

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



LSHTM Research Online

Daniel, Rm; (2009) On aspects of robustness and sensitivity in missing data methods. PhD thesis, London School of Hygiene & Tropical Medicine. DOI: <https://doi.org/10.17037/PUBS.04646537>

Downloaded from: <https://researchonline.lshtm.ac.uk/id/eprint/4646537/>

DOI: <https://doi.org/10.17037/PUBS.04646537>

Usage Guidelines:

Please refer to usage guidelines at <https://researchonline.lshtm.ac.uk/policies.html> or alternatively contact researchonline@lshtm.ac.uk.

Available under license. To note, 3rd party material is not necessarily covered under this license: <http://creativecommons.org/licenses/by-nc-nd/3.0/>

<https://researchonline.lshtm.ac.uk>

On aspects of robustness and sensitivity in missing data methods

Rhian Mair Daniel



A thesis submitted to the University of London for the degree of
Doctor of Philosophy

London School of Hygiene and Tropical Medicine, 2009

To my parents, John & Judith

Declaration

Statement of Own Work

All students are required to complete the following declaration when submitting their thesis. A shortened version of the School's definition of Plagiarism and Cheating is as follows (the full definition is given in the Research Degrees Handbook):

The following definition of plagiarism will be used:

Plagiarism is the act of presenting the ideas or discoveries of another as ones own. To copy sentences, phrases or even striking expressions without acknowledgement in a manner which may deceive the reader as to the source is plagiarism. Where such copying or close paraphrase has occurred the mere mention of the source in a biography will not be deemed sufficient acknowledgement; in each instance, it must be referred specifically to its source. Verbatim quotations must be directly acknowledged, either in inverted commas or by indenting. (University of Kent)

Plagiarism may include collusion with another student, or the unacknowledged use of a fellow student's work with or without their knowledge and consent. Similarly, the direct copying by students of their own original writings qualifies as plagiarism if the fact that the work has been or is to be presented elsewhere is not clearly stated.

Cheating is similar to plagiarism, but more serious. Cheating means submitting another student's work, knowledge or ideas, while pretending that they are your own, for formal assessment or evaluation.

Supervisors should be consulted if there are any doubts about what is permissible.

Declaration by candidate

I have read and understood the School's definition of plagiarism and cheating given in the Research Degrees Handbook. I declare that this thesis is my own work, and that I have acknowledged all results and quotations from the published or unpublished work of other people.

Signed:

Date: **August 13, 2009**

Full name: **Rhian Mair Daniel**

Abstract

Missing data are common wherever statistical methods are applied in practice. They present a problem by demanding that additional untestable assumptions be made about the mechanism leading to the incompleteness of the data. Minimising the strength of these assumptions and assessing the sensitivity of conclusions to their possible violation constitute two important aspects of current research in this area.

One attractive approach is the doubly robust (DR) weighting-based method proposed by Robins and colleagues. By incorporating two models for the missing data process, inferences are valid when at least one model is correctly specified. The balance between robustness, efficiency and analytical complexity is one which is difficult to strike, resulting in a split between the likelihood and multiple imputation (MI) school on one hand and the weighting and DR school on the other.

We propose a new method, doubly robust multiple imputation (DRMI), combining the convenience of MI with the robustness of the DR approach, and explore the use of our new estimator for non-monotone missing at random data, a setting in which, hitherto, estimators with the DR property have not been implemented. We apply the method to data from a clinical trial comparing type II diabetes drugs, where we also use MI as a tool to explore sensitivity to the missing at random assumption. Finally, we study DRMI in the longitudinal binary data setting and find that it compares favourably with existing methods.

Acknowledgements

I am extremely indebted to my supervisor, Prof. Mike Kenward, for his kindness, wisdom and inspiration over the last three years. I simply could not have wished for a better supervisor.

My sincere gratitude also goes to GlaxoSmithKline Pharmaceuticals and Dr. Paula Curtis for the access to the RECORD study dataset, and to Prof. Stuart Pocock and Dr. Duolao Wang for their guidance on this work.

I would like to thank the staff at the Medical Statistics Unit for always being generous with their expertise, and the Medical Research Council for providing my studentship.

I would also like to thank my parents, to whom this thesis is dedicated, not only for their unconditional love, but for their contagious enthusiasm for mathematics, without which I might never have found my way into the fascinating world of statistics.

The past three years would not have been nearly as enjoyable without the company of great friends and officemates. Beatriz, Sara, Richard, Fizz, Lesego, Chien-Mu, Louise and Val, and especially Jonathan for so many useful discussions: thank you to you all!

The final thank you is to my fiancé, Chris, for his love and encouragement, and—most of all—for always showing such an interest. I hope I can now do the same for him.

Acronyms and Abbreviations

AIPW Augmented Inverse Probability Weighted

ANCOVA ANalysis of COVAriance

ANOVA ANalysis Of VAriance

BR Bang and Robins

CAR Coarsened At Random

CC Complete Case

CCAR Coarsened Completely At Random

CNAR Coarsened Not At Random

CV CardioVascular

CWIEE Cluster-level Weighted Independence Estimating Equation

CWGEE Cluster-level Weighted Generalised Estimating Equation

DR Doubly Robust

DRMI Doubly Robust Multiple Imputation

EM Expectation Maximisation

GEE Generalised Estimating Equation

GLM Generalised Linear Model

GLMM Generalised Linear Mixed Model

HGLM Hierarchical Generalised Linear Model

IEE Independence Estimating Equation

IPWCC Inverse Probability Weighted Complete Case

ITT Intent To Treat

LHS Left Hand Side

LOCF Last Observation Carried Forward

MAR Missing At Random

MCAR Missing Completely At Random

MCMC Markov Chain Monte Carlo

Met Metformin

MI Multiple Imputation

MICE Multiple Imputation with Chained Equations

ML Maximum Likelihood

MLM Multivariate Linear Model

MNAR Missing Not At Random

MRMM Markov Randomised Monotone Missingness

NAR Non-compliance At Random

NCAR Non-compliance Completely At Random

NNAR Non-compliance Not At Random

OR Outcome Regression

OWIEE Observation-level Weighted Independence Estimating Equation

OWGEE Observation-level Weighted Generalised Estimating Equation

PP Per Protocol

RAL Regular Asymptotically Linear

RECORD Rosiglitazone Evaluated for Cardiac Outcomes and Regulation of
glycaemia in Diabetes

REML REstricted Maximum Likelihood

RHS Right Hand Side

RMM Randomised Monotone Missingness

Rosi Rosiglitazone

RR Robins and Rotnitzky

SE Standard Error

Su Sulfonylurea

Table of contents

DECLARATION	3
ABSTRACT	5
ACKNOWLEDGEMENTS	6
ACRONYMS AND ABBREVIATIONS	7
I Preliminaries	26
1 Introduction	27
1.1 Background	27
1.2 Outline of this thesis	30
2 The RECORD study: background	33
II Literature review & theoretical foundations	36
3 Notation and basic definitions	37
3.1 Full data and associated quantities	37
3.2 Coarsened data	40
3.3 Missing data as a special case of coarsening	41
3.4 Observed data quantities	42
3.5 Missingness indicators	43

3.6	Monotonicity and dropout	44
3.7	Semiparametric model	45
4	Missing data modelling frameworks and mechanisms	47
4.1	Modelling frameworks	47
4.2	Missing and coarsened data mechanisms	48
4.2.1	Randomised monotone missingness (RMM) processes	53
5	Simple methods	57
5.1	Complete case (CC) analysis	58
5.2	Inverse probability weighted complete case (IPWCC) analysis	59
5.3	Single imputation methods	60
6	Fully-parametric methods	65
6.1	The direct likelihood approach	66
6.2	Expectation-Maximisation algorithm	70
6.3	The linear mixed model and the multivariate normal linear model	71
6.4	The generalised linear mixed model	72
6.5	Hierarchical generalised linear models	74
6.6	Bayesian methods	74
6.7	Multiple imputation	75
6.7.1	Motivation	75
6.7.2	A formal description	76
6.7.3	Improper vs. proper imputation	77
6.7.4	Multiple imputation using chained equations (MICE)	82
7	Semiparametric methods	84
7.1	Introduction	84
7.2	Mean score method: an example of a method belonging to \mathcal{I}	86
7.3	Generalised estimating equations (GEE): an example of a method belonging to \mathcal{R}	88
7.3.1	Working correlation structure	90
7.3.2	Algorithm for fitting GEE	92
7.3.3	Estimating precision using the sandwich estimator	93

