DEALING WITH MISSING DATA
LECTURE: UNSUPERVISED LEARNING AND EVOLUTIONARY COMPUTATION USING R

Jakob Bossek

MALEO Group, Department of Computer Science, Paderborn University, Germany

19th Jan, 2025

Missing Data

“The best solution to handle missing data is to have none.”

[R. A. Fisher]

I. e., carefully design your study, databases etc. w. r. t. probable causes of missing data!

“‘If something can go wrong, it will’, or
‘If there’s more than one possible outcome of a job or task, and one of those outcomes will result in disaster or an undesirable consequence, then somebody will do it that way.‘”

[Murphy’s Law, 1950]

I.e., missing data will most probably be an issue in real-world applications.

Reasons for Missing Data in Empirical Studies

Three broad categories related to

1. Study Participants:
   E.g. participant lost interest or got offended by a question
2. Study Design:
   E.g. study required too much time of the participants
3. Interaction of Participants and Study Design:
   E.g. in clinical trials: The sickest patients were physically unable to complete exhaustive aspects of a study.

Imputation of numerical features

How can we efficiently impute the missing values in order to ensure reliable statistical analysis?

Employee Selection Data Set

	IQ	Job Performance
1	78	9
2	84	13
3	84	10
4	85	8
5	87	7
6	91	7
7	92	9
8	94	9
9	94	11
10	96	7
11	99	7
12	105	10
13	105	11
14	106	15
15	108	10
16	112	10
17	113	12
18	115	14
19	118	16
20	134	12

Hypothetical employee selection scenario (Enders 2010)

- Prospective employees complete an IQ test
- Company hires applicants in upper half of IQ distribution
- Supervisor rates job-performance (JP) after a 6-month probatory period

A Missing Data Model

Rubin (Rubin 1987) defines missingness as a random variable that has a probability distribution.

- For each individual let X_{obs} be the observed data and X_{mis} the missing part of the data
- Define a random indicator variable

$$R = \begin{cases} 1, & \text{if data is not missing} \\ 0, & \text{if data is missing} \end{cases}$$

- The missing data model considers a distribution of missingness which potentially depends on (unknown!) distribution parameters ϕ and/or X_{obs} and X_{mis}.

Missing Data Categories

Rubin (Rubin 1987) defines three categories of missing data:

1. Missing Completely at Random (MCAR)
2. Missing at Random (MAR)
3. Missing not at Random (MNAR)

We will illustrate all three concepts following (Enders 2010, Chapter 1).

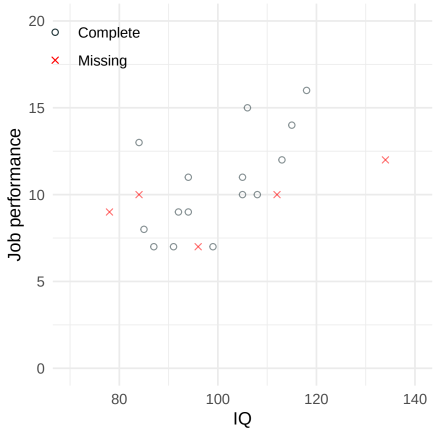
1. Missing Completely at Random (MCAR)

The probability of missing depends on the parameters θ only:

$$\Pr(R = 0 \mid \theta)$$

I.e., the observed data points form a random (unbiased) sample of the hypothetical complete data.

MCAR Example



- Job performance ratings after a 6-month probation period
- IQ was tested before hiring
- Probability of missing is **independent** of IQ or job performance

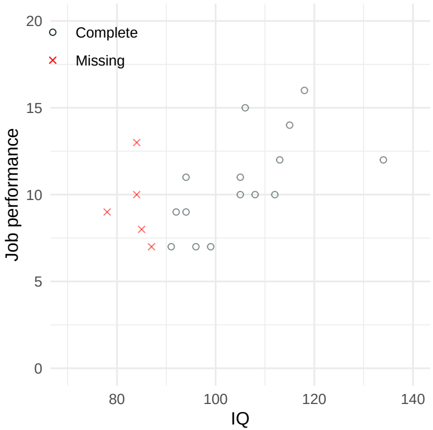
2. Missing at Random (MAR)

The probability of being missing is the same only within groups defined by the observed data.

$$\Pr(R = 0 \mid X_{\text{obs}}, \theta)$$

I.e., the probability is related to some other measured variable(s) but not to X_{mis} itself.

