
Contents

Foreword	xvii
Preface	xix
About the Author	xxi
Symbol Description	xxiii
List of Algorithms	xxv
I Basics	1
1 Introduction	3
1.1 The problem of missing data	3
1.1.1 Current practice	3
1.1.2 Changing perspective on missing data	5
1.2 Concepts of MCAR, MAR and MNAR	6
1.3 Simple solutions that do not (always) work	8
1.3.1 Listwise deletion	8
1.3.2 Pairwise deletion	9
1.3.3 Mean imputation	10
1.3.4 Regression imputation	11
1.3.5 Stochastic regression imputation	13
1.3.6 LOCF and BOFC	14
1.3.7 Indicator method	15
1.3.8 Summary	15
1.4 Multiple imputation in a nutshell	16
1.4.1 Procedure	16
1.4.2 Reasons to use multiple imputation	17
1.4.3 Example of multiple imputation	18
1.5 Goal of the book	20
1.6 What the book does not cover	20
1.6.1 Prevention	21
1.6.2 Weighting procedures	21
1.6.3 Likelihood-based approaches	22
1.7 Structure of the book	23
1.8 Exercises	23
	ix

2	Multiple imputation	25
2.1	Historic overview	25
2.1.1	Imputation	25
2.1.2	Multiple imputation	25
2.1.3	The expanding literature on multiple imputation . . .	27
2.2	Concepts in incomplete data	28
2.2.1	Incomplete data perspective	28
2.2.2	Causes of missing data	29
2.2.3	Notation	30
2.2.4	MCAR, MAR and MNAR again	31
2.2.5	Ignorable and nonignorable \spadesuit	33
2.2.6	Implications of ignorability	34
2.3	Why and when multiple imputation works	35
2.3.1	Goal of multiple imputation	35
2.3.2	Three sources of variation \spadesuit	36
2.3.3	Proper imputation	38
2.3.4	Scope of the imputation model	40
2.3.5	Variance ratios \spadesuit	41
2.3.6	Degrees of freedom \spadesuit	42
2.3.7	Numerical example	43
2.4	Statistical intervals and tests	44
2.4.1	Scalar or multi-parameter inference?	44
2.4.2	Scalar inference	44
2.5	Evaluation criteria	45
2.5.1	Imputation is not prediction	45
2.5.2	Simulation designs and performance measures	47
2.6	When to use multiple imputation	48
2.7	How many imputations?	49
2.8	Exercises	51
3	Univariate missing data	53
3.1	How to generate multiple imputations	53
3.1.1	Predict method	55
3.1.2	Predict + noise method	55
3.1.3	Predict + noise + parameter uncertainty	55
3.1.4	A second predictor	56
3.1.5	Drawing from the observed data	56
3.1.6	Conclusion	56
3.2	Imputation under the normal linear normal	57
3.2.1	Overview	57
3.2.2	Algorithms \spadesuit	57
3.2.3	Performance	59
3.2.4	Generating MAR missing data	63
3.2.5	Conclusion	64
3.3	Imputation under non-normal distributions	65

3.3.1	Overview	65
3.3.2	Imputation from the t -distribution ♣	66
3.3.3	Example ♣	67
3.4	Predictive mean matching	68
3.4.1	Overview	68
3.4.2	Computational details ♣	70
3.4.3	Algorithm ♣	73
3.4.4	Conclusion	74
3.5	Categorical data	75
3.5.1	Overview	75
3.5.2	Perfect prediction ♣	76
3.6	Other data types	78
3.6.1	Count data	78
3.6.2	Semi-continuous data	79
3.6.3	Censored, truncated and rounded data	79
3.7	Classification and regression trees	82
3.7.1	Overview	82
3.7.2	Imputation using CART models	83
3.8	Multilevel data	84
3.8.1	Overview	84
3.8.2	Two formulations of the linear multilevel model ♣	85
3.8.3	Computation ♣	86
3.8.4	Conclusion	87
3.9	Nonignorable missing data	88
3.9.1	Overview	88
3.9.2	Selection model	89
3.9.3	Pattern-mixture model	90
3.9.4	Converting selection and pattern-mixture models	90
3.9.5	Sensitivity analysis	92
3.9.6	Role of sensitivity analysis	93
3.10	Exercises	93
4	Multivariate missing data	95
4.1	Missing data pattern	95
4.1.1	Overview	95
4.1.2	Summary statistics	96
4.1.3	Influx and outflux	99
4.2	Issues in multivariate imputation	101
4.3	Monotone data imputation	102
4.3.1	Overview	102
4.3.2	Algorithm	103
4.4	Joint modeling	105
4.4.1	Overview	105
4.4.2	Continuous data ♣	105
4.4.3	Categorical data	107