7.4	Weighted GEE	93
7.4.1	High variability in the weights	96
7.5	MI-GEE	97
7.6	Improved efficiency and double robustness	98
7.6.1	Augmented inverse probability weighted (AIPW) estimator (\mathcal{I} -type)	98
7.6.2	Double robustness	100
7.6.3	AIPW estimator (\mathcal{R} -type)	103
7.6.4	Regression formulation from Bang and Robins	105
7.6.4.1	Cross-sectional univariate missing data	105
7.6.4.2	Longitudinal data with monotone missingness	108
7.6.5	A semiparametric-efficient GEE-type estimator	110
III	Multiple imputation for doubly robust estimation	114
8	Doubly robust multiple imputation	115
8.1	Motivation	115
8.2	Description of the method	116
8.3	Variance estimation	120
8.4	Discussion	124
9	Robust multiple imputation: an alternative formulation	126
9.1	Introduction	126
9.2	The proposed method	127
9.2.1	Univariate ignorable missing data	127
9.2.2	Longitudinal ignorable missing data	131
9.2.2.1	Monotone longitudinal data	132
9.2.2.2	Non-monotone longitudinal data	138
9.2.3	Non-monotone cross-sectional ignorable missing data	142
9.2.4	A closer look at Bang and Robins for longitudinal data	143
9.3	Simulation studies	145
9.3.1	Univariate ignorable missing data	145
9.3.2	Longitudinal monotone ignorable missing data	146

9.3.3	Longitudinal non-monotone ignorable missing data	149
9.3.4	Cross-sectional non-monotone ignorable missing data	153
9.4	Discussion	155
IV	The RECORD study	158
10	Doubly-robust MAR analysis	159
10.1	Introduction	159
10.2	Methods	160
10.3	Results and conclusions	161
11	MNAR sensitivity analyses	166
11.1	Aims and outline	166
11.2	Patterns of missing data and non-compliance in the 18-month RECORD data	167
11.3	What are the questions and how can we answer them?	168
11.3.1	Treatment vs. assignment to treatment	168
11.3.2	Populations	169
11.3.3	Objectives of the RECORD 18-month analysis	170
11.3.3.1	Missing data	171
11.3.3.2	Non-compliance	173
11.3.4	ITT analyses with missing data	175
11.3.4.1	Background	175
11.3.4.2	Multiple imputation and intent-to-treat	176
11.4	Sensitivity analyses	177
11.4.1	Strategy for MNAR ITT analyses on the 18-month RECORD data	178
11.4.2	Strategy for MNAR/NNAR PP analyses on the 18-month RECORD data	180
11.4.3	Results	183
11.4.3.1	ITT	183
11.4.3.2	Per protocol	183
11.4.4	Discussion	184

V Comparing methods for incomplete longitudinal binary data	187
12 Motivation and simulation studies	188
12.1 Introduction	188
12.2 Simulation study	189
12.2.1 Methods	189
12.2.2 Discussion of results	192
12.2.2.1 Convergence problems with the Robins and Rotnitzky (1995) estimator	193
12.2.2.2 Superiority of doubly robust MI over the other doubly robust procedures	194
12.2.2.3 Lack of bias in unweighted GEE	194
12.2.2.4 Differences between cluster- and observation-level weighting	194
12.2.2.5 Imputation versus weighting	195
12.2.2.6 Doubly robust methods with both models misspecified	195
13 Theoretical comparison of GEE and related methods	211
13.1 Aims and outline	211
13.2 Conditions under which observation-level weighted GEE, augmented observation-level weighted GEE and nonparametric mean quasi score imputation are numerically equivalent	212
13.3 Unweighted GEE: conditions for consistency and semiparametric efficiency	218
13.4 Cluster- versus observation-level weighting	221
13.5 MI-GEE and its relationship with observation-level weighted GEE . . .	233
13.6 A comparison of doubly robust MI and other doubly robust procedures	235
VI Discussion	238
14 Discussion	239
14.1 Main conclusions	239

14.2 Other conclusions	241
14.3 Future work	242
BIBLIOGRAPHY	245
Appendix	253
A Proofs omitted from the main text	254
A.1 Proof of Lemma 8.1	254
A.2 Proof of Lemma 8.2	258
A.3 Proof of Lemma 8.3	260
A.4 Proof of Lemma 8.5	260
B Further tables and figures	262
C Computer code	276
C.1 Robust multiple imputation: original formulation	276
C.1.1 Improper	276
C.1.2 Proper	284
C.2 Robust multiple imputation: alternative formulation	286
C.3 Binary simulation study: scenario 1	287

List of tables

- 9.1 The results of the first simulation study performed by Bang and Robins (2005) with doubly robust multiple imputation (DRMI) included in the comparison. No subscript indicates correct specification of the relevant model(s). π – false indicates that the estimator used an incorrectly-specified π -model, y – false indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y$ – false indicates that both the π - and y -models were incorrectly specified. 147
- 9.2 The results of the monotone longitudinal simulation study performed by Bang and Robins (2005) with doubly robust multiple imputation (DRMI) included in the comparison. No subscript indicates correct specification of the relevant model(s). π – false indicates that the estimator used an incorrectly-specified π -model, y – false indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y$ – false indicates that both the π - and y -models were incorrectly specified. 149
- 9.3 The results of the non-monotone longitudinal simulation study where doubly robust multiple imputation (DRMI) is compared with IPWCC and ordinary MI. No subscript indicates correct specification of the relevant model(s). π – false indicates that the estimator used an incorrectly-specified π -model, y – false indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y$ – false indicates that both the π - and y -models were incorrectly specified. 152

9.4 The results of the non-monotone cross-sectional simulation study where doubly robust multiple imputation (DRMI) is compared with IPWCC and ordinary MI. The known (true or $\sqrt{\text{true}}$) probability weights were used in the IPWCC and DRMI methods. No subscript indicates correct specification of the y -model and weights (where applicable). $\pi -$ false indicates that the square root of the weights were used, $y -$ false indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y -$ false indicates that both the weights and y -model were incorrect. 156

10.1 A comparison of the results (estimates and standard errors) from the linear mixed model and the doubly robust MI estimator. 161

12.1 The results of the first longitudinal binary simulation study, where the means model is saturated. In each case, β refers to the log odds ratio for X_1 at the third timepoint. The abbreviations used are: CWGEE (cluster-level weighted GEE), OWGEE (observation-level weighted GEE), RR (method proposed by Robins and Rotnitzky (1995)), BR (method proposed by Bang and Robins (2005)) and DRMI (doubly robust MI). No subscript indicates correct specification of the relevant model(s). $\pi -$ false indicates that the estimator used an incorrectly-specified π -model, $y -$ false indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y -$ false indicates that both the π - and y -models were incorrectly specified. 196

12.2 The results of the second longitudinal binary simulation study, where the means model is saturated but the correlation structure is different from the one used in the first set of simulations. In each case, β refers to the log odds ratio for X_1 at the third timepoint. The abbreviations used are: CWGEE (cluster-level weighted GEE), OWGEE (observation-level weighted GEE), RR (method proposed by Robins and Rotnitzky (1995)), BR (method proposed by Bang and Robins (2005)) and DRMI (doubly robust MI). No subscript indicates correct specification of the relevant model(s). $\pi -$ false indicates that the estimator used an incorrectly-specified π -model, $y -$ false indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y -$ false indicates that both the π - and y -models were incorrectly specified. 197

- 12.3 The results of the third longitudinal binary simulation study, where the means model is *not* saturated. In each case, β refers to the log odds ratio for X_1 at the third timepoint. The abbreviations used are: CWGEE (cluster-level weighted GEE), OWGEE (observation-level weighted GEE), RR (method proposed by Robins and Rotnitzky (1995)), BR (method proposed by Bang and Robins (2005)) and DRMI (doubly robust MI). No subscript indicates correct specification of the relevant model(s). $\pi -$ false indicates that the estimator used an incorrectly-specified π -model, $y -$ false indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y -$ false indicates that both the π - and y -models were incorrectly specified. 198
- B.1 Estimates and SEs of the treatment difference (change in HbA_{1c} from baseline to 18 months) between Met+Rosi and Met+Su, and Su+Rosi and Su+Met, respectively, for each of the models considered. 263

