Contents

Preface xix

Editors xxi

List of Contributors xxiii

I Preliminaries 1

1 Introduction and Preliminaries 3

Garrett M. Fitzmaurice, Michael G. Kenward, Geert Molenberghs, Geert Verbeke, and Anastasios A. Tsiatis

1.1 Introduction . . . 3
1.2 Notation and Terminology . . . 5
1.3 Missing Data Mechanisms . . . 7
1.3.1 Missing completely at random (MCAR) . . . 7
1.3.2 Missing at random (MAR) . . . 8
1.3.3 Not missing at random (NMAR) . . . 10
1.4 Joint Models for Non-Ignorable Missingness . . . 11
1.4.1 Selection models . . . 12
1.4.2 Pattern-mixture models . . . 13
1.4.3 Contrasting selection and pattern-mixture models . . . 13
1.5 Methods for Handling Missing Data . . . 15
1.5.1 Likelihood and Bayesian methods . . . 15
1.5.2 Weighting methods . . . 16
1.5.3 Multiple imputation methods . . . 18
1.6 Concluding Remarks . . . 18

2 Developments of Methods and Critique of ad hoc Methods 23

James R. Carpenter and Michael G. Kenward

2.1 Introduction . . . 23
2.2 Stream 1: Computational . . . 25
2.2.1 Direct likelihood . . . 26
2.3 Stream 2: Inferential . . . 28

Handbook of missing data methodology
2015

digitalisiert durch:
IDS Basel Bern

2.3.1 Rubin's classes 28
2.3.2 Missing at random and the conditional predictive distribution 29
2.3.3 Not missing at random and selection models 30
2.4 Stream 3: Semi-Parametric 31
2.5 Critique of *Ad Hoc* Methods 33
2.5.1 The Isolde trial 34
2.5.2 Complete records analysis 34
2.5.3 Marginal and conditional mean imputation 35
2.5.4 Last observation carried forward 38
2.6 Discussion 41

II Likelihood and Bayesian Methods 47

3 Introduction and Overview 49
Michael G. Kenward, Geert Molenberghs, and Geert Verbeke
3.1 Likelihood and Bayesian Inference and Ignorability 49
3.2 Joint Models 51

4 Perspective and Historical Overview 53
Michael G. Kenward and Geert Molenberghs
4.1 Introduction 53
4.2 Selection Models 55
4.2.1 The Heckman model 55
4.2.2 Models for categorical data 56
4.2.3 Models for continuous data 61
4.3 Pattern-Mixture Models 68
4.3.1 Introduction 68
4.3.2 Identifying restrictions 71
4.3.3 An example 74
4.4 Shared-Parameter Models 77
4.4.1 The simple shared-parameter framework 77
4.4.2 The generalized shared-parameter model 79
4.4.3 An example: Congestive heart failure 81
4.5 Concluding Remarks 84

5 Bayesian Methods 91
Michael J. Daniels and Joseph W. Hogan
5.1 Introduction 91
5.2 Primer on Bayesian Inference and Its Importance in Incomplete Data . . . 93

5.2.1 Complete data concepts and prior specification 93
5.2.2 The prior distribution and incomplete data 94
5.3 Importance of Correct Model Specification 95
5.4 Multiple Imputation and Bayesian Ignorability with and without Auxiliary Covariates 97
5.4.1 Bayesian ignorability 97
5.4.2 Connection to multiple imputation 98
5.4.3 Bayesian ignorability with auxiliary covariates 98
5.4.4 Fully Bayesian modeling with auxiliary covariates 99
5.5 Bayesian Non-Ignorability 100
5.5.1 Why mixture models? 101
5.5.2 Identifying restrictions 101
5.5.3 Identification by modeling assumptions 103
5.5.4 Some challenges for mixture models 105
5.6 General Computational Issues with the Posterior Distribution 106
5.7 Non-Parametric Bayesian Inference 109
5.8 Concluding Remarks and Future Directions 110