4.5	Fully conditional specification	108
4.5.1	Overview	108
4.5.2	The MICE algorithm	109
4.5.3	Performance	111
4.5.4	Compatibility ♣	111
4.5.5	Number of iterations	112
4.5.6	Example of slow convergence	113
4.6	FCS and JM	116
4.6.1	Relations between FCS and JM	116
4.6.2	Comparison	117
4.6.3	Illustration	117
4.7	Conclusion	121
4.8	Exercises	121
5	Imputation in practice	123
5.1	Overview of modeling choices	123
5.2	Ignorable or nonignorable?	125
5.3	Model form and predictors	126
5.3.1	Model form	126
5.3.2	Predictors	127
5.4	Derived variables	129
5.4.1	Ratio of two variables	129
5.4.2	Sum scores	132
5.4.3	Interaction terms	133
5.4.4	Conditional imputation	133
5.4.5	Compositional data ♣	136
5.4.6	Quadratic relations ♣	139
5.5	Algorithmic options	140
5.5.1	Visit sequence	140
5.5.2	Convergence	142
5.6	Diagnostics	146
5.6.1	Model fit versus distributional discrepancy	146
5.6.2	Diagnostic graphs	146
5.7	Conclusion	151
5.8	Exercises	152
6	Analysis of imputed data	153
6.1	What to do with the imputed data?	153
6.1.1	Averaging and stacking the data	153
6.1.2	Repeated analyses	154
6.2	Parameter pooling	155
6.2.1	Scalar inference of normal quantities	155
6.2.2	Scalar inference of non-normal quantities	155
6.3	Statistical tests for multiple imputation	156
6.3.1	Wald test ♣	157

6.3.2	Likelihood ratio test \spadesuit	157
6.3.3	χ^2 -test \spadesuit	159
6.3.4	Custom hypothesis tests of model parameters \spadesuit	159
6.3.5	Computation	160
6.4	Stepwise model selection	162
6.4.1	Variable selection techniques	162
6.4.2	Computation	163
6.4.3	Model optimism	164
6.5	Conclusion	166
6.6	Exercises	166

II Case studies 169

7 Measurement issues 171

7.1	Too many columns	171
7.1.1	Scientific question	172
7.1.2	Leiden 85+ Cohort	172
7.1.3	Data exploration	173
7.1.4	Outflux	175
7.1.5	Logged events	176
7.1.6	Quick predictor selection for wide data	177
7.1.7	Generating the imputations	179
7.1.8	A further improvement: Survival as predictor variable	180
7.1.9	Some guidance	181
7.2	Sensitivity analysis	182
7.2.1	Causes and consequences of missing data	182
7.2.2	Scenarios	184
7.2.3	Generating imputations under the δ -adjustment	185
7.2.4	Complete data analysis	186
7.2.5	Conclusion	187
7.3	Correct prevalence estimates from self-reported data	188
7.3.1	Description of the problem	188
7.3.2	Don't count on predictions	189
7.3.3	The main idea	190
7.3.4	Data	191
7.3.5	Application	192
7.3.6	Conclusion	193
7.4	Enhancing comparability	194
7.4.1	Description of the problem	194
7.4.2	Full dependence: Simple equating	195
7.4.3	Independence: Imputation without a bridge study	196
7.4.4	Fully dependent or independent?	198
7.4.5	Imputation using a bridge study	199
7.4.6	Interpretation	202
7.4.7	Conclusion	203

7.5	Exercises	204
8	Selection issues	205
8.1	Correcting for selective drop-out	205
8.1.1	POPS study: 19 years follow-up	205
8.1.2	Characterization of the drop-out	206
8.1.3	Imputation model	207
8.1.4	A degenerate solution	208
8.1.5	A better solution	210
8.1.6	Results	211
8.1.7	Conclusion	211
8.2	Correcting for nonresponse	212
8.2.1	Fifth Dutch Growth Study	212
8.2.2	Nonresponse	213
8.2.3	Comparison to known population totals	213
8.2.4	Augmenting the sample	214
8.2.5	Imputation model	215
8.2.6	Influence of nonresponse on final height	217
8.2.7	Discussion	218
8.3	Exercises	219
9	Longitudinal data	221
9.1	Long and wide format	221
9.2	SE Fireworks Disaster Study	223
9.2.1	Intention to treat	224
9.2.2	Imputation model	225
9.2.3	Inspecting imputations	227
9.2.4	Complete data analysis	228
9.2.5	Results from the complete data analysis	229
9.3	Time raster imputation	230
9.3.1	Change score	231
9.3.2	Scientific question: Critical periods	232
9.3.3	Broken stick model ♣	234
9.3.4	Terneuzen Birth Cohort	236
9.3.5	Shrinkage and the change score ♣	237
9.3.6	Imputation	238
9.3.7	Complete data analysis	240
9.4	Conclusion	242
9.5	Exercises	244
III	Extensions	247

10 Conclusion	249
10.1 Some dangers, some do's and some don'ts	249
10.1.1 Some dangers	249
10.1.2 Some do's	250
10.1.3 Some don'ts	251
10.2 Reporting	251
10.2.1 Reporting guidelines	252
10.2.2 Template	254
10.3 Other applications	255
10.3.1 Synthetic datasets for data protection	255
10.3.2 Imputation of potential outcomes	255
10.3.3 Analysis of coarsened data	256
10.3.4 File matching of multiple datasets	256
10.3.5 Planned missing data for efficient designs	256
10.3.6 Adjusting for verification bias	257
10.3.7 Correcting for measurement error	257
10.4 Future developments	257
10.4.1 Derived variables	257
10.4.2 Convergence of MICE algorithm	257
10.4.3 Algorithms for blocks and batches	258
10.4.4 Parallel computation	258
10.4.5 Nested imputation	258
10.4.6 Machine learning for imputation	259
10.4.7 Incorporating expert knowledge	259
10.4.8 Distribution-free pooling rules	259
10.4.9 Improved diagnostic techniques	260
10.4.10 Building block in modular statistics	260
10.5 Exercises	260
A Software	263
A.1 R	263
A.2 S-PLUS	265
A.3 Stata	265
A.4 SAS	266
A.5 SPSS	266
A.6 Other software	266
References	269
Author Index	299
Subject Index	307