Dark data, strategies & solutions

Stef van Buuren, Gerko Vink

Mar 25, 2022

Lecture topics

- Problem of dark data
- Strategies to deal with missing data
- Multiple imputation methodology to analyse incomplete data
- Synthetic data sets for disclosure protection

Why this course?

- Real data are always incomplete
- Ad-hoc fixes do not (always) work
- Multiple imputation as principled and broadly applicable approach
- Goal: get comfortable with a powerful way to deal with incomplete data
- We use the `mice` package in `R`

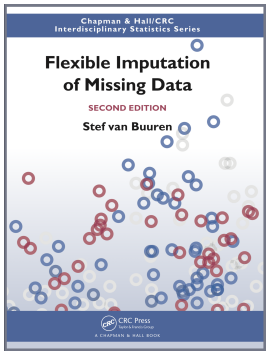
Today's schedule

Time	Activity
09.30 - 10.15	Lecture 1
10.15 - 10.30	COFFEE
10.30 - 11.15	Lecture 2
11.15 - 11.30	COFFEE
11.30 - 12.00	Lecture 3

Stef van Buuren covers the lectures. Gerko Vink covers practicals.

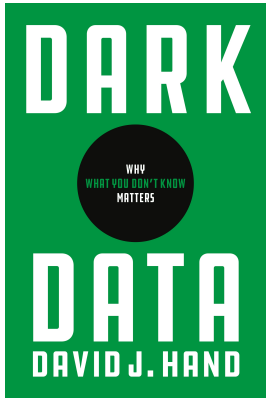
Reading materials

- 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
- Van Buuren, S. (2018). Flexible Imputation of Missing Data. Second Edition. Chapman & Hall/CRC, Boca Raton, FL. Free text: https://stefvanbuuren.name/fimd Order book: https://www.crcpress.com/Flexible-Imputation-of-Missing-Data-Second-Edition/Buuren/p/book/9781138588318



Introduction to dark data

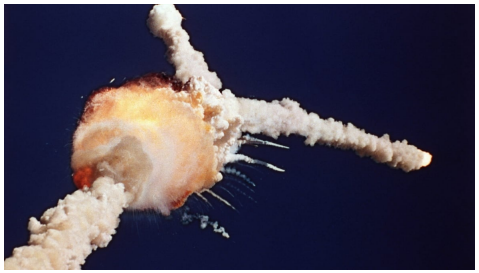
Introduction to dark data



What is dark data?

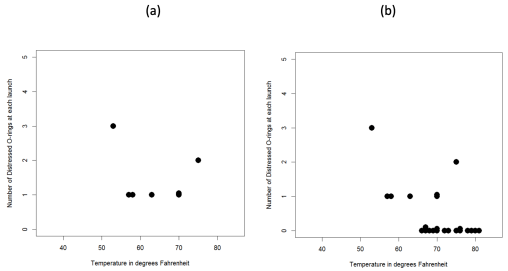
Dark data are concealed from us, and that very fact means we are at risk of misunderstanding, of drawing incorrect conclusions, and of making poor decisions.

Challenger space shuttle - 28 Jan 1986 - 7 deaths



Challenger space shuttle - 28 Jan 1986 - 7 deaths

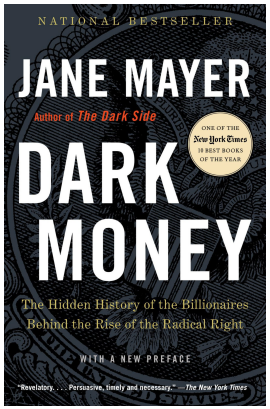
Figure 1.1 (a) Data examined in the pre-launch teleconference; **(b)** Complete data.



Dark data types (selection)

- DD-Type 1: Data We Know Are Missing
- DD-Type 2: Data We Don't Know are Missing
- DD-Type 4: Self-Selection
- DD-Type 12: Information Asymmetry
- DD-Type 13: Intentionally Darkened Data
- DD-Type 14: Fabricated and Synthetic Data

There is also "Dark money"



Definition of missing values

- Missing values are those values that are not observed
- Values do exist in theory, but we are unable to see them

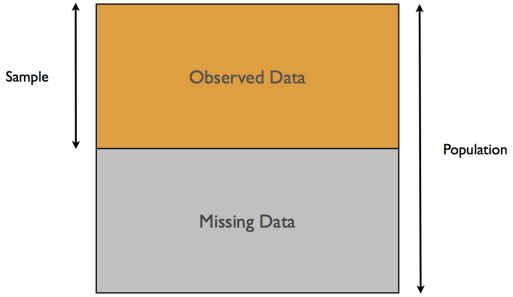
Evolving views on missing data

- "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)

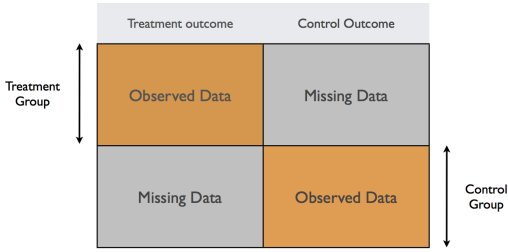
Why are missing data interesting?

