Contents

Foreword by Donald B. Rubin xv

Preface to second edition xvii

Preface to first edition xxi

About the author xxiii

List of symbols xxv

List of algorithms xxvii

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 6
- 1.2 Concepts of MCAR, MAR and MNAR 8
- 1.3 Ad-hoc solutions 9
  - 1.3.1 Listwise deletion 9
  - 1.3.2 Pairwise deletion 11
  - 1.3.3 Mean imputation 12
  - 1.3.4 Regression imputation 13
  - 1.3.5 Stochastic regression imputation 14
  - 1.3.6 LOCF and BOCF 16
  - 1.3.7 Indicator method 17
  - 1.3.8 Summary 18
- 1.4 Multiple imputation in a nutshell 19
  - 1.4.1 Procedure 19
  - 1.4.2 Reasons to use multiple imputation 20
  - 1.4.3 Example of multiple imputation 21
- 1.5 Goal of the book 23
- 1.6 What the book does not cover 23
  - 1.6.1 Prevention 24
  - 1.6.2 Weighting procedures 24
  - 1.6.3 Likelihood-based approaches 25

1.7 Structure of the book . . . 26
1.8 Exercises . . . 26

2 Multiple imputation 29
2.1 Historic overview . . . 29
2.1.1 Imputation . . . 29
2.1.2 Multiple imputation . . . 30
2.1.3 The expanding literature on multiple imputation . . . 32
2.2 Concepts in incomplete data . . . 33
2.2.1 Incomplete-data perspective . . . 33
2.2.2 Causes of missing data . . . 33
2.2.3 Notation . . . 35
2.2.4 MCAR, MAR and MNAR again . . . 36
2.2.5 Ignorable and nonignorable♠ . . . 38
2.2.6 Implications of ignorability . . . 39
2.3 Why and when multiple imputation works . . . 41
2.3.1 Goal of multiple imputation . . . 41
2.3.2 Three sources of variation♠ . . . 41
2.3.3 Proper imputation . . . 44
2.3.4 Scope of the imputation model . . . 46
2.3.5 Variance ratios♠ . . . 46
2.3.6 Degrees of freedom♠ . . . 47
2.3.7 Numerical example . . . 48
2.4 Statistical intervals and tests . . . 49
2.4.1 Scalar or multi-parameter inference? . . . 49
2.4.2 Scalar inference . . . 50
2.4.3 Numerical example . . . 50
2.5 How to evaluate imputation methods . . . 51
2.5.1 Simulation designs and performance measures . . . 51
2.5.2 Evaluation criteria . . . 52
2.5.3 Example . . . 53
2.6 Imputation is not prediction . . . 55
2.7 When not to use multiple imputation . . . 57
2.8 How many imputations? . . . 58
2.9 Exercises . . . 61

3 Univariate missing data 63
3.1 How to generate multiple imputations . . . 63
3.1.1 Predict method . . . 65
3.1.2 Predict + noise method . . . 65
3.1.3 Predict + noise + parameter uncertainty . . . 65
3.1.4 A second predictor . . . 66
3.1.5 Drawing from the observed data . . . 66
3.1.6 Conclusion . . . 66
3.2 Imputation under the normal linear normal . . . 67

3.2.1 Overview 67
3.2.2 Algorithms♠ 67
3.2.3 Performance 69
3.2.4 Generating MAR missing data 70
3.2.5 MAR missing data generation in multivariate data . . 72
3.2.6 Conclusion 73
3.3 Imputation under non-normal distributions 74
3.3.1 Overview 74
3.3.2 Imputation from the t-distribution 75
3.4 Predictive mean matching 77
3.4.1 Overview 77
3.4.2 Computational details♠ 79
3.4.3 Number of donors 81
3.4.4 Pitfalls 82
3.4.5 Conclusion 84
3.5 Classification and regression trees 84
3.5.1 Overview 84
3.6 Categorical data 87
3.6.1 Generalized linear model 87
3.6.2 Perfect prediction♠ 89
3.6.3 Evaluation 90
3.7 Other data types 91
3.7.1 Count data 91
3.7.2 Semi-continuous data 92
3.7.3 Censored, truncated and rounded data 93
3.8 Nonignorable missing data 96
3.8.1 Overview 96
3.8.2 Selection model 97
3.8.3 Pattern-mixture model 98
3.8.4 Converting selection and pattern-mixture models . . . 99
3.8.5 Sensitivity analysis 100
3.8.6 Role of sensitivity analysis 101
3.8.7 Recent developments 102
3.9 Exercises 102