6 Joint Modeling of Longitudinal and Time-to-Event Data 117
Dimitris Rizopoulos
6.1 Introduction 117
6.2 Joint Modeling Framework 118
6.2.1 Estimation 121
6.3 Missing Data Mechanism 123
6.4 Parameterizations for the Association Structure 124
6.4.1 Interaction and lagged effects 125
6.4.2 Time-dependent slopes parameterization 126
6.4.3 Cumulative-effects parameterization 126
6.4.4 Random-effects parameterization 127
6.5 Analysis of the AIDS Data 128
6.6 Discussion and Extensions 131

III Semi-Parametric Methods 137

7 Introduction and Overview 139
Garrett M. Fitzmaurice
7.1 Introduction 139
7.2 Inverse Probability Weighting 140
7.3 Double Robustness 141

7.4 Pseudo-Likelihood 143
7.5 Concluding Remarks 145

8 Missing Data Methods: A Semi-Parametric Perspective 149

Anastasios A. Tsiatis and Marie Davidian

8.1 Introduction 149
8.2 Semi-Parametric Models and Influence Functions 151
8.3 Review of m-Estimators 153
8.4 Influence Functions and Semi-Parametric Theory 155
8.5 Missing Data and Inverse Probability Weighting of Complete Cases 160
8.6 Density and Likelihood of Missing Data 165
8.7 Geometry of Semi-Parametric Missing Data Models 166
8.7.1 Observed-data nuisance tangent space 166
8.7.2 Orthogonal complement of the nuisance tangent space 169
8.7.3 Augmentation space Λ_2 with two levels of missingness 173
8.7.4 Augmentation space Λ_2 with monotone missingness 173
8.8 Optimal Observed-Data Estimating Function Associated with Full-Data Estimating Function 174
8.9 Estimating the Missing Data Process 179
8.10 Summary and Concluding Remarks 182

9 Double-Robust Methods 185

Andrea Rotnitzky and Stijn Vansteelandt

9.1 Introduction 185
9.2 Data Configuration, Goal of Inference, and Identifying Assumptions . . . 187
9.3 Dimension Reducing Strategies 188
9.3.1 Regression imputation 188
9.3.2 Inverse probability weighting 189
9.4 Double-Robust Estimation with Missing-at-Random Data 191
9.4.1 Double-robust regression imputation estimators 193
9.4.2 Double-robust sample bounded IPW estimators 195
9.4.3 Double-robust estimators that are never less efficient than IPW estimators 197
9.4.4 Double-robust estimators that are efficient over a parametric class of IPWA estimators 199
9.4.5 Double-robust sample bounded IPW estimators with enhanced efficiency under the missingness model 200
9.4.6 Double-robust regression imputation estimators with enhanced efficiency under the missingness model 201
9.5 Double-Robust Estimation in Longitudinal Studies with Attrition 202

9.6 Discussion 205
9.6.1 Double-robustness in other data structures 206
9.6.2 Model selection 207

10 Pseudo-Likelihood Methods for Incomplete Data **213**
Geert Molenberghs and Michael G. Kenward
10.1 Introduction 213
10.2 Notation and Concepts 214
10.3 Pseudo-Likelihood 215
10.3.1 Definition and properties 215
10.3.2 Pairwise pseudo-likelihood 216
10.3.3 Full conditional pseudo-likelihood 216
10.4 Pseudo-Likelihood for Incomplete Data 217
10.4.1 Estimating equations for pairwise likelihood 217
10.4.2 Precision estimation 219
10.4.3 Marginal pseudo-likelihood for binary data 220
10.4.4 The multivariate normal model 222
10.5 Case Studies 224
10.5.1 A clinical trial in onychomycosis 224
10.5.2 Analgesic trial 226
10.6 Concluding Remarks 228