List of figures

2.1	A summary of the design of the RECORD trial	34
3.1	A visual depiction of coarsened data	41
3.2	A visual depiction of missing data as a special case of coarsening	42
3.3	A diagrammatic depiction of a monotone missing data pattern (the shaded areas are unobserved)	45
4.1	A Markov randomised monotone missingness process for $J = 3$. Dependence on \mathbf{X} is implicit.	53
4.2	A Markov randomised monotone missingness process for longitudinal data	55
5.1	The bias introduced by mean imputation on the outcome in a linear regression model with outcomes missing completely at random	61
8.1	A diagrammatic representation of the robust MI formulation	118
9.1	The MRMM longitudinal process used for the longitudinal non-monotone simulation study.	150
9.2	The MRMM longitudinal process used for the longitudinal non-monotone simulation study.	153
10.1	A histogram showing the distribution of the residuals from a linear regression of HbA _{1c} on treatment group and baseline HbA _{1c} for the observed data at the final timepoint.	160
10.2	A comparison of the HbA _{1c} profiles predicted from the direct likelihood analysis and those predicted from the doubly robust multiple imputation analysis for the Met+Su arm. The green and red error bars show \pm the standard errors for the likelihood and DRMI analyses respectively.	162

-
- 10.3 A comparison of the HbA_{1c} profiles predicted from the direct likelihood analysis and those predicted from the doubly robust multiple imputation analysis for the Met+Rosi arm. The green and red error bars show \pm the standard errors for the likelihood and DRMI analyses respectively. 162
- 10.4 A comparison of the HbA_{1c} profiles predicted from the direct likelihood analysis and those predicted from the doubly robust multiple imputation analysis for the Su+Met arm. The green and red error bars show \pm the standard errors for the likelihood and DRMI analyses respectively. 164
- 10.5 A comparison of the HbA_{1c} profiles predicted from the direct likelihood analysis and those predicted from the doubly robust multiple imputation analysis for the Su+Rosi arm. The green and red error bars show \pm the standard errors for the likelihood and DRMI analyses respectively. 164
- 10.6 The differences between the HbA_{1c} profiles for the Met+Rosi and Met+Su arms. The solid green and red lines show the predicted differences from the likelihood and DRMI analyses respectively, and the dotted lines show \pm the pointwise standard errors for these differences. 165
- 10.7 The differences between the HbA_{1c} profiles for the Su+Rosi and Su+Met arms. The solid green and red lines show the predicted differences from the likelihood and DRMI analyses respectively, and the dotted lines show \pm the pointwise standard errors for these differences. 165
- 12.1 Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the first set of simulations with both models correctly specified. 199
- 12.2 Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the first set of simulations with the π -model incorrectly specified. 199
- 12.3 Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the first set of simulations with the y -model incorrectly specified. 200
- 12.4 Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the first set of simulations with both models incorrectly specified. 200

12.5	Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the second set of simulations with both models correctly specified.	201
12.6	Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the second set of simulations with the π -model incorrectly specified.	201
12.7	Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the second set of simulations with the γ -model incorrectly specified.	202
12.8	Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the second set of simulations with both models incorrectly specified.	202
12.9	Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the third set of simulations with both models correctly specified.	203
12.10	Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the third set of simulations with the π -model incorrectly specified.	203
12.11	Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the third set of simulations with the γ -model incorrectly specified.	204
12.12	Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the third set of simulations with both models incorrectly specified.	204
12.13	Kernel density plots comparing unweighted and observation-level weighted GEE. These estimates are from the first set of simulations with both models correctly specified.	205
12.14	Kernel density plots comparing unweighted and observation-level weighted GEE. These estimates are from the second set of simulations with both models correctly specified.	205

12.15	Kernel density plots comparing unweighted and observation-level weighted GEE. These estimates are from the third set of simulations with both models correctly specified.	206
12.16	Kernel density plots comparing cluster- and observation-level weighted GEE. These estimates are from the first set of simulations with both models correctly specified.	206
12.17	Kernel density plots comparing cluster- and observation-level weighted GEE. These estimates are from the second set of simulations with both models correctly specified.	207
12.18	Kernel density plots comparing cluster- and observation-level weighted GEE. These estimates are from the third set of simulations with both models correctly specified.	207
12.19	Kernel density plots comparing MI-GEE and observation-level weighted GEE. These estimates are from the first set of simulations with both models correctly specified.	208
12.20	Kernel density plots comparing MI-GEE and observation-level weighted GEE. These estimates are from the second set of simulations with both models correctly specified.	208
12.21	Kernel density plots comparing MI-GEE and observation-level weighted GEE. These estimates are from the third set of simulations with both models correctly specified.	209
12.22	Kernel density plots comparing the three doubly robust procedures. These estimates are from the first set of simulations with both models correctly specified.	209
12.23	Kernel density plots comparing the three doubly robust procedures. These estimates are from the second set of simulations with both models correctly specified.	210
12.24	Kernel density plots comparing the three doubly robust procedures. These estimates are from the third set of simulations with both models correctly specified.	210

13.1	The variance of the weighted average of two second timepoint observations from a three timepoint CWIEE. The corresponding OWIEE variance occurs when $w = 1$	223
13.2	The 8 eigenvalues of (13.8) with 2 timepoints evaluated for 1,000 different datasets.	233
13.3	The 12 eigenvalues of (13.8) with 3 timepoints evaluated for 1,000 different datasets, with the 8 eigenvalues for the 2 timepoints case superimposed.	234
13.4	The 16 eigenvalues of (13.8) with 4 timepoints evaluated for 1,000 different datasets, with the 12 eigenvalues for the 3 timepoints case and the 8 eigenvalues for the 2 timepoints case superimposed.	235
13.5	A comparison of the cluster-level and observation-level weights for the first simulation in the first set.	236
13.6	A comparison of the cluster-level and observation-level weights for the first simulation in the third set.	237
B.1	The profiles (mean \pm SE) implied by the MAR per protocol analysis	264
B.2	The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (0,0,0)$	264
B.3	The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (-0.25, 0.25, 0)$	265
B.4	The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (-0.5, 0.5, 0)$	265
B.5	The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (-1, 1, 0)$	266
B.6	The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (0.25, -0.25, 0)$	266
B.7	The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (0.5, -0.5, 0)$	267
B.8	The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (1, -1, 0)$	267
B.9	HbA _{1c} at the 12-month timepoint: imputed vs. observed for the “best” combination, $(\gamma_1, \gamma_2, \gamma_3) = (0.4, 0.2, 0)$	268

B.10	The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (0.4, 0.2, 0)$	268
B.11	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (0, 0, 0, 0)$	269
B.12	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (0.25, 0.25, 0.25, 0.25)$	269
B.13	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (0.5, 0.5, 0.5, 0.5)$	270
B.14	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (1, 1, 1, 1)$	270
B.15	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (0.25, -0.25, -0.25, 0.25)$	271
B.16	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (0.5, -0.5, -0.5, 0.5)$	271
B.17	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (1, -1, -1, 1)$	272
B.18	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-0.25, 0.25, 0.25, -0.25)$	272
B.19	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-0.5, 0.5, 0.5, -0.5)$	273
B.20	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-1, 1, 1, -1)$	273
B.21	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-0.25, -0.25, -0.25, -0.25)$	274
B.22	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-0.5, -0.5, -0.5, -0.5)$	274
B.23	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-1, -1, -1, -1)$	275
B.24	The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (3, 3, 3, 3)$	275

Nothing is wrong with making assumptions; on the contrary, they are the strands that join the field of statistics to scientific disciplines. The quality of these assumptions, not their existence, is the issue.

Little and Rubin (2000)

Part I

Preliminaries

1

Introduction

1.1 Background

Missing data are common wherever statistical methods are applied in practice. The problems they present are manifested in different ways in different contexts, but can be summarised by the following observation: if some data are missing, additional assumptions must be made about the mechanism leading to the incompleteness of the data and the relationship between the observed and unobserved data. These assump-

tions are inherently untestable and no amount of sophisticated mathematics can save us from this fact.

A common example of missing data in medicine is patient dropout in a clinical trial. Some examples of assumptions that might be made about the missing data in this context are

- A. that the outcome of interest for patients who drop out is, on average, equal to the outcome of interest for patients who remain in the study, or
- B. that the outcome of interest for those who dropped out, had they not dropped out, would have remained constant at the last observed value.

Both of these are now considered to be implausible and unnecessarily strong in most settings, and over the last two decades, much work has been done to develop methods that rely on weaker assumptions than these. For example, a weaker version of assumption A is that

- A'. if two subjects exhibit identical behaviour up to some point, whereafter one continues in the study but the other drops out, then the subsequent (unobserved) behaviour of the latter is, in distribution, equal to the subsequent behaviour of the former.

It transpires that given a fully-parametric model for the (unobserved) full data, models for the incomplete data based on assumption A' are identifiable from the observed data alone, and lead to valid inferences if the full-data model is correct and the assumption holds. Furthermore, assumption A' is minimal in this respect, in the sense that any model for the incomplete data based on a weaker assumption, such as

A". if two subjects exhibit identical behaviour up to some point, whereafter one continues in the study but the other drops out, then the subsequent (unobserved) behaviour of the latter is, on average, worse than or equal to the subsequent behaviour of the former,

cannot be identified from the observed data alone. Informally, the matter of ‘how much worse?’ in assumption A" is one which cannot be determined from the observed data.