- MISSING DATA ARE THE HEART OF STATISTICS
- Taking a sample
- Estimating a causal effect
- Predicting future outcome
- Combining data from different sources

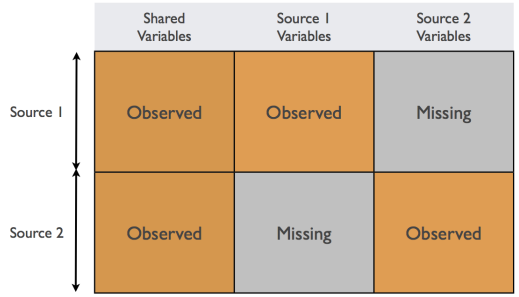
Sampling example



Experiment example



Matching example



Reasons

Missing data can occur for a lot of reasons. For example

- death, dropout, refusal, concealed
- routing, experimental design
- join, merge, bind
- too far away, too small to observe
- power failure, budget exhausted, bad luck

Why are missing values problematic?

- Cannot calculate, not even the mean
- Less information than planned
- Enough statistical power?
- Different analyses, different n's
- Systematic biases in the analysis
- Appropriate confidence interval, P-values?

Missing data can severely complicate interpretation and analysis

Some confusing terminology

Complete data = Observed data + Unobserved data

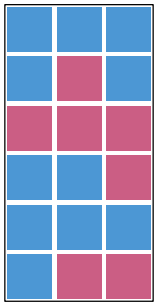
Incomplete data = Observed data

Missing data = Unobserved data

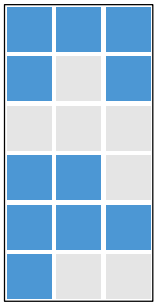
Complete cases = subset of rows in the observed data without missing values

Complete variables = subset of columns in the observed data without missing values

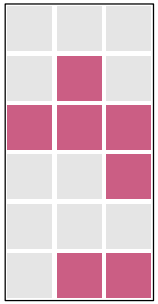
Complete data



Incomplete data = observed data



Missing data = unobserved data



Notation: Y, R, X

- Y random variable with missing data
- Y^{obs} true and observed values of Y
- Y^{mis} true but unobserved values of Y, missing values
- R response indicator
- $R = 1$ if Y is observed
- $R = 0$ if Y is missing
- X complete covariate

Missing data mechanism

- Process that governs which Y's are observed and which Y's are unobserved (Rubin, 1976)
- Sometimes we know this process (e.g.~experimental design, sampling)
- Alternatively, model by response probability $P(R|Y^{\text{obs}}, Y^{\text{mis}}, X)$
- Also called **missing data model**

MCAR: Missing Completely at Random

- Probability to be missing is not related to any data

$$P(R|Y^{\text{obs}}, Y^{\text{mis}}, X, \psi) = P(R|\psi)$$

- Examples
  - data transmission error
  - random sample

MAR: Missing at Random

- Probability to be missing depends on known data

$$P(R|Y^{\text{obs}}, Y^{\text{mis}}, X, \psi) = P(R|Y^{\text{obs}}, X, \psi)$$

- Examples
  - Income, where we have X related to wealth
  - Branch patterns (e.g. how old are your children?)

MNAR: Missing Not at Random

- Probability to be missing depends on unknown data

$P(R|Y^{\text{obs}}, Y^{\text{mis}}, X, \psi)$ does not simplify

- Examples
  - Income, without covariates related to income
  - Body weight report

Strategies to deal with missing data

Strategies to deal with missing data

- Prevention
- Ad-hoc methods, e.g., single imputation, complete cases
- Weighting methods
- Likelihood methods, EM-algorithm
- Multiple imputation

Prevention

- Design: Time intervals, Number of variables, Pilot study
- Collection: Incentives, Match interviewer-respondent, Quick follow-up, Retrieve missing data
- Measures: Use short forms, Minimize intrusive measures, Clarity, Layout
- Treatment: Minimize burden and intensity

Listwise deletion, complete-case analysis

- Analyze only the complete records
- Advantages
  - Simple (default in most software)
  - Unbiased under MCAR
  - Conservative standard errors, significance levels
  - Two special properties in regression

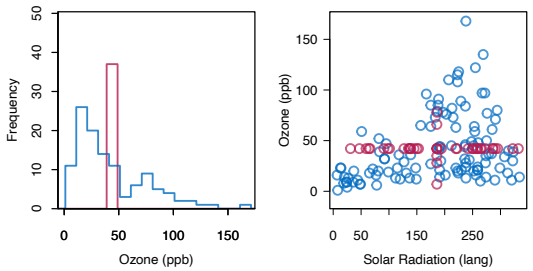
Listwise deletion, complete-case analysis

