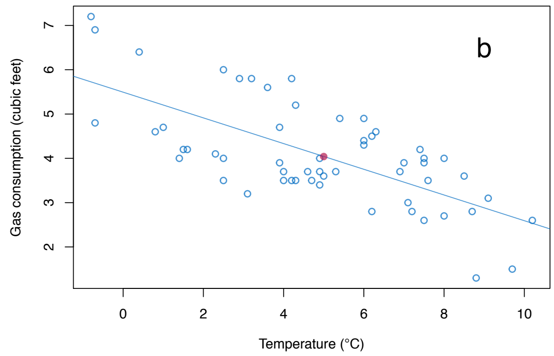
Relation between temperature and gas consumption



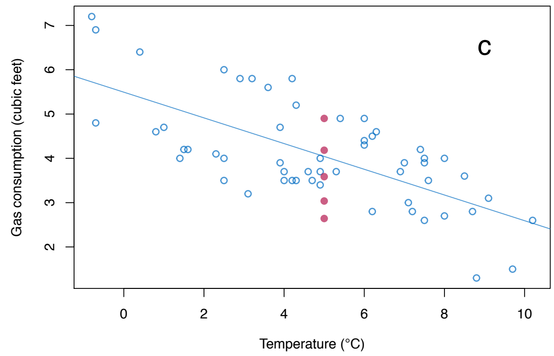
We delete gas consumption of observation 47



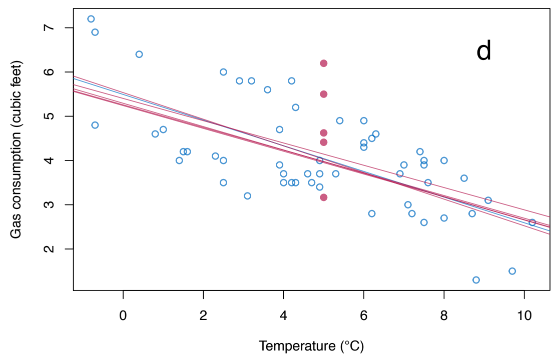
Predict imputed value from regression line



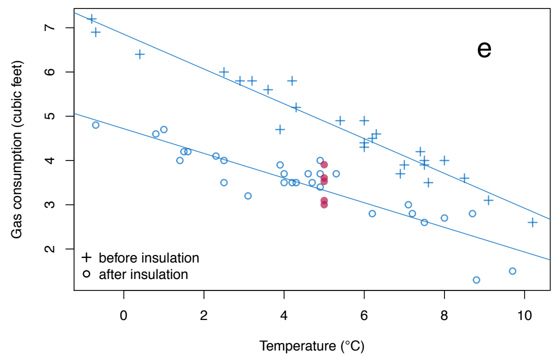
Predicted value + noise



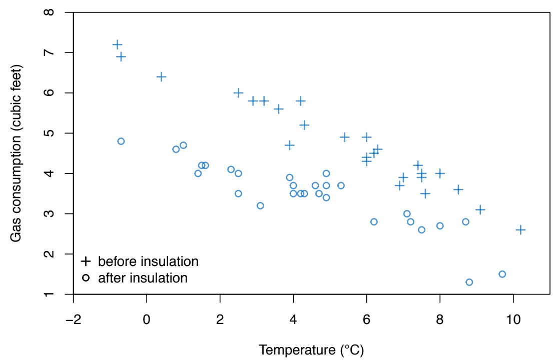
Predicted value + noise + parameter uncertainty