MAR Example



- Missing ratings for low IQ scores
- Probably those people not hired
- Probability of missing ratings is solely a function of IQ scores and unrelated to job performance

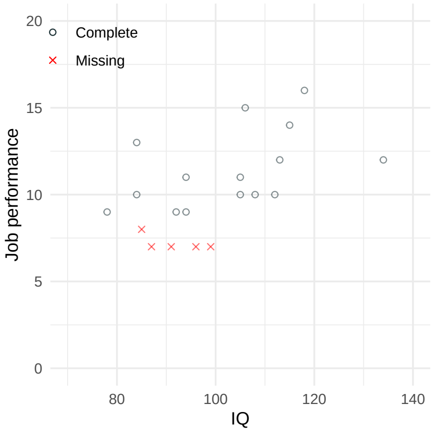
3. Missing not at Random (MNAR)

The probability of being missing is related to the variable itself, even after controlling for other variables.

$$\Pr(R = 0 \mid X_{\text{obs}}, X_{\text{mis}}, \theta)$$

I.e., the probability is related to unobserved information including X_{mis} itself.

MNAR Example



- 16 employees hired
- Some employees were fired prior to 6-month evaluation due to bad performance
- Probability of missing performance depends on the performance, even after controlling for IQ.

Examples

Tabular data given by Enders ()Enders 2010; for completeness):

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	NA	NA	9
84	13	13	NA	13
84	10	NA	NA	10
85	8	8	NA	NA
87	7	7	NA	NA
91	7	7	7	NA
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	NA	7	NA
99	7	7	7	NA
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	NA	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	NA	12	12

Implications of Missing Data Mechanisms

- There is unfortunately no way to definitely determine the missing data mechanism
- However, missing data mechanism theory supports analysis of reasons for missingness
- Imputation techniques desired which are ideally suitable for all mechanisms

Central Question

Under which condition is it possible to accurately estimate substantive parameters without the knowledge of the missing data distribution?
Most ad-hoc imputation techniques require the MCAR assumption to yield unbiased estimates.

Throw in some ideas!

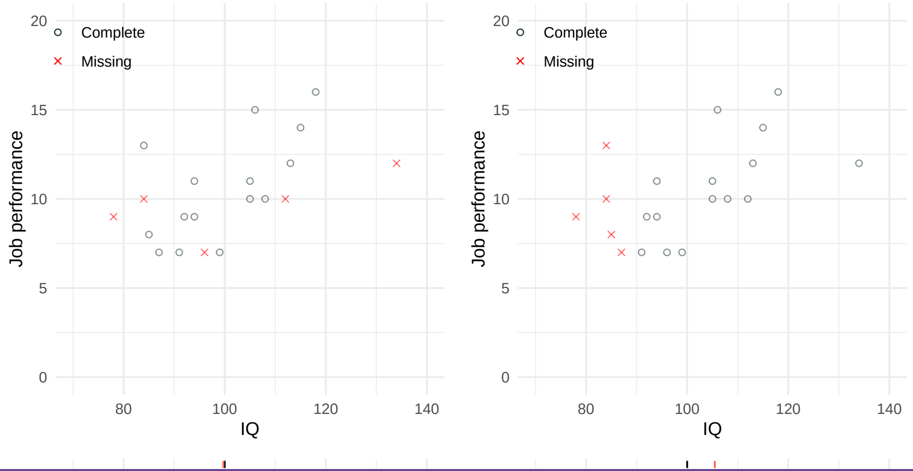
Complete-Case Analysis

Complete-case analysis (listwise deletion) removes all observations with at least one missing value.

- Common set of cases for all analysis
- Requires MCAR data!
  Otherwise remaining complete-cases will in general be unrepresentative of the hypothetically complete data $\rightsquigarrow$ biased estimates will be the rule rather the exception
- Very wasteful (obviously)
- Unfortunately, often used in social and behavioral sciences <**empty citation**> ☹
- Consequence: should not be used at all!
  Exception if there are really few data points with missing values.

Complete-Case Analysis

Caution: biased parameter estimates under MAR/MNAR (unbiased only under MCAR).



Biased estimators? Some theory.