Even if we are prepared to make assumption A', the validity of our inferences remains reliant on the correct specification of a model for the full data. This, of course, is the case for any statistical analysis, but if these assumptions are violated, the consequences (for example, induced bias) are more serious for incomplete data than they are for fully-observed data. Informally, when there are missing data, we compensate by relying more heavily on the assumptions of our full data model. This has led to the view, held by many in this field, that full-data models should be more robust (e.g. semiparametric) when the data are incomplete.

As is common to all areas of statistics, however, there is a trade-off between robustness and efficiency, and exactly where the correct balance lies is a matter of considerable contention. This has led to two broad schools of thought, as Molenberghs explains in response to Davidian *et al.* (2005):

“...the academic research community is divided between two rather opposing schools: the likelihood-oriented school of Rubin and co-workers, on the one hand, and the weighting-based school of Robins, Rotnitzky and co-workers, on the other hand. Exchanges between these two schools can certainly be entertaining, but when the debates are too fierce and go on too long, the winner is likely to be a third party. In this case, the third party may well be *last observation carried forward* (LOCF), *complete case analysis* (CC) and related simplistic methods.”

LOCF and CC are, respectively, the names given to assumptions B and A above. Rubin and Robins would no doubt agree that the broad aims of any researcher embarking on the analysis of an incomplete dataset should be

1. to understand and make explicit the assumptions made,
2. to minimise the strength of these assumptions wherever possible, whilst preserving other desirable properties of the analysis, such as efficiency and computational manageability, and
3. to investigate the sensitivity of any conclusions drawn to possible violations of the assumptions made.

Any disagreement between the two would be confined to the second of these aims and to where exactly the line should be drawn.

1.2 Outline of this thesis

Robins and his colleagues have introduced an attractive set of methods based on inverse probability weighting, the idea being that the bias induced by missing observations can be mitigated by weighting up the subjects most similar to those who are missing. By incorporating, in addition to the model for these weights, a model for the relationship between the observed and missing data, estimates are doubly robust (DR) in the sense that inferences are valid when at least one of the two models is correctly specified. The fact that these two additional models are specified allows the model for the full data to be less restrictive than might otherwise be the case, and the full-data models proposed by Robins and colleagues are semiparametric or even nonparametric, leading to the commonly used label of *semiparametric* for methods following this approach. Within their specified semiparametric (or nonparametric) classes, Robins and colleagues have shown their proposed estimators to be asymptotically optimally efficient.

On the other hand, one of the main methods proposed and advocated by Rubin, *multiple imputation* (MI), is also attractive. In MI, rather than weighting observed values, the focus is on filling in the missing values with appropriate ‘guesses’, whilst taking this into account in the subsequent inference so as not to fall into the trap of ‘counting the same information more than once’. MI uses only one of the two additional models employed in DR procedures: a model for the missing data conditional on the observed. A model for the weights is not contemplated. The method relies therefore on the correct specification of the model for the missing data conditional on the observed data for valid inferences, causing MI to be less robust than Robins’s DR methods. However, MI’s great strength is its flexibility and convenience in practice. Whereas DR methods need, in general, to be derived individually for each situation, MI is largely a ‘one size fits all’ approach, where, once the imputations have been drawn, valid inferences are obtained using a few simple general formulæ.

In this thesis, we propose a new method, doubly robust multiple imputation (DRMI), which combines the convenience of MI with the robustness of the weighting-based approach. Our aim is to use the computational flexibility of MI to provide a general framework for constructing DR estimators, extending to settings where, hitherto, estimators with this property have not been implemented. We apply the new method to data from the RECORD study, a clinical trial comparing type II diabetes drugs.

This thesis is divided into six parts. The remaining chapter of part 1 introduces the RECORD study dataset used in the remainder of the thesis. Part 2 lays down the mathematical foundations for the work contained in the thesis, and gives a detailed account of existing methods in the missing data literature, focussing on those methods which form the basis for this research. In part 3, we introduce our new method, doubly robust multiple imputation, and exhibit its properties both theoretically and using simulations. In part 4 (Chapter 10) we apply DRMI to the RECORD study dataset and (in Chapter 11) we then explore these data further, using MI to assess the sensitivity to the assumptions made about the missing data mechanism. Part 5 focusses on repeated binary data, exploring DRMI in this context, but also explaining theoretically some aspects of other methods hitherto not well understood. The final

part summarises the main conclusions of the thesis and suggests possible avenues of further research emanating from this work. Proofs, tables and figures considered not to be central to the main thrust of the arguments presented have been moved to the appendices at the end of the thesis, along with the computer code for any novel analysis.

2

The RECORD study: background

A clinical trial in type II diabetes mellitus patients carried out by GlaxoSmithKline motivates much of the work presented in this thesis. We introduce this example here.

Two well-established drugs prescribed to patients with type II diabetes are Metformin (Met) and Sulfonylurea (Su). The progressive nature of the disease, coupled with the setting of more stringent HbA_{1c} targets by practitioners, means that an increasing number of patients are taking combination therapies, such as both Met and Su. The primary aim of the RECORD (Rosiglitazone Evaluated for Cardiac Outcomes and

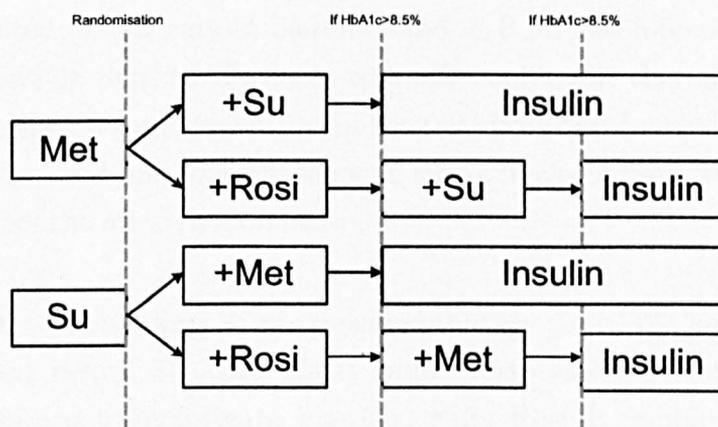


Figure 2.1: A summary of the design of the RECORD trial

Regulation of glycaemia in Diabetes) study was to evaluate the safety (by looking at cardiac outcomes) of a third drug, Rosiglitazone (Rosi), when used in combination with either Met or Su. See, for example, the article by Nissen and Wolski (2007) for some background on these cardiovascular safety concerns. However, a secondary analysis was also planned, to assess the efficacy of Rosi as regards glycaemic control. It is with this secondary analysis, in particular the measurement of HbA_{1c} (a measure of the average level of glucose in the blood over the 8–12 weeks prior to measurement), that we are concerned.

The RECORD trial recruited 4458 patients from 330 centres in 23 countries, all of whom were taking either Met or Su (monotherapy) prior to the start of the trial. The Met and Su arms were subsequently treated as two separate strata, with patients in the Met arm randomised to receive either additional Su or additional Rosi, and patients in the Su arm randomised to receive either additional Met or additional Rosi. If dual therapy proved not to be effective ($\text{HbA}_{1c} \geq 8.5\%$ in two consecutive measurements >1 month apart and at least 8 weeks from titration to maximum dose of study medication), patients in the Met+Rosi or Su+Rosi arm would be given additional Su or Met respectively, whereas patients in the two non-Rosi arms would be put straight onto insulin; that is, the protocol for the Rosi and non-Rosi arms differed. This is summarised in Fig. 2.1.

HbA_{1c} was collected on patients at baseline, and at 8 further follow-up visits: at 2, 4, 6, 8, 10, 12, 15 and 18 months. Patients who left dual/triple therapy were considered to belong only to the CV-outcome phase of the trial from the date at which they ceased dual/triple therapy, and only one subsequent HbA_{1c} measurement was taken on these patients: at 12 months after randomisation.

Since recruitment was on-going, it was specified that the first 1122 patients randomised (those randomised before 15 April 2002) would form the cohort for this 18-month analysis. The aim was to investigate whether or not Rosi in combination with Met or Su is as good as Met+Su for achieving glycaemic control. The pre-specified endpoint (see Home *et al.*, 2005) was to look at the change in mean HbA_{1c} from baseline at 18 months after randomisation, but any difference in the trajectories between the different combination therapies is also of interest. The non-inferiority criterion (upper band 95% CI of difference) was set at 0.4%.