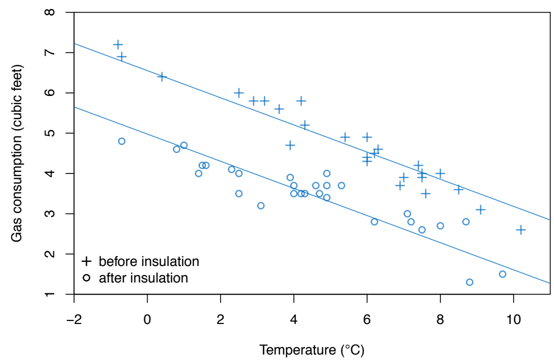
Imputation based on two predictors



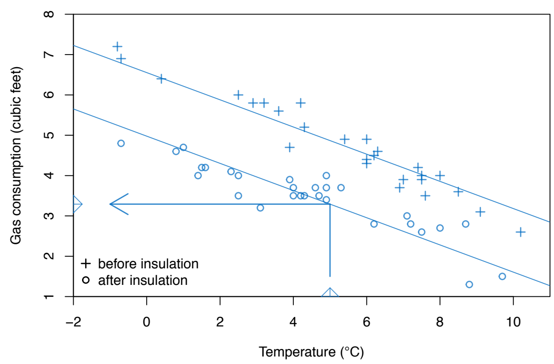
Predictive mean matching: Y given X



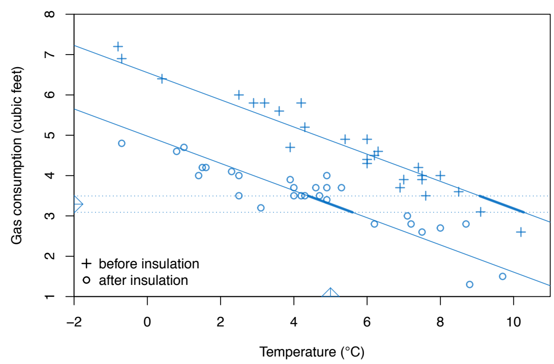
Add two regression lines



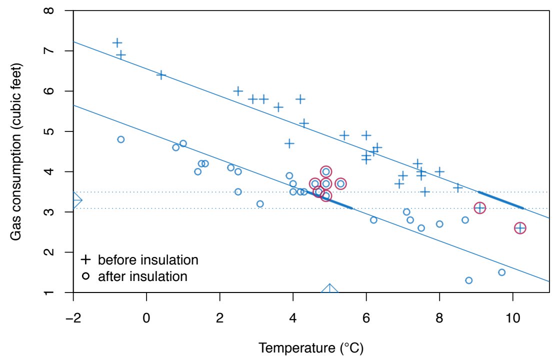
Predicted given 5° C, 'after insulation'



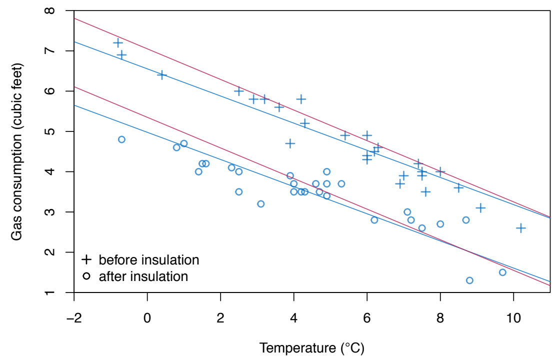
Define a matching range $\hat{y} \pm \delta$