IV Multiple Imputation **233**

11 Introduction **235**
Michael G. Kenward
11.1 Introduction 235

12 Multiple Imputation: Perspective and Historical Overview **239**
John B. Carlin
12.1 What Is Multiple Imputation? 239
12.1.1 Introduction 239
12.1.2 A brief overview 240
12.1.3 Illustrative example 242
12.1.4 Early history 246
12.2 Foundations and Basic Theory 247
12.2.1 Multiple imputation as approximate Bayesian inference 248
12.2.2 Proper imputation and repeated sampling properties 250
12.2.3 The role of the missing data model 251

12.2.4 Further technical aspects of multiple imputation inference 251
12.2.5 How many imputations? 253
12.3 Imputation Methods 254
12.3.1 Regression-based imputation 254
12.3.2 Imputation under a joint model 256
12.3.3 Imputation using fully conditional specification 257
12.3.4 General principles for building imputation models 258
12.4 Multiple Imputation in Practice 258
12.4.1 Early applications: Expert imputation of large-scale survey data and other tailored applications 259
12.4.2 Multiple imputation enters the toolkit for practical data analysis . . 260
12.4.3 New applications and outstanding problems 261

13 Fully Conditional Specification **267**

Stef van Buuren

13.1 Introduction 267
13.1.1 Overview 267
13.1.2 Notation 268
13.2 Practical Problems in Multivariate Imputation 268
13.3 Missing Data Patterns 270
13.3.1 Overview 270
13.3.2 Ways to quantify the linkage pattern 271
13.4 Multivariate Imputation Models 273
13.4.1 Overview 273
13.4.2 Imputation of monotone missing data 274
13.4.3 Imputation by joint modeling 274
13.5 Fully Conditional Specification (FCS) 275
13.5.1 Overview 275
13.5.2 Chained equations: The MICE algorithm 276
13.5.3 Properties 276
13.5.4 Compatibility 277
13.5.5 Number of iterations 279
13.5.6 Performance 279
13.6 Modeling in FCS 280
13.6.1 Overview 280
13.6.2 MAR or NMAR? 281
13.6.3 Model form 282
13.6.4 Predictors 283

13.6.5 Derived variables 285
13.6.6 Visit sequence 287
13.7 Diagnostics 288
13.8 Conclusion 289

14 Multilevel Multiple Imputation **295**
Harvey Goldstein and James R. Carpenter
14.1 Introduction 295
14.2 Multilevel MI with Mixed Response Types 296
14.2.1 Congeniality 298
14.2.2 Conditioning on fully observed variables in imputation 299
14.2.3 Choice of auxiliary variables 299
14.3 Imputing Mixed Response Types 299
14.3.1 Some comments 302
14.4 Interactions and General Functions of Covariates 303
14.5 Coarsened Data 305
14.6 Record Linkage 307
14.7 Applications 309
14.7.1 Modelling class size data with missing values 309
14.7.2 Individual patient data meta-analysis 311
14.8 Conclusions and Software 313

V Sensitivity Analysis **317**

15 Introduction and Overview **319**
Geert Molenberghs, Geert Verbeke, and Michael G. Kenward
15.1 Sensitivity 319
15.2 Sensitivity Analysis 320
15.3 Sensitivity Analysis for Parametric Models 320
15.4 Sensitivity Analysis in a Semi-Parametric Setting 322

16 A Likelihood-Based Perspective **325**
Geert Verbeke, Geert Molenberghs, and Michael G. Kenward
16.1 Introduction 326
16.2 Motivating Examples 327
16.2.1 The orthodontic growth data 327
16.2.2 The Slovenian Public Opinion Survey 328
16.2.3 The rat data 331
16.2.4 A clinical trial in onychomycosis 332