- Disadvantages
  - Wasteful
  - May not be possible
  - Larger 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



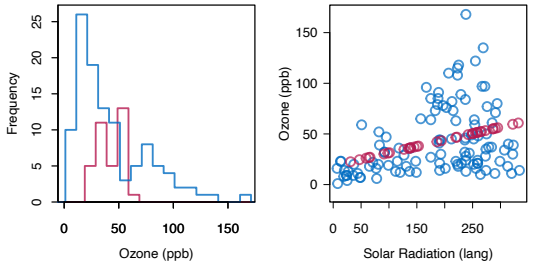
Mean imputation

- 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
  - Under MAR, unbiased estimates of regression coefficients
  - Good approximation to the (unknown) true data if explained variance is high
- Favourite among data scientists and machine learners

Regression imputation



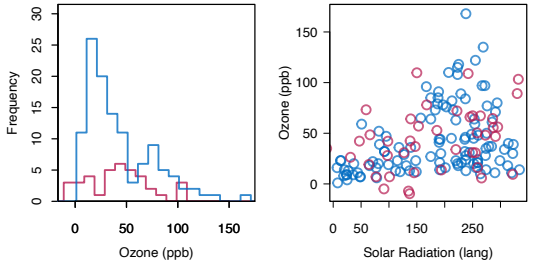
Regression imputation

- 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

Overview of assumptions needed

		Unbiased		Standard Error
	Mean	Reg Weight	Correlation	
Listwise	MCAR	MCAR	MCAR	Too large
Pairwise	MCAR	MCAR	MCAR	Complicated
Mean	MCAR	–	–	Too small
Regression	MAR	MAR	–	Too small
Stochastic	MAR	MAR	MAR	Too small
LOCF	–	–	–	Too small
Indicator	–	–	–	Too small

Multiple imputation

- Imputes each missing value m times
- Variation between the m imputed values reflects our ignorance about the true value

Acceptance of multiple imputation

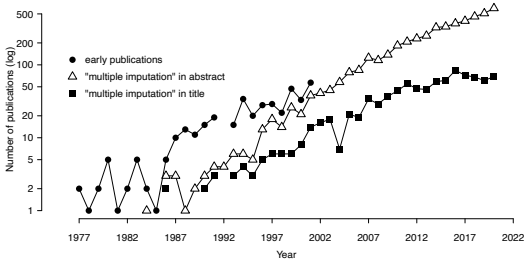
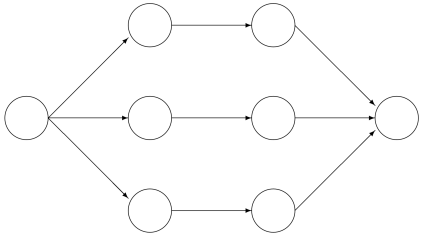


Figure 1: Source: Scopus (May 27, 2021)

Multiple imputation



Multiple imputation

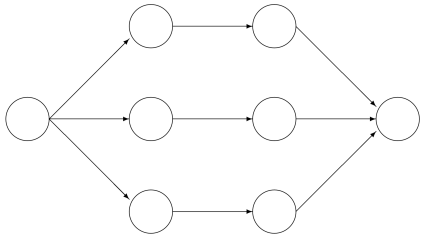
- Advantages
  - Correct point and variance estimates
  - Splits missing data problem from complete-data analysis
  - Theoretical properties well established
  - Flexible, widely applicable
  - Extensible to MNAR
- Disadvantages
  - Need to create and work with multiple imputed data sets
  - May not always be most efficient

Multiple imputation topics

- General idea of multiple imputation
- Statistical inference on multiply-imputed data
- Creating univariate imputations
- Creating multivariate imputations, MICE algorithm

General idea of multiple imputation

Multiple imputation



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, represented as a $k \times 1$ column vector

Pooled estimate $\bar{Q}$ is simply the average

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

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 ℓ-th imputation

$\bar{U}_\ell$ is the variance is the estimate, *not* the variance in the data

Within-imputation variance is large if the sample is small

Between-imputation variance

Variance between the m complete-data estimates is given by

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

where $\bar{Q}$ is the pooled estimate.

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

$$\begin{aligned} T &= \bar{U} + B + B/m \\ &= \bar{U} + \left(1 + \frac{1}{m}\right) B \end{aligned} \tag{1}$$

for the total variance of $\bar{Q}_m$, 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 $\bar{Q}_m$ itself is based on finite m.

Variance ratio's (1)

Proportion of the variation attributable to the missing data

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

Relative increase in variance due to nonresponse

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

These are related by $r = \lambda/(1 - \lambda)$.

Variance ratio's (2)

Fraction of information about Q missing due to nonresponse

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

This measure needs an estimate of the degrees of freedom ν (c.f. section 2.3.6)

Relation between γ and λ

$$\gamma = \frac{\nu + 1}{\nu + 3} \lambda + \frac{2}{\nu + 3}.$$

The literature often confuses γ and λ.