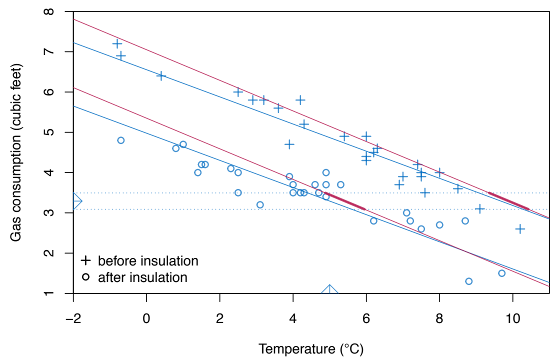
Select potential donors



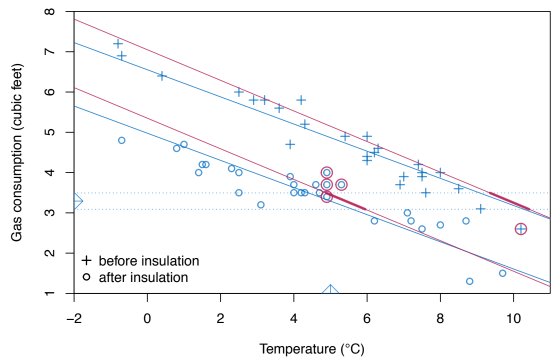
Bayesian PMM: Draw a line



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



Select potential donors

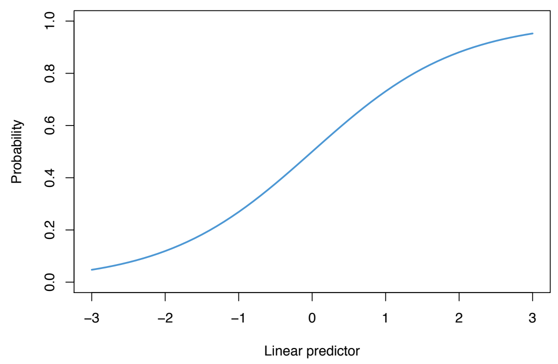


Imputation of a binary variable

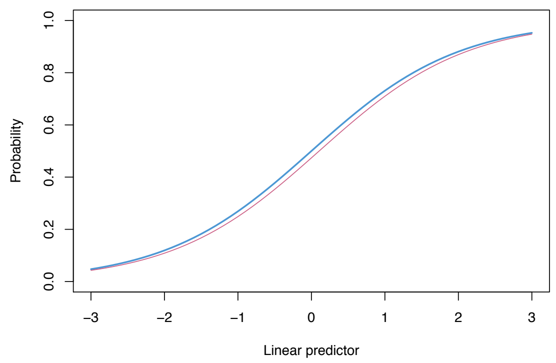
- *logistic regression*

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

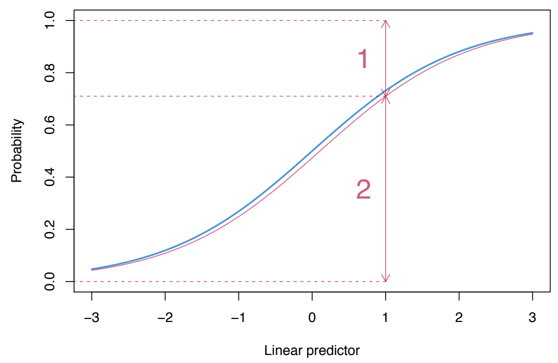
Fit logistic model



Draw parameter estimate



Read off the probability

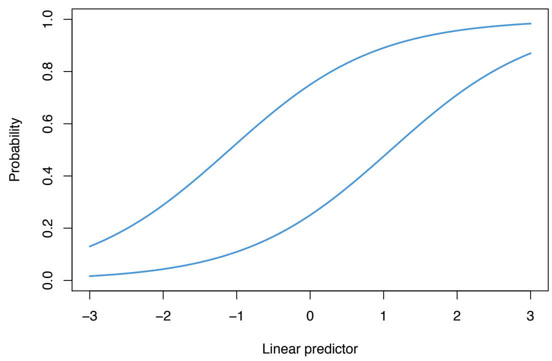


Impute ordered categorical variable

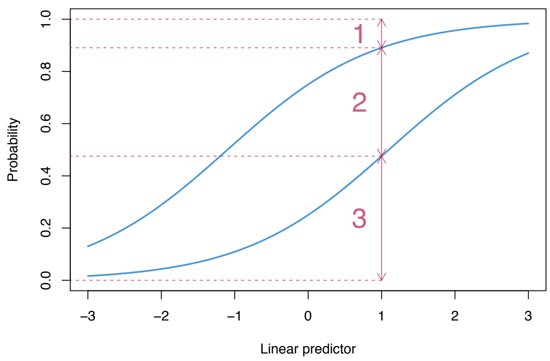
- K ordered categories $k = 1, \ldots, K$
- *ordered logit model*, or
- *proportional odds model*

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

Fit ordered logit model



Read off the probability



Other types of variables

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

Univariate imputation in `mice`

Method	Description	Scale type
pmm	Predictive mean matching	numeric*
norm	Bayesian linear regression	numeric
norm.nob	Linear regression, non-Bayesian	numeric
norm.boot	Linear regression with bootstrap	numeric
mean	Unconditional mean imputation	numeric
2L.norm	Two-level linear model	numeric
logreg	Logistic regression	factor, 2 levels*
logreg.boot	Logistic regression with bootstrap	factor, 2 levels
polyreg	Multinomial logit model	factor, > 2 levels*
polr	Ordered logit model	ordered, > 2 levels*
lda	Linear discriminant analysis	factor
sample	Simple random sample	any