16.2.5 A depression trial 333
16.3 Notation and Concepts 333
16.4 What Is Different When Data Are Incomplete? 334
16.4.1 Problems with model selection and assessment with incomplete data 334
16.4.2 The BRD family of models 335
16.4.3 Model selection and assessment with incomplete data 337
16.4.4 Model assessment for the orthodontic growth data 339
16.4.5 Model assessment for the Slovenian Public Opinion Survey 340
16.5 Interval of Ignorance, Global Influence, and Local Influence 343
16.5.1 Interval of ignorance 344
16.5.2 Global influence 345
16.5.3 Local influence 346
16.5.4 How sensitive is the proportion of "Yes" voters? 350
16.5.5 Local influence for Gaussian data 352
16.5.6 Analysis and sensitivity analysis of the rat data 354
16.5.7 Local influence methods and their behavior 355
16.6 A Sensitivity Analysis for Shared-Parameter Models 357
16.6.1 An extended shared-parameter model 358
16.6.2 SPM models for the onychomycosis data 359
16.6.3 A sensitivity analysis in the shared-parameter framework 362
16.7 A Latent-Class Mixture Model for Incomplete Longitudinal Gaussian Data 364
16.7.1 Classification 366
16.7.2 Analysis of the depression trial 367
16.7.3 A sensitivity analysis for the depression trial 370
16.8 Further Methods 370
16.9 Concluding Remarks 372

17 A Semi-Parametric Perspective **379**
Stijn Vansteelandt
17.1 Why Semi-Parametric Sensitivity Analyses? 379
17.2 Non-Parametric Identification 381
17.2.1 Why non-parametric identification? 383
17.2.2 The curse of dimensionality 384
17.3 Case Study on Non-Monotone Missing Follow-Up Data 386
17.3.1 International Breast Cancer Study Group, Trial VI 386
17.3.2 Missing data assumptions 387
17.3.3 Estimation 390
17.3.4 Sensitivity analysis with a scalar sensitivity parameter 393

17.3.5 Summarizing the sensitivity analysis results 394
17.3.6 Sensitivity analysis with a vector of sensitivity parameters 396
17.4 Discussion 398

18 Bayesian Sensitivity Analysis **405**
Joseph W. Hogan, Michael J. Daniels, and Liangyuan Hu
18.1 Introduction 405
18.2 Notation and Definitions 406
18.2.1 Outcome variables, covariates, and response indicators 406
18.2.2 Distributions and models 407
18.2.3 Missing data mechanisms 408
18.3 Inference from Under-Identified Models 409
18.3.1 Overview 409
18.3.2 Assessing veracity of modeling assumptions 410
18.3.3 Non-parametric (non-)identifiability 411
18.3.4 Parametric identification 412
18.3.5 Summary 413
18.4 Sensitivity Analysis from a Bayesian Perspective 414
18.4.1 Likelihood parameterization 414
18.4.2 Sensitivity parameters 415
18.4.3 Priors 417
18.4.4 Summary 418
18.5 Empirical Illustrations 419
18.5.1 Overview 419
18.5.2 Inference about proportions 419
18.5.3 Inference about continuous bivariate distribution 427
18.5.4 Summary 431
18.6 Summary and Discussion 431

19 Sensitivity Analysis with Multiple Imputation **435**
James R. Carpenter and Michael G. Kenward
19.1 Introduction 435
19.2 Review of NMAR Modelling 437
19.3 Pattern-Mixture Modelling with Multiple Imputation 438
19.3.1 Modifying the MAR imputation distribution 439
19.3.2 Missing covariates 442
19.3.3 Application to survival analysis 444
19.4 Longitudinal Clinical Trial Data 447

19.4.1 Estimands . . . 447
19.4.2 Deviations . . . 449
19.4.3 Change in slope post-deviation: The 'Δ-method' . . . 449
19.4.4 Example: Asthma Study . . . 450
19.4.5 Reference-based imputation . . . 452
19.4.6 Constructing joint distributions of pre- and post-deviation data . . . 453
19.4.7 Technical details . . . 454
19.4.8 Example: Asthma Study . . . 456
19.4.9 Distinguishing two types of sensitivity analysis . . . 459
19.5 Approximating a Selection Model by Importance Weighting . . . 460
19.5.1 Algorithm for approximate sensitivity analysis by re-weighting . . . 462
19.5.2 Reliability of the approximation . . . 462
19.5.3 Further developments . . . 465
19.6 Discussion . . . 466