Statistical inference on multiply-imputed data

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_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

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

where $F_{1,\nu}$ is an F distribution with 1 and ν degrees of freedom.

How large should m be?

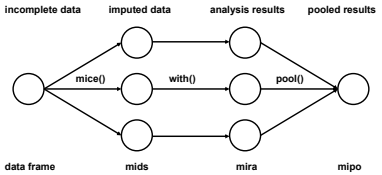
Classic advice: $m = 3, 5, 10$. More recently: set m higher: 20–100.

Some advice:

- Use $m = 5$ or $m = 10$ if the fraction of missing information is low, $\gamma < 0.2$.
- Develop your model with $m = 5$. Do final run with m equal to percentage of incomplete cases.

Multiple imputation in `mice`

Generic workflow in `mice`



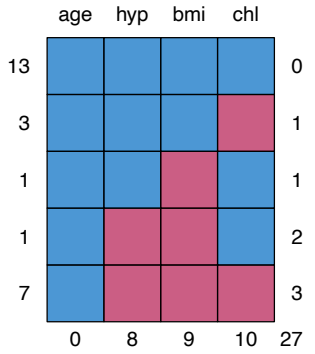
Inspect the data

```
library("mice")
head(nhanes)
```

```
##   age  bmi hyp chl
## 1   1   NA  NA  NA
## 2   2 22.7   1 187
## 3   1   NA   1 187
## 4   3   NA  NA  NA
## 5   1 20.4   1 113
## 6   3   NA  NA 184
```

Inspect missing data pattern

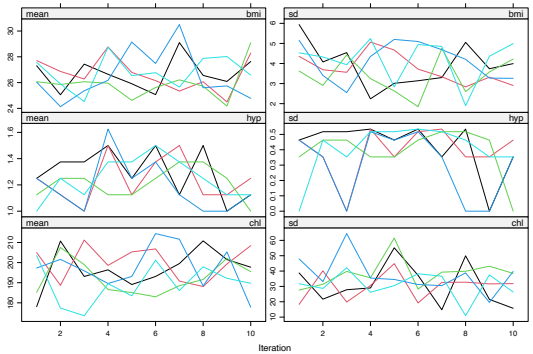
```
md.pattern(nhanes)
```



Multiply impute the data

```
imp <- mice(nhanes, print = FALSE, maxit=10, seed = 24415)
```

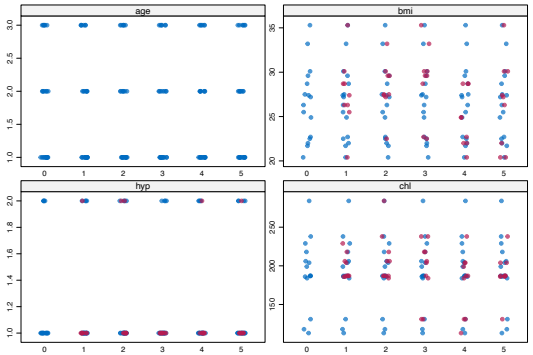
Inspect the trace lines for convergence



Stripplot of observed and imputed data

```
stripplot(imp, pch = 20, cex = 1.2)
```

Stripplot of observed and imputed data



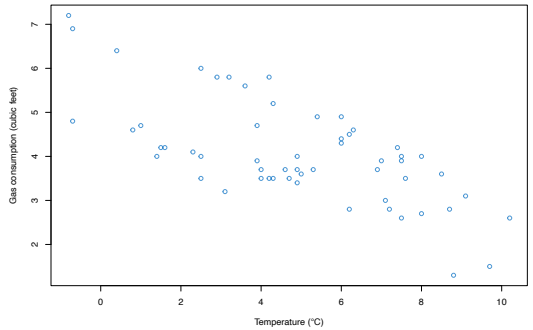
Fit the complete-data model

```
fit <- with(imp, lm(bmi ~ age))
est <- pool(fit)
summary(est)
```

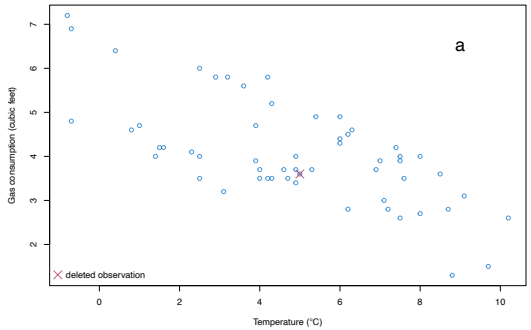
```
##          term estimate std.error statistic   df  p.value
## 1 (Intercept)     30.5      2.45     12.46  7.2 3.94e-06
## 2         age     -2.1      1.12     -1.87 10.8 8.89e-02
```

