# Missing Data 1

MSBBSS01: Survey data analysis

Stef van Buuren, Gerko Vink

Nov 9, 2020

Course Overview

Nature and impact of missing data

Ad-hoc techniques

Multiple imputation

## Stef van Buuren

## Gerko Vink

## Course Overview

## Why deal with missing data?

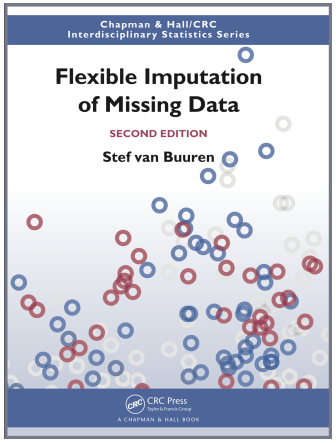
- Missing data are everywhere
- Missing data are the **heart of statistics**
- Ad-hoc fixes do not (always) work
- **Multiple imputation** is broadly applicable, yields correct statistical inferences
- Goal: get you comfortable with use of `mice` for imputing survey data


## Course materials

- INCLUDE URL HERE

## 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. https://stefvanbuuren.name/fimd

## mice software

- CRAN: `mice 3.11.0`
- `install.packages("mice")`
- Github: `mice 3.12.0`
- `devtools::install_github("amices/mice")`
- Results differ because of update in random sampler
- Slides are generated with `mice 3.12.0`

## Schedule

| Slot | Time | What | Topic |
|---|---|---|---|
| A | 10.00-10.45 | L | Missing data, ad-hoc methods |
| | 10.45-11.00 | | COFFEE/TEA |
| B | 11.00-11.45 | L | Multiple imputation, univariate |
| | 11.45-12.00 | | COFFEE/TEA |
| C | 12.00-13.00 | P | Three vignettes |

## Nature and impact of missing data

## 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

## Lion Air Indonesia - 29 Oct 2018 - 189 deaths

## Ethiopian Airlines - 10 Mar 2019 - 157 deaths

## What caused Boeing 737 Max to crash?

- Max introduced MCAS, a new course correction system
- MCAS was not mentioned in the flight manuals
- No action upon several "nosing down" reports made during 2018
- Lion Air & Ethiopian Air: Sensor produced faulty/missing input data
- MCAS wasn't prepared to deal with the faulty/missing data

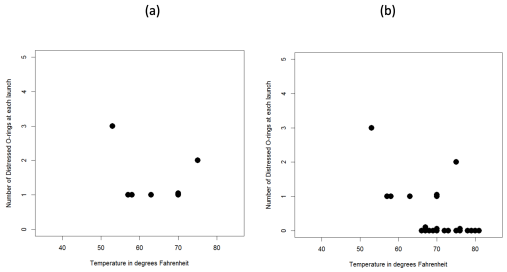
https://www.youtube.com/watch?v=H2tuKiiznsY

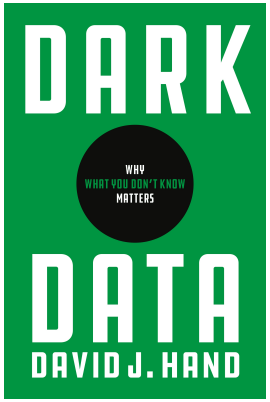
## Challenger space shuttle - 28 Jan 1986 - 7 deaths

## Challenger space shuttle - 28 Jan 1986 - 7 deaths

- What made the Challenger crash?