As in almost all longitudinal studies, there were patients lost to follow-up and also patients who failed to comply with their treatment protocol for the duration of the follow-up time. The original analysis of these data (Home *et al.*, 2007) made certain assumptions about the mechanisms leading to the dropout and noncompliance. In Chapter 11 we assess the robustness of their conclusions to possible violations of these assumptions.

HbA_{1c} in type II diabetic patients, as it is a measure of glycaemic control, is likely to fall outside the range considered to be normal in the general population. However, the nature of the disease means that HbA_{1c} levels in type II diabetic patients are far more likely to be abnormally high than abnormally low. As such, we may expect HbA_{1c} measurements in this population to be non-normal and to exhibit some right skewness. The original analysis carried out by Home *et al.* (2007) assumed multivariate normality for the repeated HbA_{1c} measurements conditional on baseline HbA_{1c}. In Chapter 10, we assess the sensitivity of their conclusions by relaxing the multivariate normality assumption, an assumption which carries extra weight when the data are incomplete.

Part II

Literature review & theoretical foundations

3

Notation and basic definitions

3.1 Full data and associated quantities

Definition 3.1 (Full-data model). The *full data* are the data we would collect in an ideal setting, if there were no missing data. Following the notation used by Tsiatis (2006), we write these full data as

$$\{ \mathbf{Z}_i : i = 1, \dots, n \}$$

where n is the number of subjects and \mathbf{Z}_i is the full data for subject i . We assume* that the \mathbf{Z}_i 's are independent and identically distributed random vectors from a parametric model with true density of the form

$$p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta})$$

for some true value $\boldsymbol{\theta}_0$ of the q -dimensional parameter $\boldsymbol{\theta}$.

Our aim is usually either to make inference about $\boldsymbol{\theta}$, a subset $\boldsymbol{\psi}$ of $\boldsymbol{\theta}$, or a function $\boldsymbol{\psi}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$.

We will often distinguish between outcomes (\mathbf{Y}) and covariates (\mathbf{X}), where $\mathbf{Z}_i = (\mathbf{X}_i^T, \mathbf{Y}_i^T)^T$. Occasionally we will further distinguish between the covariates of interest (\mathbf{X}) and the *auxilliary* covariates (\mathbf{V}), and write $\mathbf{Z}_i = (\mathbf{X}_i^T, \mathbf{V}_i^T, \mathbf{Y}_i^T)^T$.

Definition 3.2 (Saturated model). If the covariates \mathbf{X}_i on which we are conditioning are all categorical, and $\boldsymbol{\theta}$ (or a transformation of $\boldsymbol{\theta}$) contains a separate parameter for the mean of \mathbf{Y}_i for every possible combination of the categorical covariates, then the model is said to be *saturated*.

The subscript $i \in \{1, \dots, n\}$ always indexes subjects, with $j \in \{1, \dots, J\}$ and $t \in \{1, \dots, T\}$ used to index the constituent random variables of \mathbf{Z}_i , e.g. $\mathbf{Z}_i = (Z_{1,i}, \dots, Z_{j,i}, \dots, Z_{J,i})^T$. t is reserved for longitudinal data, and j is used either for cross-sectional data, or in the general case where we do not wish to specify whether the data are longitudinal or cross-sectional. That is, $\{Z_{t,i} : 1 \leq t \leq T\}$ are repeated measurements of the same outcome on a given subject, and $\{Z_{j,i} : 1 \leq j \leq J\}$ could either be different variables measured on the same subject simultaneously, or either of the two cases when we do not wish to specify. In the longitudinal setting, $Z_{t,i}$ (or $Z_{j,i}$) could be vector-valued, and denoted by $\mathbf{Z}_{t,i}$ (or $\mathbf{Z}_{j,i}$).

*Even when using a semiparametric or nonparametric approach, we assume that there is *some* parametric distribution that gave rise to the data under consideration, even if we choose not to postulate the form of $p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta})$.

Unless specified otherwise, the superscript \cdot^T is used to denote matrix transpose and should not be mistaken for T , the upper bound of t .

Definition 3.3 (Full-data score). The *full-data score vector*, $\mathbf{S}_\theta^F(\mathbf{Z}, \theta)$, is defined as:

$$\mathbf{S}_\theta^F(\mathbf{Z}, \theta) = \frac{\partial \log [p_{\mathbf{Z}}(\mathbf{Z}, \theta)]}{\partial \theta}$$

Definition 3.4 (Full-data information). The *full-data information matrix*, $\mathbf{I}_{\theta\theta}^F(\theta)$, is defined as:

$$\mathbf{I}_{\theta\theta}^F(\theta) = \mathbb{E} \left\{ -\frac{\partial^2 \log [p_{\mathbf{Z}}(\mathbf{Z}, \theta)]}{\partial \theta \partial \theta^T} \right\} = \mathbb{E} \left\{ -\frac{\partial [\mathbf{S}_\theta^F(\mathbf{Z}, \theta)^T]}{\partial \theta} \right\}$$

where the expectation is with respect to the true distribution of \mathbf{Z} .

Definition 3.5 (Asymptotically linear estimator). An *asymptotically linear estimator* $\hat{\theta}_n$ of θ is one that satisfies the following:

$$n^{\frac{1}{2}} (\hat{\theta}_n - \theta_0) = n^{-\frac{1}{2}} \sum_{i=1}^n \varphi(\mathbf{Z}_i) + o_p(1)$$

where $\varphi(\mathbf{Z}_i)$ is called the *ith influence function* of $\hat{\theta}_n$ and must satisfy the following:

$$\mathbb{E}[\varphi(\mathbf{Z}_i)] = \mathbf{0}$$

$$\mathbb{E}[\varphi(\mathbf{Z}_i) \varphi(\mathbf{Z}_i)^T] < \infty$$

Definition 3.6 (Regular asymptotically linear (RAL) estimator). Under certain regularity conditions (see Tsiatis, 2006), all asymptotically linear estimators share the property that

$$n^{\frac{1}{2}} (\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathbf{N}(0, \Sigma) \quad (3.1.1)$$

as $n \rightarrow \infty$, where $\xrightarrow{\mathcal{D}}$ is used to mean *tends in distribution* in the measure-theoretic sense as described by Williams (1991). We will call such estimators *regular asymptotically linear* or *RAL*. Conversely, it can be shown that any estimator that satisfies (3.1.1) is

RAL. Furthermore, this association between RAL estimators and influence functions is one-to-one.

3.2 Coarsened data

The notion of coarsened data was first introduced by Heitjan and Rubin (1991).

Definition 3.7 (Injective function). A function $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is *injective* if $f(x) = f(y) \Rightarrow x = y$ for any $x, y \in \mathbb{R}^m$.

Definition 3.8 (Non-injective function). A function $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ which is not injective is called *non-injective*. That is, f is non-injective if there exist at least one pair $x, y \in \mathbb{R}^m$ such that $x \neq y$ and $f(x) = f(y)$.

Definition 3.9 (Coarsened data). Suppose that, for each $i = 1, \dots, n$, instead of observing \mathbf{Z}_i , we observe a *coarsening indicator* $C_i \in \mathbb{Z}^+ \cup \{0, \infty\}$, together with $G_{C_i}(\mathbf{Z}_i)$, where $G_{C_i}(\mathbf{Z}_i)$ is a *non-injective* function of \mathbf{Z}_i . Then

$$\{ [C_i, G_{C_i}(\mathbf{Z}_i)] : i = 1, \dots, n \}$$

are known as the *coarsened data*. We will also refer to these as the *observed data* and write the *observed-data density* as $p_{C, G_C(\mathbf{Z})}(c, \mathbf{g}_c, \boldsymbol{\theta}, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is a set of additional parameters that govern the distribution of C given \mathbf{Z} .

Intuitively, we can use Fig. 3.1 to conceptualise coarsened data. Consider a discrete (and finite) example where the sample space Ω consists only of 25 possible one-dimensional values (labelled $p1$ – $p25$). Now suppose that Ω is divided into subspaces $A1, A2, \dots, A8$, and that rather than directly observing $\mathbf{Z}_i = Z_i$, we observe only the subspace ($A1, A2, \dots$, or $A8$) in which Z_i lies. For example, if $Z_i = p14$, then $G_{c_i}(Z_i) = A6$. For each observation, a different partition of the sample space is allowed and these different partitions are indexed by $c \in \mathbb{Z}^+ \cup \{0, \infty\}$.

We use $C = \infty$ to denote the case where we observe the full data, i.e. $G_\infty(\mathbf{Z}) = \mathbf{Z}$.