Definition (Unbiased estimator)

Formally, an estimator $\hat{S}$ for a statistic S is unbiased if

$$E(\hat{S}) = S.$$

I.e., if the estimation on average corresponds to the true value!

Example

Think of S being the *expected value* of a distribution and $\hat{S}$ being the *arithmetic mean of an empirical sample*. Unbiased means, that if we sample n times and get estimations $\hat{s}_1, \dots, \hat{s}_n$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{s}_i \xrightarrow{n\to\infty} S.$$

Available-Case Analysis

Drop or keep data on a analyis-by-analysis basis:

- For each analysis keep all available data! E.g., if we have two features x, y use all observed x values to calculate $\bar{x}$. Likewise keep all observed y values to calculate $\bar{y}$.[1]
- Different subset for each analysis possible $\rightsquigarrow$ Can cause severe problems. E.g.,

$$r_{XY} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N \cdot \sqrt{{s_x}^2 \cdot {s_y}^2}}$$

  may yield values $\notin [-1, 1]$.
- Again, requires MCAR assumption to hold. Otherwise strong bias is possible.
- Consequence: should not be used at all!

[1] In R we can use `mean(x, na.rm = TRUE)` and `mean(y, na.rm = TRUE)`.

Deletion Methods

Remarks

- Plain simple and convenient and thus very tempting!
- Implemented in many statistical libraries
- Very wasteful if there is much missing data
- Reduces statistical power
- Very likely to produce biased estimates if MCAR is violated

“The two popular methods for dealing with missing data that are found in basic statistical packages – listwise and pairwise deletion of missing values – are among the worst methods available for practical applications.”

[Wilkinson & Task Force on Statistical Inference, 1999, p. 598]

Imputation is . . .

Cambridge Dictionary[2]

1. ‘A suggestion that someone is guilty of something or has a particular bad quality’
   Not what we do here.
2. ‘A way of calculating something when you do not have the full or correct data’
   Exactly what we are going to do.

[2] https://www.dictionary.cambridge.org

Another data set

Airquality

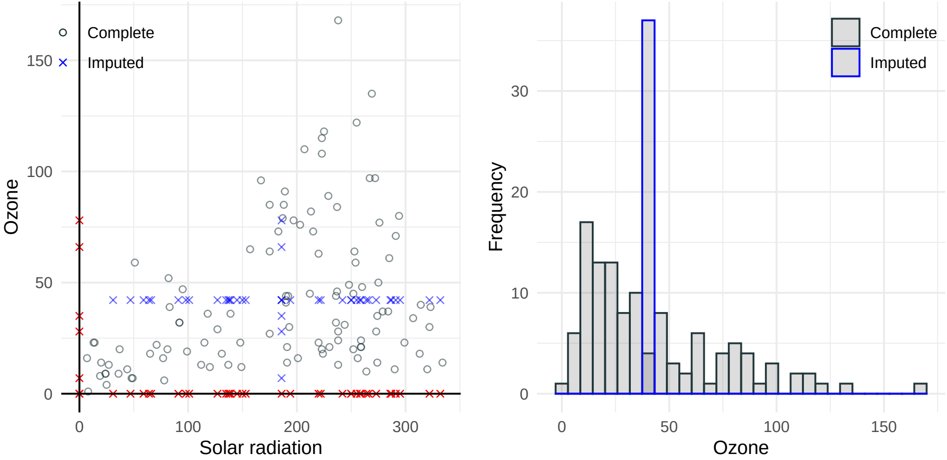
Daily air quality measurement[3] in New York, May to September 1973 (Chambers et al. 1983)

Feature	Description
Ozone (**35 NAs**)	Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
Solar.R. (**7 NAs**)	Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park
Wind	Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
Temp	Maximum daily temperature in degrees Fahrenheit at La Guardia Airport

[3] `data(airquality)` in R.

Mean Imputation

Replace mssing values with the mean of the respective observed values.