Creating univariate imputations

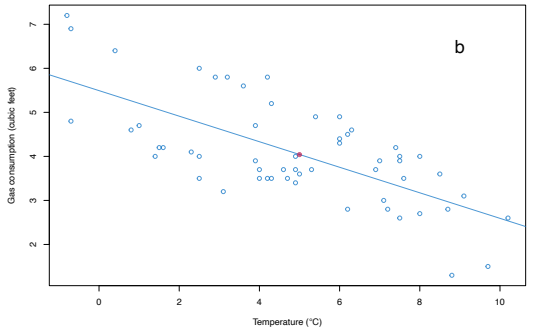
Relation between temperature and gas consumption



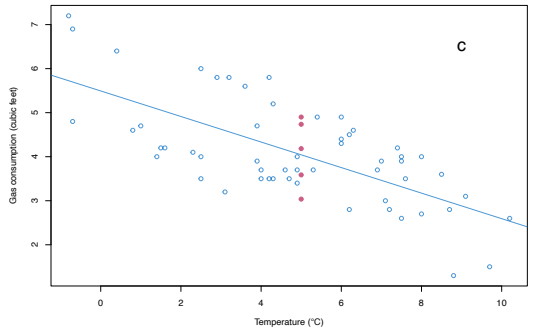
We delete gas consumption of observation 47



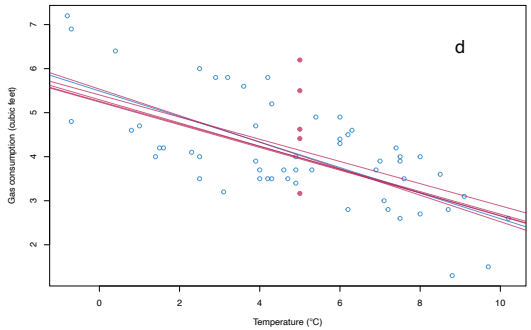
Predict imputed value from regression line



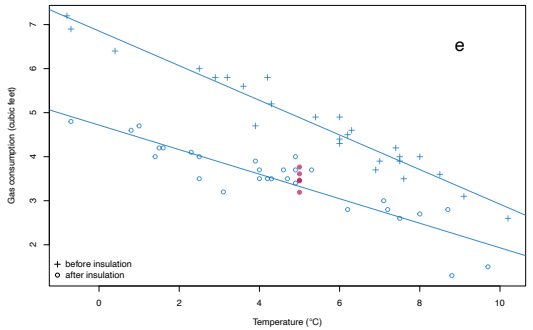
Predicted value + noise



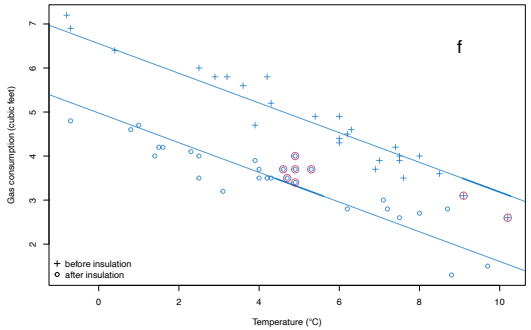
Predicted value + noise + parameter uncertainty