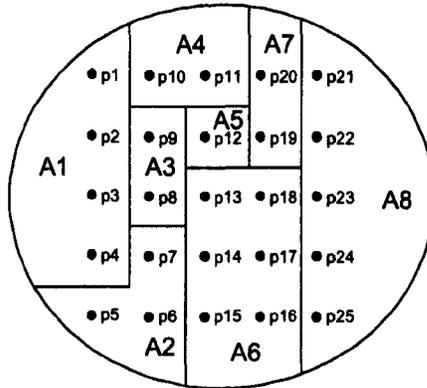


Figure 3.1: A visual depiction of coarsened data

3.3 Missing data as a special case of coarsening

Suppose $\mathbf{Z}_i = (Z_{1,i}, Z_{2,i}, \dots, Z_{J,i})^T$, then we can view missing data as a special case of coarsening if we let $\{G_c(\mathbf{Z}_i) : c = 0, \dots, 2^J - 1\}$ be all possible subvectors of \mathbf{Z}_i , indexed by c . A conventional way of ordering the subvectors is to let $G_0(\mathbf{Z}_i) = \emptyset$ and $G_{2^J-1}(\mathbf{Z}_i) = \mathbf{Z}_i$ so that $G_{2^J-1}(\mathbf{Z}_i) = G_\infty(\mathbf{Z}_i)$.

In the case where $\mathbf{Z}_i = Z_i$ is univariate, each member of the population is either observed ($C_i = \infty$) or not observed ($C_i = 0$). Fig. 3.2 illustrates missing data as a special case of coarsened data for univariate Z_i . In this case, p_1 - p_{15} are all observed ($G_{c_i}(p_k) = A_k = p_k$ for $k = 1, \dots, 15$), whereas $G_{c_i}(p_k) = A_{16}$ for $k = 16, \dots, 25$, and all we observe is that $C_i = 0 \forall k \in \{16, \dots, 25\}$. So for subject i , if $Z_i = p_k$, $k \in \{1, \dots, 15\}$ then $C_i = \infty$ but if $Z_i = p_k$, $k \in \{16, \dots, 25\}$ then $C_i = 0$. For two different subjects, these two sets ($\{1, \dots, 15\}$ and $\{16, \dots, 25\}$) could be defined differently.

Definition 3.10. Sometimes it is also useful to consider the set

$$\{(C_i, \mathbf{Z}_i) : i = 1, \dots, n\}$$

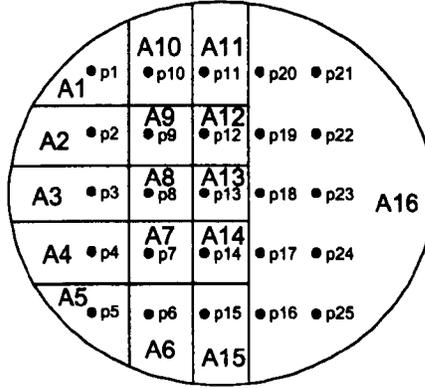


Figure 3.2: A visual depiction of missing data as a special case of coarsening

which we call the *unobservable data*. These have density

$$p_{C,\mathbf{Z}}(c, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\xi}) \quad (3.3.1)$$

3.4 Observed data quantities

The observed-data score vector and observed-data information matrix are defined in an analogous way to the full-data equivalents.

Definition 3.11 (Observed-data score). The *observed-data score vector*, $\mathbf{S}_{\boldsymbol{\theta}}[C, G_C(\mathbf{Z}), \boldsymbol{\theta}, \boldsymbol{\xi}]$, is defined as:

$$\mathbf{S}_{\boldsymbol{\theta}}[C, G_C(\mathbf{Z}), \boldsymbol{\theta}, \boldsymbol{\xi}] = \frac{\partial \log [p_{C, G_C(\mathbf{Z})}(c, \mathbf{g}_c, \boldsymbol{\theta}, \boldsymbol{\xi})]}{\partial \boldsymbol{\theta}}$$

Definition 3.12 (Observed-data information). The *observed-data information matrix*, $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta})$, is defined as:

$$\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbb{E} \left\{ -\frac{\partial^2 \log [p_{C, G_C(\mathbf{Z})}(c, \mathbf{g}_c, \boldsymbol{\theta}, \boldsymbol{\xi})]}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\} = \mathbb{E} \left(-\frac{\partial \left\{ \mathbf{S}_{\boldsymbol{\theta}}[C, G_C(\mathbf{Z}), \boldsymbol{\theta}, \boldsymbol{\xi}]^T \right\}}{\partial \boldsymbol{\theta}} \right)$$

where the expectation is with respect to the true distribution of $[C, G_C(\mathbf{Z})]$.

3.5 Missingness indicators

Although it will sometimes be useful to derive results under the more general notion of coarsening and subsequently apply these results to the special case of missing data, coarsened data are not the direct focus of any part of this thesis. For this reason, it is useful to define *missingness indicators* as well as the coarsening indicator C_i .

As we have seen, the observed data as a function of the full data \mathbf{Z}_i can be unambiguously described using the univariate coarsening indicator $C_i \in \mathbb{Z}^+ \cup \{0, \infty\}$. However, often more useful in practice is the following:

Definition 3.13 (Missingness indicator vector). Let $\mathbf{Z}_i = (Z_{1,i}, Z_{2,i}, \dots, Z_{J,i})^T$. The *missingness indicator vector*, \mathbf{R}_i is given by

$$\mathbf{R}_i = [\mathbb{1}(Z_{1,i} \text{ is observed}), \mathbb{1}(Z_{2,i} \text{ is observed}), \dots, \mathbb{1}(Z_{J,i} \text{ is observed})]^T$$

and $R_{j,i}$ is the j^{th} element of \mathbf{R}_i , i.e. $R_{j,i} = \mathbb{1}(Z_{j,i} \text{ is observed})$.

Thus, whenever the condition $C_i = c_i$ is used in the coarsened data formulation, it can be translated as $\mathbf{R}_i = \mathbf{r}_i$ in the missing data formulation. Correspondingly, $[C_i, G_{C_i}(\mathbf{Z}_i)]$, the observed data in the coarsened data formulation, can be translated as $(\mathbf{R}_i, \mathbf{Z}_i^{\text{obs}})$ in the missing data formulation, where $\mathbf{Z}_i^{\text{obs}} = \{Z_{j,i} : R_{j,i} = 1\}$ (and $\mathbf{Z}_i^{\text{mis}} = \{Z_{j,i} : R_{j,i} = 0\}$).

Furthermore, if $\{(C_i, \mathbf{Z}_i) : i = 1, \dots, n\}$ are the unobservable data under coarsening, then $\{(\mathbf{R}_i, \mathbf{Z}_i) : i = 1, \dots, n\}$ are the unobservable data under missingness, with associated density

$$p_{\mathbf{R}, \mathbf{Z}}(\mathbf{r}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\zeta}) \tag{3.5.1}$$

3.6 Monotonicity and dropout

Definition 3.14 (Monotone missing data pattern). If

$$\mathbf{Z}_i = (Z_{1,i}, Z_{2,i}, \dots, Z_{J,i})^T$$

and there exists a permutation $\{p_1, p_2, \dots, p_J\}$ of $\{1, 2, \dots, J\}$ with

$$\mathbf{Z}'_i = (Z_{p_1,i}, Z_{p_2,i}, \dots, Z_{p_J,i})^T$$

such that if $R_{p_j,i} = 0$ then $R_{p_k,i} = 0$ for all $k > j$, then the missing data pattern is said to be *monotone*.

Definition 3.15 (General missing data pattern). If the missing data pattern is not monotone, then it is said to be *general*.

Monotonicity implies that the variables can be ordered as shown in Fig. 3.3. The condition trivially holds if only one variable is incomplete and can otherwise occur in clinical trials where data are only missing because of loss-to-follow-up, where patients leave and never return, in which case the variables would be ordered with increasing time. For cross-sectional data with more than one incomplete variable, monotonicity is unlikely to hold.

Definition 3.16 (Dropout indicator). Under monotonicity, $\mathbf{R}_i \in \{0, 1\}^T$ (where T is the upper bound of t and not used here to denote matrix transpose) is restricted to a vector which must contain a series of $D_i - 1$, say, ones, followed by $T - D_i + 1$ zeros. $D_i \in \{1, \dots, T + 1\}$ is called the *dropout indicator* and represents the first time at which subject i was not observed.

Definition 3.17 (History). For any vector $\mathbf{W}_i = (W_{1,i}, W_{2,i}, \dots, W_{K,i})^T$, let the *history* of \mathbf{W}_i up to time k be

$$\bar{\mathbf{W}}_{k,i} = (W_{1,i}, W_{2,i}, \dots, W_{k,i})^T$$

Under monotonicity, with $D_i = d_i$, $\mathbf{Z}_i^{\text{obs}} = \bar{\mathbf{Z}}_{d_i-1,i}$.