Mean Imputation

Remarks

- Super simple!
- Way better than deletion methods
- Standard deviation of imputed feature decreases

  Because we impute the same value for every single missing value.
- Correlation decreases

  Same reason!
- In total the entire distribution gets distorted (very undesireble)

Regression Imputation (RI)

Train a (linear) regression model

$$X_i = \sum_{\substack{j=0,\\ j\neq i}}^{p} \beta_j \cdot X_j + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

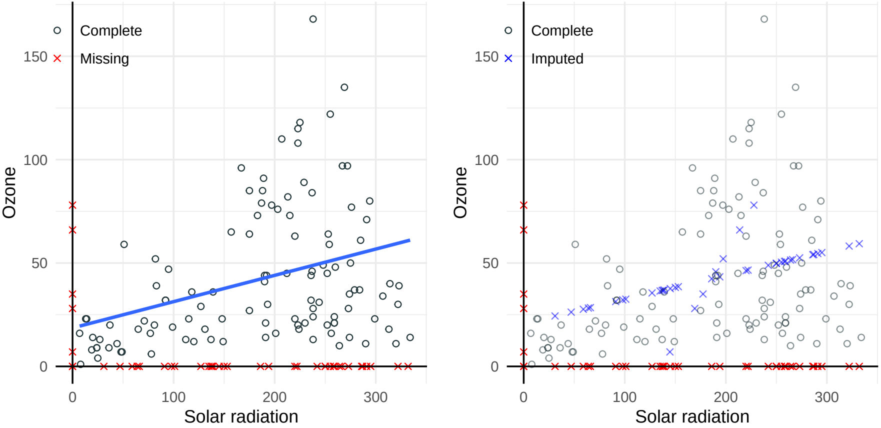
where X_i is the target variable (which contains missing values) and the model is trained on the remaining variables

- Training on complete cases only
- Use the fitted model's predictions to impute missing X_i observations

  Imputed values most likely under fitted model.

Regression Imputation (RI)

Model(s): $\texttt{Ozone} = \beta_0 + \beta_1 \cdot \texttt{Solar.R} + \varepsilon$ and $\texttt{Solar.R} = \beta_0 + \beta_1 \cdot \texttt{Ozone} + \varepsilon$



Regression Imputation (RI)

Remarks

- Appealing approach!

  Using knowledge encoded in other variables.

- If model is very accurate, realistic imputation values possible

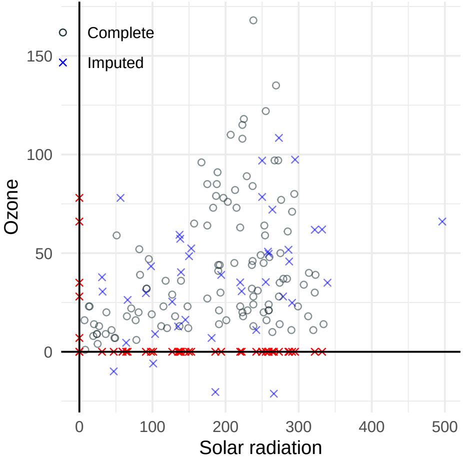
  For linear models this would be the case if there was a very strong correlation between fitted variable and variables used for training.

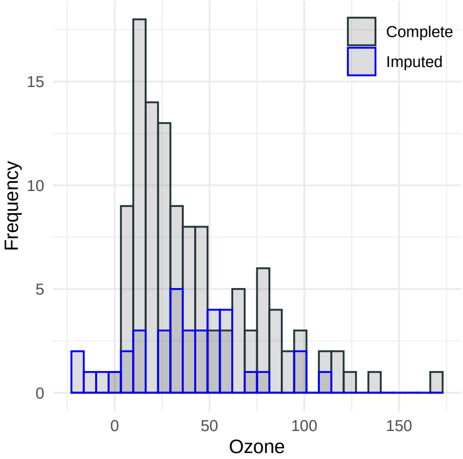
- Not restricted to linear models[4]

[4] Polynomial models perfectly possible.

Stochastic Regression Imputation (SRI)

Regression imputation with additional noise w. r. t. error distribution of the linear model.