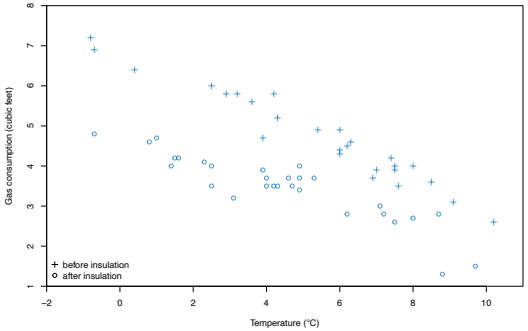
Imputation based on two predictors



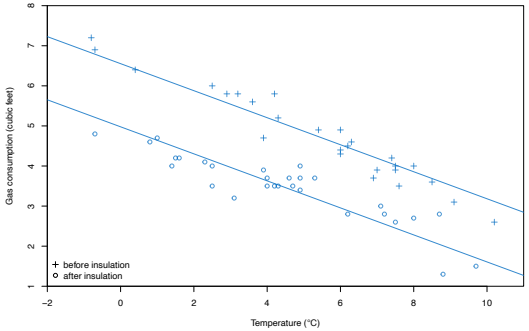
Drawing from the observed data



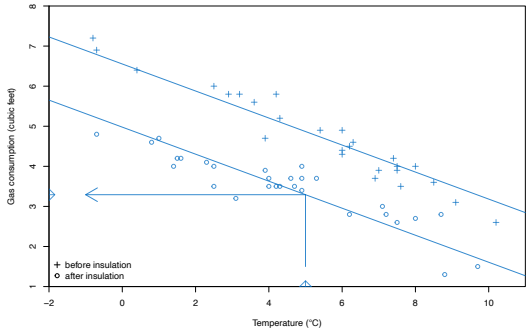
Predictive mean matching



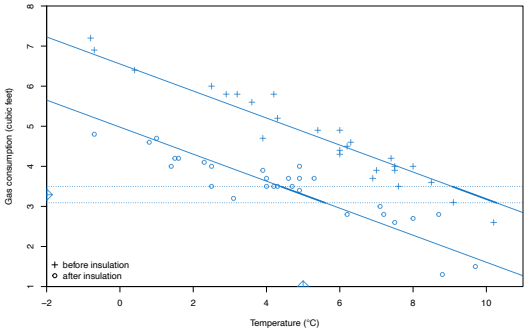
PMM: Add two regression lines



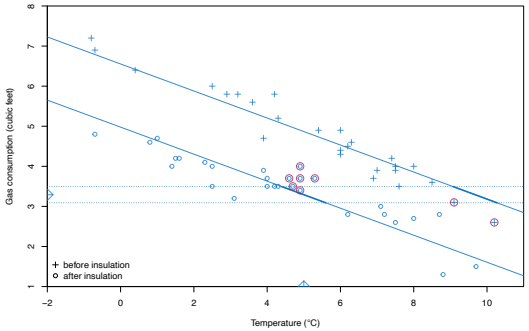
PMM: Predicted given 5°,C, 'after insulation'



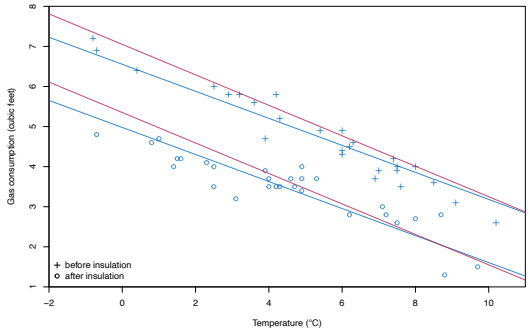
PMM: Define a matching range $\hat{y} \pm \delta$