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

## Dark data types (1/2)

- DD-Type 1: **Data We Know Are Missing**
- DD-Type 2: Data We Don't Know are Missing
- DD-Type 3: Choosing Just Some Cases
- DD-Type 4: Self-Selection
- DD-Type 5: Missing What Matters
- DD-Type 6: **Data Which Might Have Been**
- DD-Type 7: Changes with Time
- DD-Type 8: Definitions of Data
- DD-Type 9: Summaries of Data
- DD-Type 10: Measurement Error and Uncertainty

## Dark data types (2/2)

- DD-Type 11: Feedback and Gaming
- DD-Type 12: Information Asymmetry
- DD-Type 13: **Intentionally Darkened Data**
- DD-Type 14: Fabricated and Synthetic Data
- DD-Type 15: Extrapolating beyond Your Data

## 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
- One possible reasons is non-response

## Types of non-response

Two types of non-response

- unit non-response: no observed response at all for a case
- item non-response: some, but not all, responses are missing for a case

You can classify missing values in three groups:

- Missing values that should have been observed (unintentional)
- Missing values that should not have been observed (intentional)
- Missing values whose true value can be deduced from the observed data (deductive missings)

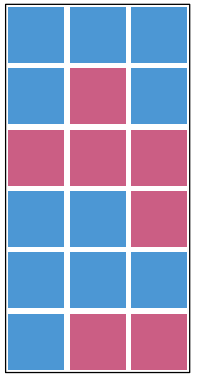
## Intentionality vs Response

| | Intentional | Unintentional |
|---|---|---|
| Unit nonresponse | Sampling | Refusal Self-selection |
| Item nonresponse | Branching Matrix Sampling | Skip question Coding error |

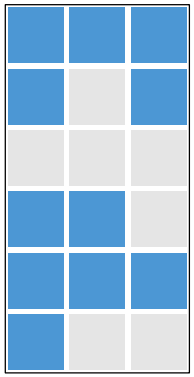
## Some confusing terminology

- 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

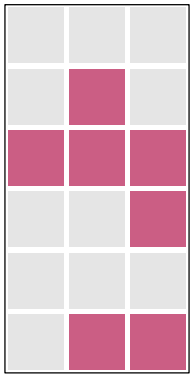
## Complete data



## Incomplete data = observed data



## Missing data = unobserved data



## Why values can be missing

Missingness can occur for a lot of reasons. For example

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

## Consequences of missing data

- 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

## Strategies to deal with missing data

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

## Ad-hoc techniques

## 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

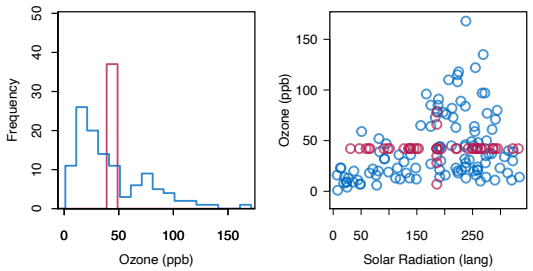
## Listwise deletion, complete-case analysis

- 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



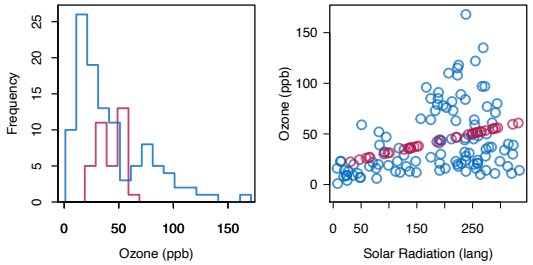

## Mean imputation

- 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 $Y^{\text{obs}}$ under listwise deletion
  - Predict $Y^{\text{mis}}$ 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



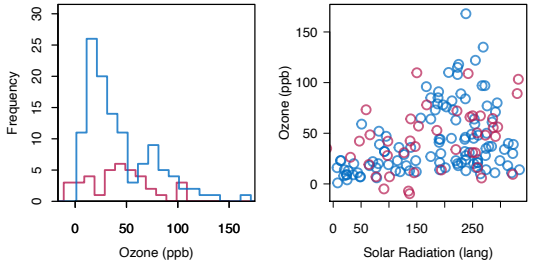

## Regression imputation

- 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 $Y^{\text{obs}}$
  - Preserves the correlation between $Y$ and $X$ in the imputed data