Stochastic Regression Imputation (SRI)

Remarks

- Reasonable extension of regression imputation
- Problem: noise may introduce infeasible values

  E.g., negative values for physical measurements,negative income, negative age etc.
- Problem: does not account for heteroscedastic data

  I.e., data where the variance is non-constant
- Basis of many more advanced imputation techniques

Predictive Mean Matching (PMM)

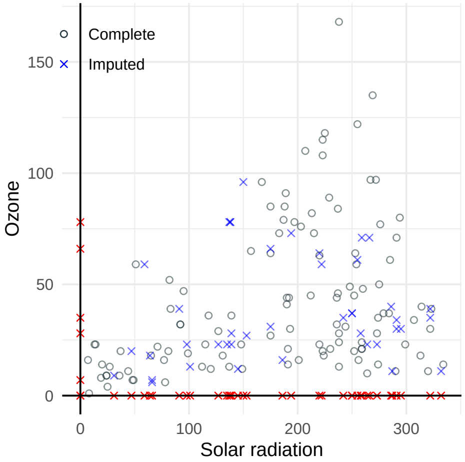
Predictive Mean Matching (PMM) (PMM) in a simplified version:

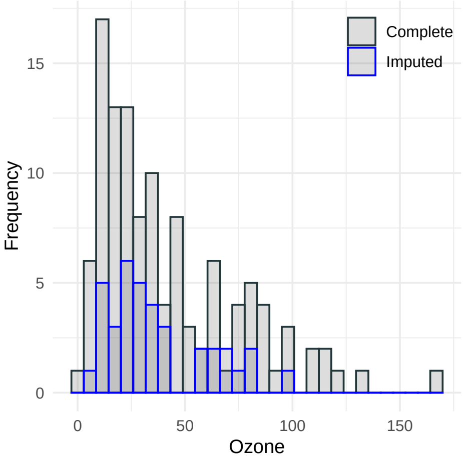
- Train regression model and get prediction for missing value x
- Select the k data points whose predictions with the model are closest to the x. These points serve as potential *donors*[5]
- Sample one of these k donors uniformly at random and use its observed value (not the predicted value) for imputation ("nearest neighbor hot deck" prediction)
- Ensures that imputed values are realistic

  Since true observations are used.
- Thus avoids meaningless imputation (see SRI)

[5] This is just one possiblity. Many more extensions have been proposed.

Predictive Mean Matching (PMM)





Predictive Mean Matching (PMM)

Remarks

- Recommended method with various modifications in particular how to choose the donor(s) (Siddique and Belin 2008; Schenker and Taylor 1996)
- Quite robust against misspecification of the imputation model
- Can be used for discrete data
- Well suited for large samples
- Imputations possess characteristics of the complete data
- Cannot extrapolate beyond the data range
- Severe problems if data is sparse in certain parts
- Not suited for small data sets
- Risk of choosing the same donors repeatedly

Comparison of Classical Imputation Methods

Assumptions on the missing data mechanism which underlies each method so that unbiased estimates are generated.

	Unbiased			**Standard Error**
	Mean	Reg. Coeff	Correlation	
Complete Case	MCAR	MCAR	MCAR	Too large
Mean Imputation	MCAR			Too small
Regression Imp.	MAR	MAR		Too small
Stoch. Reg. Imp.	MAR	MAR	MAR	Too small

No method is perfect!

⇝ Multiple imputation is a promising solution.

Core Idea

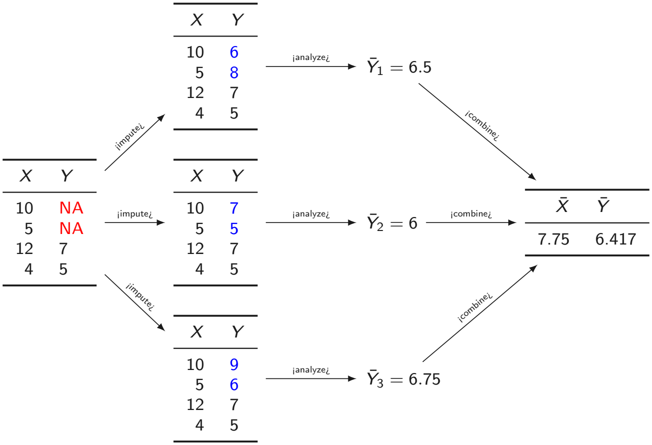
Do multiple indepenent ”runs” of imputation. I.e.,

$$\begin{tikzcd}
 & X_1 \arrow[r] & \theta_1 \arrow[ddr] & \\
 & X_2 \arrow[r] & \theta_2 \arrow[dr] & \\
X \arrow[uur] \arrow[ur] \arrow[r] \arrow[dr] \arrow[ddr] & X_3 \arrow[r] & \theta_3 \arrow[r] & \bar{\theta} \\
 & \cdots \arrow[r] & \cdots \arrow[ur] & \\
 & X_m \arrow[r] & \theta_m \arrow[uur] &
\end{tikzcd}$$