4 Multivariate missing data **105**
4.1 Missing data pattern 105
4.1.1 Overview 105
4.1.2 Summary statistics 107
4.1.3 Influx and outflux 109
4.2 Issues in multivariate imputation 111
4.3 Monotone data imputation 112
4.3.1 Overview 112
4.3.2 Algorithm 113
4.4 Joint modeling 115

4.4.1 Overview 115
4.4.2 Continuous data 115
4.4.3 Categorical data 117
4.5 Fully conditional specification 119
4.5.1 Overview 119
4.5.2 The MICE algorithm 120
4.5.3 Compatibility♠ 122
4.5.4 Congeniality or compatibility? 124
4.5.5 Model-based and data-based imputation 125
4.5.6 Number of iterations 126
4.5.7 Example of slow convergence 126
4.5.8 Performance 129
4.6 FCS and JM 130
4.6.1 Relations between FCS and JM 130
4.6.2 Comparisons 130
4.6.3 Illustration 131
4.7 MICE extensions 135
4.7.1 Skipping imputations and overimputation 135
4.7.2 Blocks of variables, hybrid imputation 135
4.7.3 Blocks of units, monotone blocks 136
4.7.4 Tile imputation 136
4.8 Conclusion 137
4.9 Exercises 137

5 Analysis of imputed data 139
5.1 Workflow 139
5.1.1 Recommended workflows 140
5.1.2 Not recommended workflow: Averaging the data . . . 142
5.1.3 Not recommended workflow: Stack imputed data . . . 144
5.1.4 Repeated analyses 144
5.2 Parameter pooling 145
5.2.1 Scalar inference of normal quantities 145
5.2.2 Scalar inference of non-normal quantities 146
5.3 Multi-parameter inference 147
5.3.1 D_1 Multivariate Wald test 147
5.3.2 D_2 Combining test statistics♠ 149
5.3.3 D_3 Likelihood ratio test♠ 150
5.3.4 D_1, D_2 or D_3? 152
5.4 Stepwise model selection 153
5.4.1 Variable selection techniques 153
5.4.2 Computation 154
5.4.3 Model optimism 155
5.5 Parallel computation 157
5.6 Conclusion 158
5.7 Exercises 158

II Advanced techniques 161

6 Imputation in practice 163
6.1 Overview of modeling choices 163
6.2 Ignorable or nonignorable? 165
6.3 Model form and predictors 166
6.3.1 Model form 166
6.3.2 Predictors 167
6.4 Derived variables 170
6.4.1 Ratio of two variables 170
6.4.2 Interaction terms 175
6.4.3 Quadratic relations♠ 176
6.4.4 Compositional data♠ 177
6.4.5 Sum scores 181
6.4.6 Conditional imputation 182
6.5 Algorithmic options 184
6.5.1 Visit sequence 184
6.5.2 Convergence 187
6.6 Diagnostics 189
6.6.1 Model fit versus distributional discrepancy 190
6.6.2 Diagnostic graphs 190
6.7 Conclusion 194
6.8 Exercises 195

7 Multilevel multiple imputation 197
7.1 Introduction 197
7.2 Notation for multilevel models 197
7.3 Missing values in multilevel data 200
7.3.1 Practical issues in multilevel imputation 201
7.3.2 Ad-hoc solutions for multilevel data 202
7.3.3 Likelihood solutions 203
7.4 Multilevel imputation by joint modeling 204
7.5 Multilevel imputation by fully conditional specification . . . 205
7.5.1 Add cluster means of predictors 206
7.5.2 Model cluster heterogeneity 207
7.6 Continuous outcome 207
7.6.1 General principle 208
7.6.2 Methods 209
7.6.3 Example 209
7.7 Discrete outcome 214
7.7.1 Methods 214
7.7.2 Example 215
7.8 Imputation of level-2 variable 218
7.9 Comparative work 219
7.10 Guidelines and advice 220

7.10.1 Intercept-only model, missing outcomes 222
7.10.2 Random intercepts, missing level-1 predictor 222
7.10.3 Random intercepts, contextual model 224
7.10.4 Random intercepts, missing level-2 predictor 226
7.10.5 Random intercepts, interactions 228
7.10.6 Random slopes, missing outcomes and predictors . . . 232
7.10.7 Random slopes, interactions 234
7.10.8 Recipes 238
7.11 Future research 240