## Stochastic regression imputation




## Stochastic regression imputation

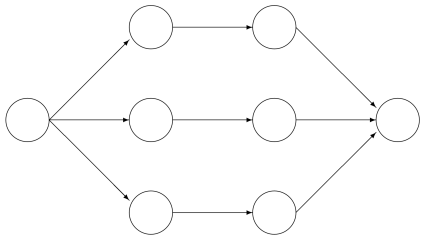
- Disadvantages
  - Symmetric and constant error restrictive
  - Single imputation: does not take uncertainty imputed data into account, and incorrectly treats them as real
  - Not so simple anymore

## 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

## Multiple imputation




## Acceptance of multiple imputation

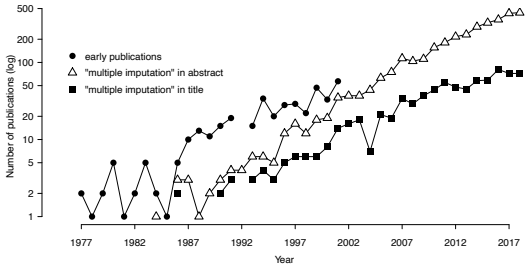



Figure 1: Source: Scopus (April 3, 2019)

## Pooled estimate $\bar{Q}$

$\hat{Q}_\ell$ is the estimate of the $\ell$-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 $\ell$-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 $\nu$ (c.f. section 2.3.6)

Relation between $\gamma$ and $\lambda$

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

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.

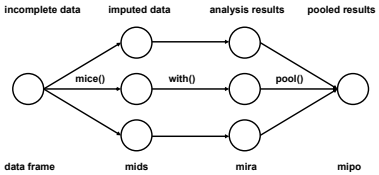
## How large should $m$ be?

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`



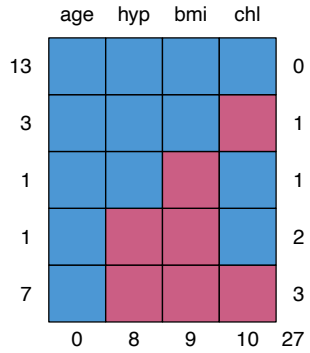

## Inspect the data

```
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

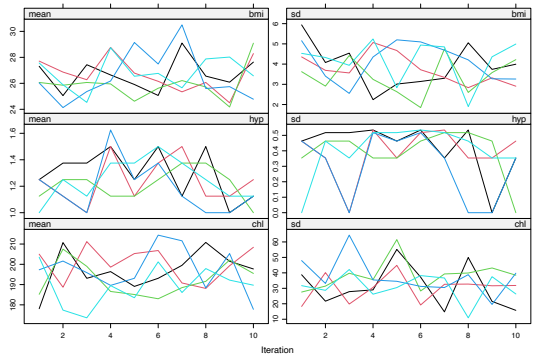
```
md.pattern(nhanes)
```




## Multiply impute the data

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

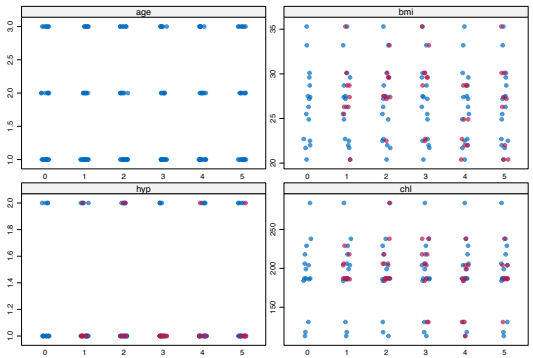
## Inspect the trace lines for convergence



## Stripplot of observed and imputed data

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

## Stripplot of observed and imputed data

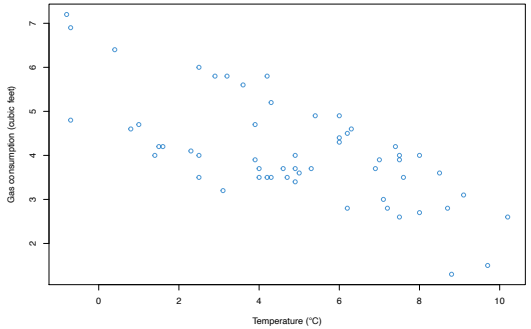



## Fit the complete-data model

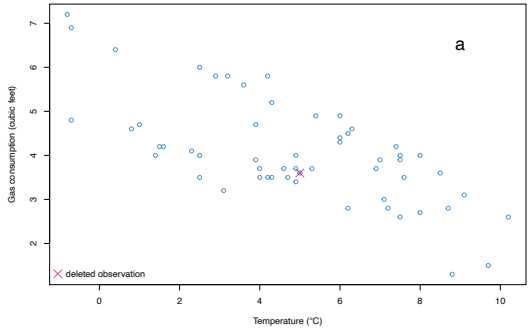
```
fit <- with(imp, lm(bmi ~ age))
est <- pool(fit)
summary(est)
```