20 The Elicitation and Use of Expert Opinion **471**

Ian R. White

20.1 Introduction . . . 471
20.2 Background on Elicitation . . . 472
20.3 How to Parameterise a Model to Elicit Expert Opinion . . . 473
20.3.1 Pattern-mixture model . . . 473
20.3.2 Prior . . . 474
20.3.3 Estimation: Sensitivity analysis . . . 474
20.3.4 Estimation: Bayesian analysis . . . 475
20.4 Eliciting Expert Opinion about a Single Sensitivity Parameter . . . 475
20.4.1 The peer review trial . . . 475
20.4.2 Elicitation . . . 477
20.4.3 Analysis . . . 477
20.5 A Spreadsheet to Elicit Expert Opinion about Two Sensitivity Parameters 478
20.5.1 The Down Your Drink (DYD) Trial . . . 478
20.5.2 Spreadsheet for each arm . . . 479
20.5.3 Spreadsheet for the correlation . . . 480
20.5.4 Use of the elicited values . . . 481
20.6 Practicalities . . . 481
20.6.1 Choice of experts . . . 481
20.6.2 Format of elicitation . . . 482
20.6.3 Feedback . . . 482
20.6.4 Pooling experts . . . 484

20.6.5 When to elicit 484
20.7 Discussion 485

VI Special Topics **491**

21 Introduction and Overview **493**
Geert Molenberghs
21.1 Introduction 493
21.2 Missing Data in Clinical Trials 493
21.3 Missing Data in Surveys 495
21.4 Model Diagnostics 495

22 Missing Data in Clinical Trials **499**
Craig Mallinckrodt
22.1 Introduction 499
22.1.1 Clinical trials 499
22.1.2 History 501
22.1.3 National Research Council recommendations 502
22.2 Preventing Missing Data 502
22.3 Estimands 504
22.4 Analyses 506
22.4.1 Primary analysis 506
22.4.2 Model diagnostics 507
22.4.3 Sensitivity analyses 508
22.5 Example 512
22.5.1 Data 512
22.5.2 Primary analyses 512
22.5.3 Model diagnostics 513
22.5.4 Sensitivity analyses 515
22.6 Discussion 521

23 Missing Data in Sample Surveys **525**
Thomas R. Belin and Juwon Song
23.1 Design-Based versus Model-Based Inference 526
23.2 Design-Based versus Model-Based Perspectives on Missing-Data Uncertainty in Estimation of Census Undercount Rates 526
23.3 Weighting and Imputation as Strategies for Unit and Item Nonresponse . . 528
23.4 Strategies for Producing Imputations 529
23.5 Imputation Based on an Explicit Bayesian Model 529

23.6 Hot-Deck Imputation 531
23.6.1 Hot-deck imputation within adjustment cells 531
23.6.2 Hot-deck imputation using distance metrics 532
23.6.3 Maintaining relationships between variables in multivariate data . . 533
23.7 Sampling Weights 534
23.8 Multiple Imputation 535
23.9 When the Imputer's Model Differs from the Analyst's Model 536
23.9.1 When the analyst considers fewer variables than the imputer 536
23.9.2 When the imputer assumes more than the analyst 536
23.10 Variance Estimation with Imputed Data 537
23.10.1 Applying explicit variance formulae 537
23.10.2 Resampling methods 538
23.10.3 Multiple imputation 539
23.11 Discussion 539

24 Model Diagnostics **547**

Dimitris Rizopoulos, Geert Molenberghs, and Geert Verbeke

24.1 Introduction 547
24.2 Multivariate Models 548
24.2.1 Marginal and mixed-effects models 548
24.2.2 Analysis of the AIDS and PBC datasets 550
24.3 Residuals for Mixed-Effects and Marginal Models 551
24.3.1 Definitions 551
24.3.2 Residuals for the AIDS and PBC datasets 552
24.3.3 Dropout and residuals 552
24.4 Multiple Imputation Residuals 555
24.4.1 Fixed visit times 555
24.4.2 Random visit times 557
24.5 Discussion and Extensions 561

Index **565**