8 Individual causal effects 241
8.1 Need for individual causal effects 241
8.2 Problem of causal inference 243
8.3 Framework 245
8.4 Generating imputations by FCS 246
8.4.1 Naive FCS 246
8.4.2 FCS with a prior for ρ 247
8.4.3 Extensions 253
8.5 Bibliographic notes 254

III Case studies 257

9 Measurement issues 259
9.1 Too many columns 259
9.1.1 Scientific question 260
9.1.2 Leiden 85+ Cohort 260
9.1.3 Data exploration 261
9.1.4 Outflux 263
9.1.5 Finding problems: `loggedEvents` 265
9.1.6 Quick predictor selection: `quickpred` 267
9.1.7 Generating the imputations 268
9.1.8 A further improvement: Survival as predictor variable 270
9.1.9 Some guidance 270
9.2 Sensitivity analysis 271
9.2.1 Causes and consequences of missing data 272
9.2.2 Scenarios 274
9.2.3 Generating imputations under the δ-adjustment . . . 274
9.2.4 Complete-data model 275
9.2.5 Conclusion 277
9.3 Correct prevalence estimates from self-reported data 277
9.3.1 Description of the problem 277
9.3.2 Don't count on predictions 278
9.3.3 The main idea 280
9.3.4 Data 281
9.3.5 Application 281

9.3.6 Conclusion 283

9.4 Enhancing comparability 283

9.4.1 Description of the problem 283

9.4.2 Full dependence: Simple equating 284

9.4.3 Independence: Imputation without a bridge study . . 286

9.4.4 Fully dependent or independent? 288

9.4.5 Imputation using a bridge study 289

9.4.6 Interpretation 292

9.4.7 Conclusion 293

9.5 Exercises 294

10 Selection issues **295**

10.1 Correcting for selective drop-out 295

10.1.1 POPS study: 19 years follow-up 295

10.1.2 Characterization of the drop-out 296

10.1.3 Imputation model 296

10.1.4 A solution “that does not look good” 299

10.1.5 Results 301

10.1.6 Conclusion 302

10.2 Correcting for nonresponse 302

10.2.1 Fifth Dutch Growth Study 303

10.2.2 Nonresponse 303

10.2.3 Comparison to known population totals 304

10.2.4 Augmenting the sample 304

10.2.5 Imputation model 306

10.2.6 Influence of nonresponse on final height 307

10.2.7 Discussion 308

10.3 Exercises 309

11 Longitudinal data **311**

11.1 Long and wide format 311

11.2 SE Fireworks Disaster Study 313

11.2.1 Intention to treat 314

11.2.2 Imputation model 315

11.2.3 Inspecting imputations 317

11.2.4 Complete-data model 318

11.2.5 Results from the complete-data model 319

11.3 Time raster imputation 320

11.3.1 Change score 321

11.3.2 Scientific question: Critical periods 322

11.3.3 Broken stick model♠ 324

11.3.4 Terneuzen Birth Cohort 326

11.3.5 Shrinkage and the change score♠ 328

11.3.6 Imputation 328

11.3.7 Complete-data model 330

11.4 Conclusion 332
11.5 Exercises 334

IV Extensions 337

12 Conclusion 339
12.1 Some dangers, some do's and some don'ts 339
12.1.1 Some dangers 339
12.1.2 Some do's 340
12.1.3 Some don'ts 341
12.2 Reporting 342
12.2.1 Reporting guidelines 343
12.2.2 Template 344
12.3 Other applications 345
12.3.1 Synthetic datasets for data protection 345
12.3.2 Analysis of coarsened data 345
12.3.3 File matching of multiple datasets 346
12.3.4 Planned missing data for efficient designs 346
12.3.5 Adjusting for verification bias 347
12.4 Future developments 347
12.4.1 Derived variables 347
12.4.2 Algorithms for blocks and batches 347
12.4.3 Nested imputation 348
12.4.4 Better trials with dynamic treatment regimes 348
12.4.5 Distribution-free pooling rules 348
12.4.6 Improved diagnostic techniques 349
12.4.7 Building block in modular statistics 349
12.5 Exercises 349

References 351

Author index 393

Subject index 405