```
##          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
```

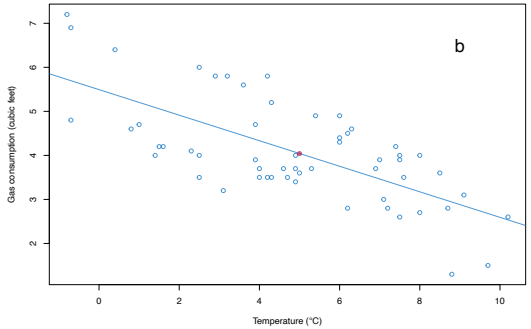
## Relation between temperature and gas consumption



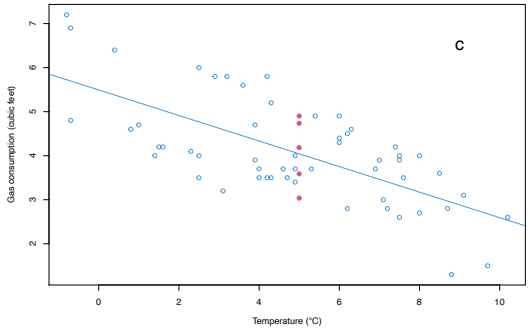

## We delete gas consumption of observation 47



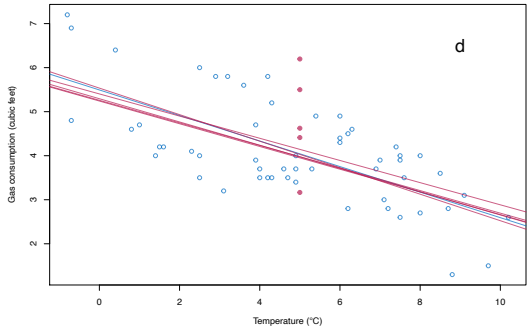

## Predict imputed value from regression line



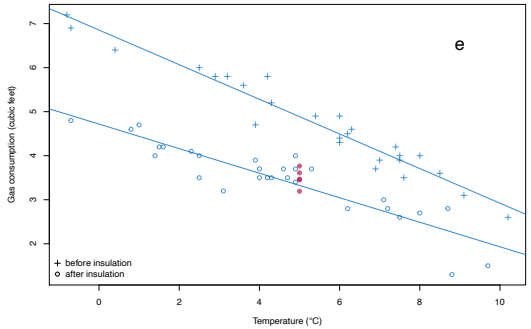

## Predicted value + noise



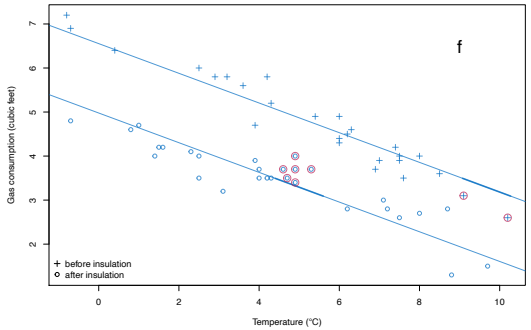

## Predicted value + noise + parameter uncertainty




## Imputation based on two predictors



## Drawing from the observed data




## Next week

- Predictive mean matching
- Categorical data
- Aproaches to multivariate missing data
- MICE algorithm
- Pooling