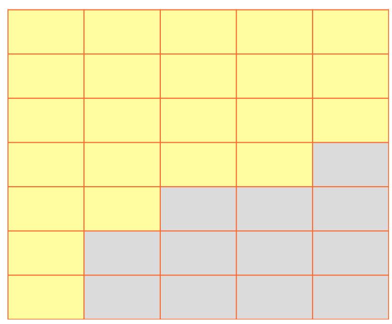
Problems in multivariate imputation

- 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

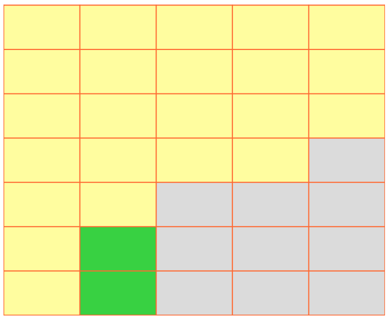
Three general strategies

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

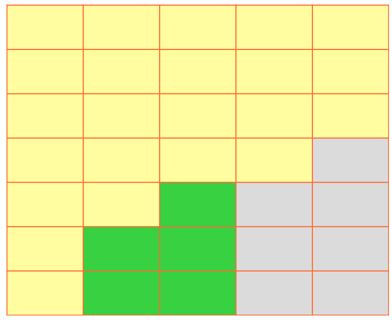
Imputation of monotone pattern



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 $\dot{Y}_{\text{mis}}$

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

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

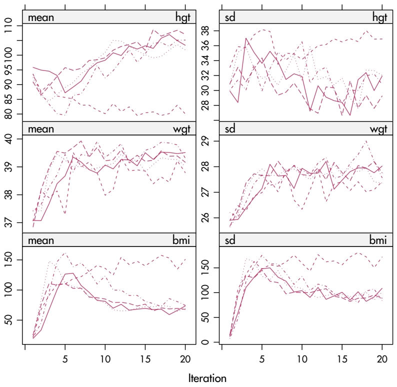
Fully Conditional Specification (FCS): Software

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`

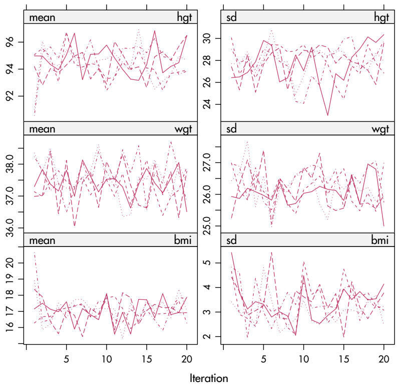
How many iterations?

- 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



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: whr=wgt/hgt 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

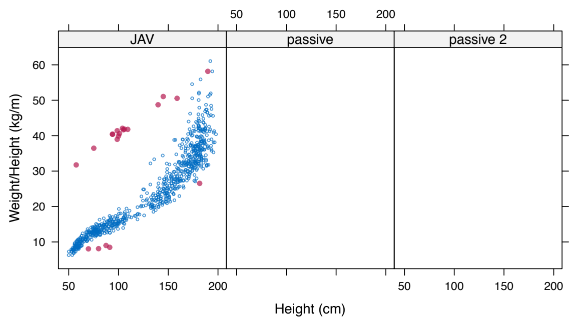
Method POST

```
> imp1 <- mice(boys)
> long <- complete(imp1, "long", inc = TRUE)
> long$whr <- with(long, wgt/(hgt/100))
> imp2 <- long2mids(long)
```


Method JAV: Just another variable

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


Method JAV



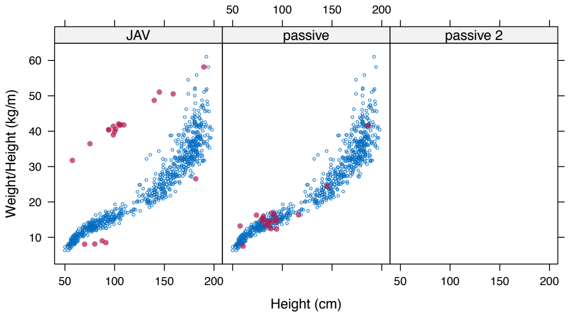
Method PASSIVE

```
> meth["whr"] <- "~I(wgt/(hgt/100))"
```