Figure 3.3: A diagrammatic depiction of a monotone missing data pattern (the shaded areas are unobserved)

It is common to impose the additional restriction $D_i > 1$ —that every subject in the dataset is observed on at least one occasion.

3.7 Semiparametric model

Definition 3.18 (Semiparametric model). Implicit in our definition (3.5.1) of the unobservable data density $p_{\mathbf{R}, \mathbf{Z}}(\mathbf{r}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\zeta})$ is that $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ are both finite-dimensional. In a *semiparametric* model, this assumption is relaxed, and either $\boldsymbol{\theta}$ or $\boldsymbol{\zeta}$ (or both) can be infinite-dimensional. Usually, the parameter of interest is finite-dimensional, e.g. if $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$ and $\boldsymbol{\theta}$ is infinite-dimensional, then $\boldsymbol{\beta}$, the parameter of interest, is finite-dimensional, and $\boldsymbol{\eta}$, the nuisance parameter, is infinite-dimensional. We use \mathcal{S}

to denote this broad class of semiparametric models.

Definition 3.19 (Semiparametric model for incomplete data). Models for incomplete data often require (in addition to a model for the full data) the specification of either a model for the probability of observing the data conditional on the observed data or a model for the full data conditional on the observed data, or—in some cases—both. When a method for analysing incomplete data relies on a model for the full data conditional on the observed data, the method can be semiparametric in the sense that the distribution of the full data given the observed data (or the aspect of this distribution needed for the analysis) is estimated non-parametrically from the observed data. The full-data model remains parametric. We use $\mathcal{I} \subset \mathcal{S}$ to denote this subclass of semiparametric models. An example is given in §7.2.

Definition 3.20 (Restricted moment model). We use $\mathcal{R} \subset \mathcal{S}$ to denote the subclass of semiparametric models when only the mean of the distribution is modelled, both in the full-data model and (in an incomplete data problem) the model for the full data conditional on the observed.

Definition 3.21 (Parametric submodel). Let \mathcal{F} denote a family of unobservable semiparametric densities of the form $p_{\mathbf{R},\mathbf{Z}}(\mathbf{r}, \mathbf{z}, \boldsymbol{\theta}_\infty, \boldsymbol{\zeta}_\infty)$ where $\boldsymbol{\theta}_\infty$ and $\boldsymbol{\zeta}_\infty$ are potentially infinite-dimensional. In truth, the data have been generated from a parametric density $p_0 = p_{\mathbf{R},\mathbf{Z}}(\mathbf{r}, \mathbf{z}, \boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ where $\boldsymbol{\theta}_0$ and $\boldsymbol{\zeta}_0$ are finite-dimensional. A *parametric submodel* $\mathcal{F}_{\boldsymbol{\theta}_f, \boldsymbol{\zeta}_f}$ is a family of unobservable parametric densities of the form $p_{\mathbf{R},\mathbf{Z}}(\mathbf{r}, \mathbf{z}, \boldsymbol{\theta}_f, \boldsymbol{\zeta}_f)$ where $\boldsymbol{\theta}_f$ and $\boldsymbol{\zeta}_f$ are finite-dimensional such that

1. $p_0 \in \mathcal{F}_{\boldsymbol{\theta}_f, \boldsymbol{\zeta}_f}$, and
2. $\mathcal{F}_{\boldsymbol{\theta}_f, \boldsymbol{\zeta}_f} \subset \mathcal{F}$

Definition 3.22 (Semiparametric-efficient estimator). Within each parametric submodel $\mathcal{F}_{\boldsymbol{\theta}_f, \boldsymbol{\zeta}_f}$ of a semiparametric family \mathcal{F} , let v be the variance of the most efficient estimator in the family. Then the *semiparametric efficiency bound* (Newey, 1990) is the supremum of v across all parametric submodels of \mathcal{F} and a *semiparametric-efficient estimator* is an estimator with asymptotic variance achieving this bound.

4

Missing data modelling frameworks and mechanisms

4.1 Modelling frameworks

Recall that the unobservable data $\{(\mathbf{R}_i, \mathbf{Z}_i) : i = 1, \dots, n\}$ have density $p_{\mathbf{R}, \mathbf{Z}}(\mathbf{r}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\zeta})$. Three different modelling approaches arise from three different factorisations of this density.

- **Selection** modelling (Heckman, 1976), where

$$p_{\mathbf{R},\mathbf{Z}}(\mathbf{r}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\zeta}) = p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) p_{\mathbf{R}|\mathbf{Z}}(\mathbf{r}|\mathbf{z}, \boldsymbol{\zeta})$$

As we see in the next section, under certain circumstances, $p_{\mathbf{R}|\mathbf{Z}}(\mathbf{r}|\mathbf{z}, \boldsymbol{\zeta})$ can be *ignored* in the analysis, making this framework appealing.

- **Pattern-mixture** modelling (Little, 1993), where

$$p_{\mathbf{R},\mathbf{Z}}(\mathbf{r}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\zeta}) = p_{\mathbf{Z}|\mathbf{R}}(\mathbf{z}|\mathbf{r}, \boldsymbol{\theta}) p_{\mathbf{R}}(\mathbf{r}, \boldsymbol{\zeta})$$

This is arguably the most natural framework in longitudinal studies. For example, in the RECORD study, where patients attend a clinic every few months to be measured, we could think of their data as being generated in two stages. First, the patient decides on the morning of the visit whether or not he or she will attend. Then, given that the patient attends, an HbA_{1c} measurement is taken and observed. If the patient does not attend, we imagine a *counterfactual* observation: that which would have been observed had the patient attended.

- and **shared parameter** modelling (Wu and Carroll, 1988), where

$$p_{\mathbf{R},\mathbf{Z}}(\mathbf{r}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\zeta}) = \int_{\mathbf{b}} p_{\mathbf{Z}|\mathbf{B}}(\mathbf{z}|\mathbf{b}, \boldsymbol{\theta}) p_{\mathbf{R}|\mathbf{B}}(\mathbf{r}|\mathbf{b}, \boldsymbol{\zeta}) p_{\mathbf{B}}(\mathbf{b}, \boldsymbol{\omega}) d\mathbf{b}$$

Here, \mathbf{B} is a set of latent random effects (with realisations \mathbf{b}) governing the distribution of both the full data and the missingness mechanism. \mathbf{Z} and \mathbf{R} are *conditionally independent* given \mathbf{B} .

4.2 Missing and coarsened data mechanisms

The classification of missing data as Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR) dates back to Rubin (1976) and is most naturally described in the selection modelling framework of the previous

section.

Informally, the classification can be thought of as follows. If the probability of observing a particular data point is completely independent of all other observations in the dataset (observed and unobserved), then the missing data are said to be MCAR. If the probability of observing a particular data point depends on other observed data, but (after conditioning on the observed data) is independent of the unobserved data, then the missing data are MAR. If, even after conditioning on the observed data, the probability of observing a particular data point depends on the unobserved data, then the missing data are MNAR.

More generally, an analogous classification for coarsened data (Heitjan and Rubin, 1991) can be formally defined as follows:

Definition 4.1 (Coarsening completely at random (CCAR)). If $\mathbb{P}(C = c|\mathbf{Z}) = \pi$, the data are said to be *coarsened completely at random*.

Definition 4.2 (Coarsening at random (CAR)). If $\mathbb{P}(C = c|\mathbf{Z}) = \pi[c, G_c(\mathbf{Z})]$, the data are said to be *coarsened at random*.

Definition 4.3 (Coarsening not at random (CNAR)). If $\mathbb{P}(C = c|\mathbf{Z}) = \pi(c, \mathbf{Z})$, the data are said to be *coarsened not at random*.

This classification has remained at the heart of almost all the work in the missing data area since its introduction in 1976. MCAR is often implausible, and analyses assuming only MNAR are usually difficult to implement and often rely on additional information, external to the data. This means that the MAR assumption has become a central part of much of the literature on missing data methods. This is not to say that the assumption is justified as often as it is used. In fact, justifying the MAR assumption from the observed data alone is impossible: a premise formally shown by Gill *et al.* (1996) and further explored by Molenberghs *et al.* (2008). Whether or not the missingness mechanism depends intrinsically on the unobserved data after conditioning on the observed data is *counterfactual* in the sense that it cannot be determined without

knowing what the unobserved data *would have* been had we observed them. This means that in order to proceed, additional assumptions must be made, and consequently, the sensitivity of any conclusions to violations of the additional assumptions should be investigated.