1. **Imputation**: Generate $m > 1$ independent versions of the imputed data (usually $m \approx 5$)
   $\rightsquigarrow$ m copies of data completed with different imputed values.
2. **Analysis**: Do whatever statistical analysis on each data set, i.e., calculate parameters of interest
   $\rightsquigarrow$ ”pool” of m parameter estimates
3. **Pooling**: Combine m estimates into a single estimate (pooling)

Example

Parameter of interest $\bar{Y}$ via MI with $m = 3$:



X	Y
10	NA
5	NA
12	7
4	5

¡impute¿

X	Y
10	6
5	8
12	7
4	5

X	Y
10	7
5	5
12	7
4	5

X	Y
10	9
5	6
12	7
4	5

¡analyze¿

$\bar{Y}_1 = 6.5$

$\bar{Y}_2 = 6$

$\bar{Y}_3 = 6.75$

¡combine¿

$\bar{X}$	$\bar{Y}$
7.75	6.417

Multiple Imputation

Properties

- Deals with inherent uncertainty of the (imputed) values. I.e., accounts for the statistical uncertainty in the imputations.
- Solves the problem of too small standard errors
- Given certain conditions, pooled estimates are unbiased
- Sounds tedious, right?

  Luckily, multiple imputation packages automate the process, e.g., `mice` Buuren and Groothuis-Oudshoorn 2011 in R.

Pooled Parameter Estimation

Given $m > 1$ estimates

$$\theta_1, \ldots, \theta_m$$

the *pooled estimate* $\bar{\theta}$ is simply the average (Rubin 1987)

$$\bar{\theta} = \frac{1}{m} \sum_{i=1}^{m} \theta_i.$$

Pooled Standard Error / Variation

Slightly more involved since there are multiple sources of variation:

1. Within-Imputation Variance: Sampling error that would have resulted had the data been complete

2. Between-Imputation Variance: Additional sampling error introduced by imputed data

Pooled Standard Error / Variation

Within-Imputation Variance

$$V_W = \frac{1}{m} \sum_{i=1}^{m} \operatorname{Var} X_i$$

$\rightsquigarrow$ measures the variability of the imputed data sets

Between-Imputation Variance

$$V_B = \operatorname{Var}(\theta_1, \ldots, \theta_m) = \frac{1}{m-1} \sum_{i=1}^{m} \left(\theta_i - \bar{\theta}\right)^2$$

$\rightsquigarrow$ measures the variability of the parameter estimate over all m imputed data sets

Pooled Standard Error / Variation

Total Sampling Variance

$$V = V_W + V_B + \underbrace{\frac{V_B}{m}}_{\text{correction}}$$

The correction term is due to V_B using $\bar{\theta}$ and the latter being also subject to variation. However,

$$\frac{V_B}{m} \xrightarrow{m \to \infty} 0.$$

Conclusion

- Missing data should be avoided through careful experimental study design
- Sobering fact: in practice we will be confronted with missing data
  Unanswered questions, defect sensors, failed computational experiments etc.
- Deletion methods should be used only if number of missing observations is very low
- Inappropriate handling of missing values can induce severe bias and lead to faulty interpretations
- Multiple imputation in combination with PMM is recommended

References I

Enders, C.K. (2010). *Applied Missing Data Analysis*. Methodology in the social sciences. Guilford Publications. ISBN: 9781606236390.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, p. 258.

Chambers, J.M. et al. (1983). "Graphical Methods for Data Analysis". In: *The Wadsworth Statistics/Probability Series. Boston, MA: Duxury*.

Siddique, Juned and Thomas R. Belin (2008). "Multiple imputation using an iterative hot-deck with distance-based donor selection.". In: *Statistics in medicine* 27 1, pp. 83–102.

Schenker, Nathaniel and Jeremy M.G. Taylor (1996). "Partially parametric techniques for multiple imputation". In: *Computational Statistics & Data Analysis* 22.4, pp. 425–446. ISSN: 0167-9473. DOI: `https://doi.org/10.1016/0167-9473(95)00057-7`.

Buuren, Stef van and Karin Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3, pp. 1–67. DOI: `10.18637/jss.v045.i03`.