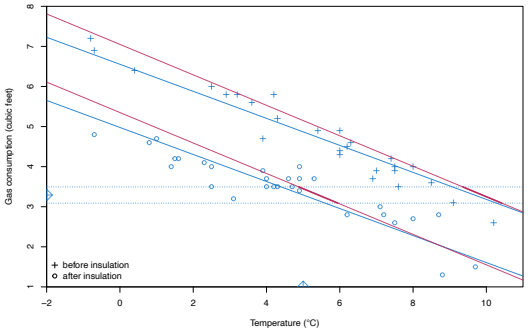
PMM: Select potential donors



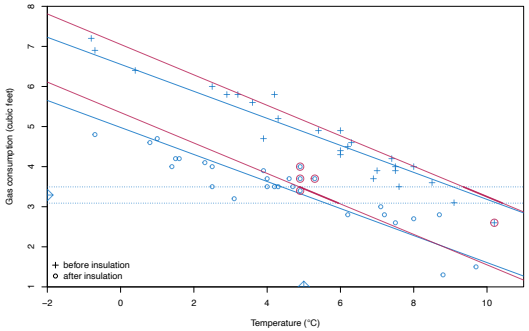
PMM: Bayesian PMM: Draw a line



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



PMM: Select potential donors

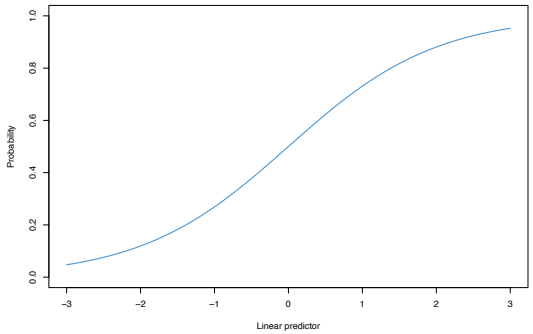


Imputation of a binary variable

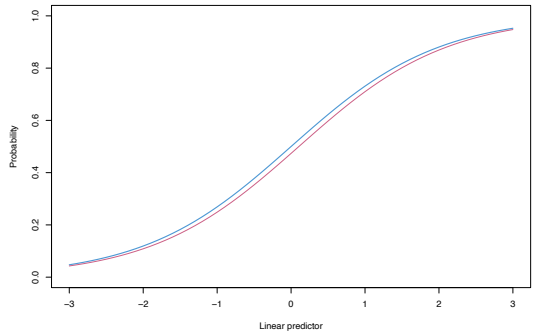
- Logistic regression

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

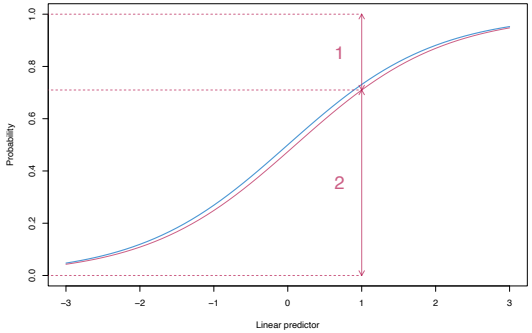
Fit logistic model



Draw parameter estimate



Read off the probability



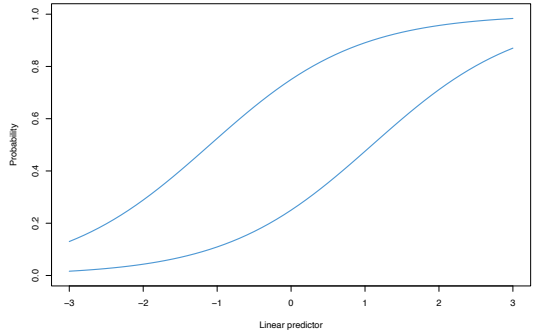
Impute ordered categorical variable

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

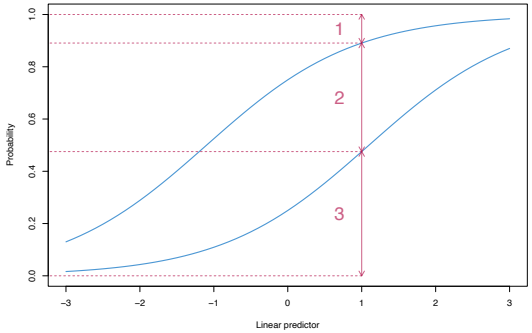
$$\Pr(y_i = k|X_i, \beta) = \frac{\exp(\tau_k + X_i\beta)}{\sum_{k=1}^{K} \exp(\tau_k + X_i\beta)}$$

-

Fit ordered logit model



Read off the probability



Built-in imputation functions

https://amices.org/mice/reference/index.html

Creating multivariate imputations, MICE algorithm

Issues in multivariate imputation

- The predictors Y_{-j} themselves can contain missing values;
- "Circular" dependence can occur, where Y_j^{mis} depends on Y_h^{mis}, and vice versa;
- Especially with large p and small n, collinearity or empty cells can occur;
- Derived variables;
- The ordering of the rows and columns can be meaningful, e.g., as in longitudinal data;
- Imputation can create impossible combinations, such as pregnant grandfathers.

Missing data patterns

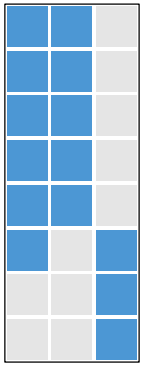
Three general strategies

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

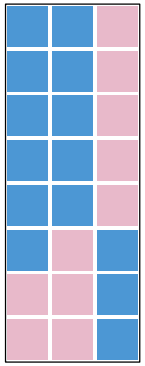
Fully conditional specification (FCS), MICE algorithm

- Imputes multivariate missing data on a variable-by-variable basis
- Requires a specification of an imputation model for each incomplete variable
- Creates imputations per variable in an iterative fashion

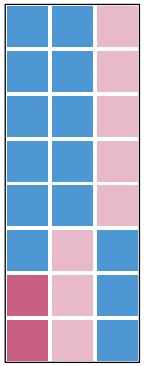
Imputation by fully conditional specification



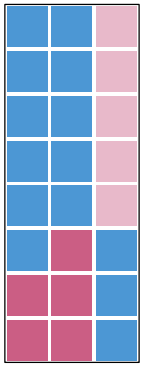
Imputation by fully conditional specification



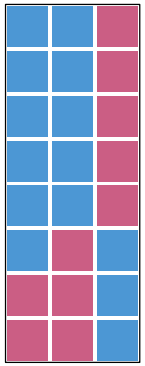
Imputation by fully conditional specification



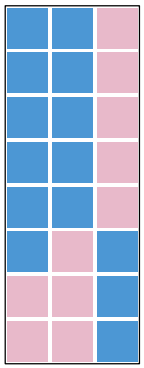
Imputation by fully conditional specification



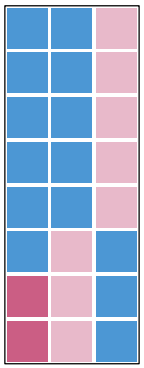
Imputation by fully conditional specification



Imputation by fully conditional specification - next iteration



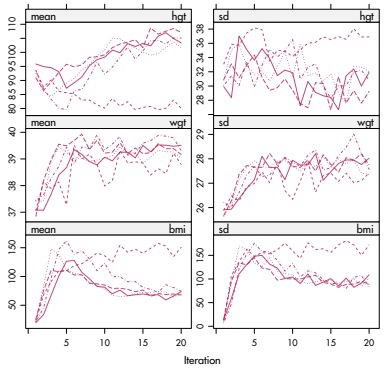
Imputation by fully conditional specification - next iteration



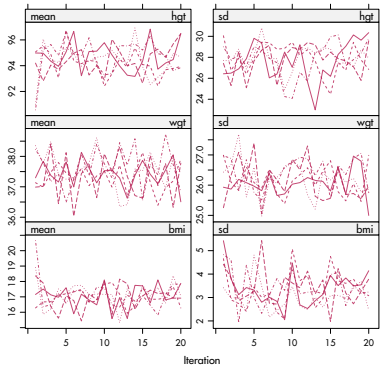
How many iterations?