Method PASSIVE, predictor matrix

```
    age hgt wgt bmi hc gen phb tv reg whr
age   0   0   0   0  0   0   0  0   0   0
hgt   1   0   1   0  1   1   1  1   1   0
wgt   1   1   0   0  1   1   1  1   1   0
bmi   1   1   1   0  1   1   1  1   1   0
hc    1   1   1   1  0   1   1  1   1   1
gen   1   1   1   1  1   0   1  1   1   1
phb   1   1   1   1  1   1   0  1   1   1
tv    1   1   1   1  1   1   1  0   1   1
reg   1   1   1   1  1   1   1  1   0   1
whr   1   1   1   0  1   1   1  1   1   0
```

Method PASSIVE



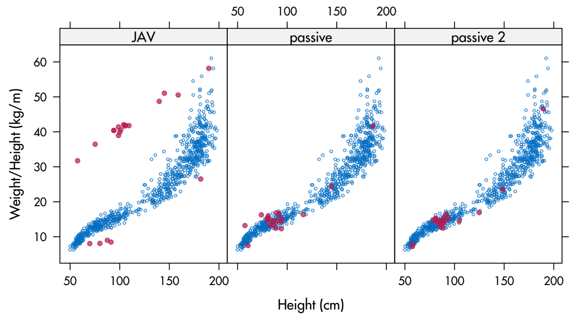
Method PASSIVE2

```
> 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

	age	hgt	wgt	bmi	hc	gen	phb	tv	reg	whr
age	0	0	0	0	0	0	0	0	0	0
hgt	1	0	1	0	1	1	1	1	1	0
wgt	1	1	0	0	1	1	1	1	1	0
bmi	1	1	1	0	1	1	1	1	1	0
hc	1	1	1	0	0	1	1	1	1	0
gen	1	0	0	1	0	0	1	1	1	0
phb	1	0	0	1	0	1	0	1	1	0
tv	1	0	0	1	0	1	1	0	1	0
reg	1	1	1	0	1	1	1	1	0	0
whr	1	1	1	1	1	1	1	1	1	0

Method PASSIVE2



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

Standard diagnostic plots in `mice`

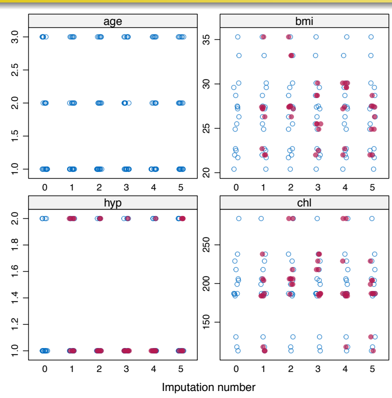
Since `mice` 2.5, plots for imputed data:

- 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))
```

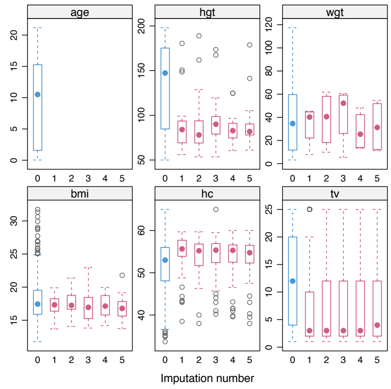

`stripplot(imp, pch=c(1,19))`



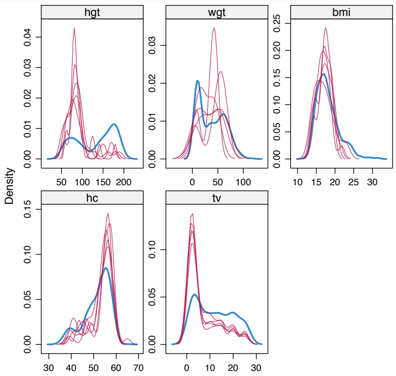
A larger data set

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


bwplot(imp)



densityplot(imp)



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