- 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



Number of iterations

Watch out for situations where

- the correlations between the Y_j's are high;
- the missing data rates are high; or
- constraints on parameters across different variables exist.

Multiple imputation

- Disadvantages
  - Need to create and work with multiple imputed data sets
  - May not always be most efficient
- Advantages
  - Correct point and variance estimates
  - Splits missing data problem from complete-data analysis
  - Theoretical properties well established
  - Flexible, widely applicable
  - Extensible to MNAR
  - De facto standard

Advanced feature: Which cells to impute?

- By default, `mice` imputes the missing (`NA`) data
- The `where` argument in `mice 3.0.0` specifies which cells are imputed
- *Overimputation*: Impute everything, create synthetic data
- *Skip imputation*: Skip imputation of selected cells (e.g. BP for the dead)
- *Monotone block imputation*:
  - Impute only cells that destroy the monotone pattern
  - Impute only cells that conform to the monotone pattern

Why synthetic data?

- Cannot be traced back to the original
- Protects privacy
- Promotes data sharing
- Bypasses GDPR
- Eases data sharing
- Yields approximately same analytic results
- Multiple synthetic data: unbiased & proper coverage

Multiple synthetic data for nhanes: data

```
library(mice)
head(nhanes, 3)
```

```
##   age  bmi hyp chl
## 1   1   NA  NA  NA
## 2   2 22.7   1 187
## 3   1   NA   1 187
```

Synthetic data for nhanes: summary

```
summary(nhanes[, 2:4])
```

```
##       bmi             hyp             chl
##  Min.   :20.4   Min.   :1.00   Min.   :113
##  1st Qu.:22.6   1st Qu.:1.00   1st Qu.:185
##  Median :26.8   Median :1.00   Median :187
##  Mean   :26.6   Mean   :1.24   Mean   :191
##  3rd Qu.:28.9   3rd Qu.:1.00   3rd Qu.:212
##  Max.   :35.3   Max.   :2.00   Max.   :284
##  NA's   :9      NA's   :8      NA's   :10
```

Synthetic data creation

```
where <- make.where(nhanes, "all")
imp <- mice(nhanes, where = where, m = 2, seed = 5151, pri
head(complete(imp, 1), 3)
```

```
##   age  bmi hyp chl
## 1   1 26.3   2 238
## 2   2 27.4   2 187
## 3   1 26.3   1 118
```

```
head(complete(imp, 2), 3)
```

```
##   age  bmi hyp chl
## 1   1 22.0   1 113
## 2   1 22.5   1 131
## 3   1 27.2   1 238
```

Synthetic data inspection: dataset 1

```
summary(complete(imp, 1)[, 2:4])
```

```
##       bmi             hyp             chl
##  Min.   :20.4   Min.   :1.00   Min.   :113
##  1st Qu.:22.5   1st Qu.:1.00   1st Qu.:186
##  Median :26.3   Median :1.00   Median :187
##  Mean   :25.9   Mean   :1.48   Mean   :184
##  3rd Qu.:27.4   3rd Qu.:2.00   3rd Qu.:206
##  Max.   :35.3   Max.   :2.00   Max.   :238
```

Synthetic data inspection: dataset 2

```
summary(complete(imp, 2)[, 2:4])
```

```
##       bmi             hyp             chl
##  Min.   :20.4   Min.   :1.00   Min.   :113
##  1st Qu.:22.5   1st Qu.:1.00   1st Qu.:186
##  Median :26.3   Median :1.00   Median :204
##  Mean   :26.1   Mean   :1.24   Mean   :193
##  3rd Qu.:28.7   3rd Qu.:1.00   3rd Qu.:218
##  Max.   :35.3   Max.   :2.00   Max.   :238
```

Synthetic data analysis: fit models

```
fit <- with(imp, lm(bmi ~ age + chl))
```

Synthetic data analysis: pool estimates

```
est <- pool.syn(fit)
summary(est)
```

```
##          term estimate std.error statistic       df  p.v
## 1 (Intercept)  24.7115    4.5642     5.414 1.59e+06 6.16
## 2         age  -1.2228    1.2148    -1.007 1.36e+01 3.3
## 3         chl   0.0176    0.0224     0.786 4.99e+02 4.3
```

Synthetic data - future??

- Synthetic data: Projected to be multi-billion dollar market
- mice is free and has all the proper tools

Conclusion

- Dark data are a fact of life, and actually interesting
- There are many ways to treat missing data, only few are valid
- Always try to prevent missing data
- Use ad-hoc methods with caution
- Multiple imputation is an all-round general purpose method
- Many applications possible

That's it!