Acquiring an intuitive feel for the MAR assumption is easiest when considering monotone missing data. In this setting, the MAR assumption states that if two subjects, i_1 and i_2 , exhibit identical behaviour up to some point $d - 1$, whereafter subject i_1 continues in the study, while subject i_2 drops out, then the subsequent (unobserved) behaviour of subject i_2 is identical *in distribution* to the subsequent behaviour of subject i_1 . More formally, as was shown by Molenberghs *et al.* (1998),

Proposition 4.1 (MAR under dropout). *Under the MAR assumption, if the non-response is monotone,*

$$\begin{aligned} p_{Z_d, Z_{d+1}, \dots, Z_T | \bar{\mathbf{z}}_{d-1}, D} (z_d, z_{d+1}, \dots, z_T | \bar{\mathbf{z}}_{d-1}, D = d) \\ = p_{Z_d, Z_{d+1}, \dots, Z_T | \bar{\mathbf{z}}_{d-1}, D} (z_d, z_{d+1}, \dots, z_T | \bar{\mathbf{z}}_{d-1}, D > d) \end{aligned}$$

We must first prove the following lemma:

Lemma 4.2. *Under MAR and monotonicity, $p_{\bar{\mathbf{R}}_d | \mathbf{z}} (R_d = 1 | \mathbf{z})$ is independent of z_d, z_{d+1}, \dots, z_T .*

Proof (by induction). Suppose that $p_{\bar{\mathbf{R}}_{d-1} | \mathbf{z}} (R_{d-1} = 1 | \mathbf{z})$ is independent of z_{d-1}, z_d, \dots, z_T . Then,

$$\begin{aligned} p_{\bar{\mathbf{R}}_d | \mathbf{z}} (R_d = 1 | \mathbf{z}) &= p_{R_d | \mathbf{z}, \bar{\mathbf{R}}_{d-1}} (R_d = 1 | \mathbf{z}, R_{d-1} = 1) p_{\bar{\mathbf{R}}_{d-1} | \mathbf{z}} (R_{d-1} = 1 | \mathbf{z}) \\ &= \left[1 - p_{R_d | \mathbf{z}, \bar{\mathbf{R}}_{d-1}} (R_d = 0 | \mathbf{z}, R_{d-1} = 1) \right] p_{\bar{\mathbf{R}}_{d-1} | \mathbf{z}} (R_{d-1} = 1 | \mathbf{z}) \\ &= p_{\bar{\mathbf{R}}_{d-1} | \mathbf{z}} (R_{d-1} = 1 | \mathbf{z}) - p_{\mathbf{R} | \mathbf{z}} (R_{d-1} = 1, R_d = 0 | \mathbf{z}) \quad (4.2.1) \end{aligned}$$

The first term in (4.2.1) is independent of z_{d-1}, \dots, z_T (and hence z_d, \dots, z_T) by the inductive step, and the second term is independent of z_d, \dots, z_T by the MAR assumption.

$p_{R_1|\mathbf{z}}(R_1 = 1|\mathbf{z})$ is usually assumed to be 1 (hence independent of \mathbf{z}), but even when this convention is not followed,

$$p_{R_1|\mathbf{z}}(R_1 = 1|\mathbf{z}) = 1 - p_{R_1|\mathbf{z}}(R_1 = 0|\mathbf{z})$$

which must be independent of \mathbf{z} by the MAR assumption. \square

Now we prove Proposition 4.1.

Proof.

$$\begin{aligned}
 & p_{Z_d, Z_{d+1}, \dots, Z_T | \bar{\mathbf{z}}_{d-1}, D}(z_d, z_{d+1}, \dots, z_T | \bar{\mathbf{z}}_{d-1}, d) \\
 &= p_{Z_d, Z_{d+1}, \dots, Z_T | \bar{\mathbf{z}}_{d-1}, \mathbf{R}}(z_d, z_{d+1}, \dots, z_T | \bar{\mathbf{z}}_{d-1}, R_{d-1} = 1, R_d = 0) \\
 & \hspace{15em} (4.2.2) \\
 &= \frac{p_{\mathbf{Z}, \mathbf{R}}(\mathbf{z}, R_{d-1} = 1, R_d = 0)}{p_{\bar{\mathbf{z}}_{d-1}, \mathbf{R}}(\bar{\mathbf{z}}_{d-1}, R_{d-1} = 1, R_d = 0)} \\
 &= \frac{p_{\mathbf{R}|\mathbf{z}}(R_{d-1} = 1, R_d = 0|\mathbf{z}) p_{\mathbf{z}}(\mathbf{z})}{p_{\mathbf{R}|\bar{\mathbf{z}}_{d-1}}(R_{d-1} = 1, R_d = 0|\bar{\mathbf{z}}_{d-1}) p_{\bar{\mathbf{z}}_{d-1}}(\bar{\mathbf{z}}_{d-1})} \\
 &= \frac{p_{\mathbf{R}|\bar{\mathbf{z}}_{d-1}}(R_{d-1} = 1, R_d = 0|\bar{\mathbf{z}}_{d-1}) p_{\mathbf{z}}(\mathbf{z})}{p_{\mathbf{R}|\bar{\mathbf{z}}_{d-1}}(R_{d-1} = 1, R_d = 0|\bar{\mathbf{z}}_{d-1}) p_{\bar{\mathbf{z}}_{d-1}}(\bar{\mathbf{z}}_{d-1})} \\
 &= \frac{p_{\mathbf{z}}(\mathbf{z})}{p_{\bar{\mathbf{z}}_{d-1}}(\bar{\mathbf{z}}_{d-1})}
 \end{aligned}$$

by the MAR assumption.

Similarly,

$$\begin{aligned}
 & p_{Z_d, Z_{d+1}, \dots, Z_T | \bar{\mathbf{z}}_{d-1}, D} (z_d, z_{d+1}, \dots, z_T | \bar{\mathbf{z}}_{d-1}, D > d) \\
 &= p_{Z_d, Z_{d+1}, \dots, Z_T | \bar{\mathbf{z}}_{d-1}, \bar{\mathbf{R}}_d} (z_d, z_{d+1}, \dots, z_T | \bar{\mathbf{z}}_{d-1}, R_d = 1) \quad (4.2.3) \\
 &= \frac{p_{\mathbf{z}, \bar{\mathbf{R}}_d} (\mathbf{z}, R_d = 1)}{p_{\bar{\mathbf{z}}_{d-1}, \bar{\mathbf{R}}_d} (\bar{\mathbf{z}}_{d-1}, R_d = 1)} \\
 &= \frac{p_{\bar{\mathbf{R}}_d | \mathbf{z}} (R_d = 1 | \mathbf{z}) p_{\mathbf{z}} (\mathbf{z})}{p_{\bar{\mathbf{R}}_d | \bar{\mathbf{z}}_{d-1}} (R_d = 1 | \bar{\mathbf{z}}_{d-1}) p_{\bar{\mathbf{z}}_{d-1}} (\bar{\mathbf{z}}_{d-1})}
 \end{aligned}$$

In (4.2.2) we use the subscript \mathbf{R} , whereas in (4.2.3) we use the subscript $\bar{\mathbf{R}}_d$. This is since $\{R_{d-1} = 1, R_d = 0\}$ completely specifies \mathbf{R} , but $R_d = 1$ does not. This means that we cannot immediately use the MAR assumption to write

$$p_{\bar{\mathbf{R}}_d | \bar{\mathbf{z}}_{d-1}} (R_d = 1 | \bar{\mathbf{z}}_{d-1}) = p_{\bar{\mathbf{R}}_d | \mathbf{z}} (R_d = 1 | \mathbf{z}) \quad (4.2.4)$$

However, (4.2.4) holds by Lemma 4.2. Thus,

$$\begin{aligned}
 & p_{Z_d, Z_{d+1}, \dots, Z_T | \bar{\mathbf{z}}_{d-1}, D} (z_d, z_{d+1}, \dots, z_T | \bar{\mathbf{z}}_{d-1}, D > d) \\
 &= \frac{p_{\bar{\mathbf{R}}_d | \bar{\mathbf{z}}_{d-1}} (R_d = 1 | \bar{\mathbf{z}}_{d-1}) p_{\mathbf{z}} (\mathbf{z})}{p_{\bar{\mathbf{R}}_d | \bar{\mathbf{z}}_{d-1}} (R_d = 1 | \bar{\mathbf{z}}_{d-1}) p_{\bar{\mathbf{z}}_{d-1}} (\bar{\mathbf{z}}_{d-1})} \\
 &= \frac{p_{\mathbf{z}} (\mathbf{z})}{p_{\bar{\mathbf{z}}_{d-1}} (\bar{\mathbf{z}}_{d-1})}
 \end{aligned}$$

□

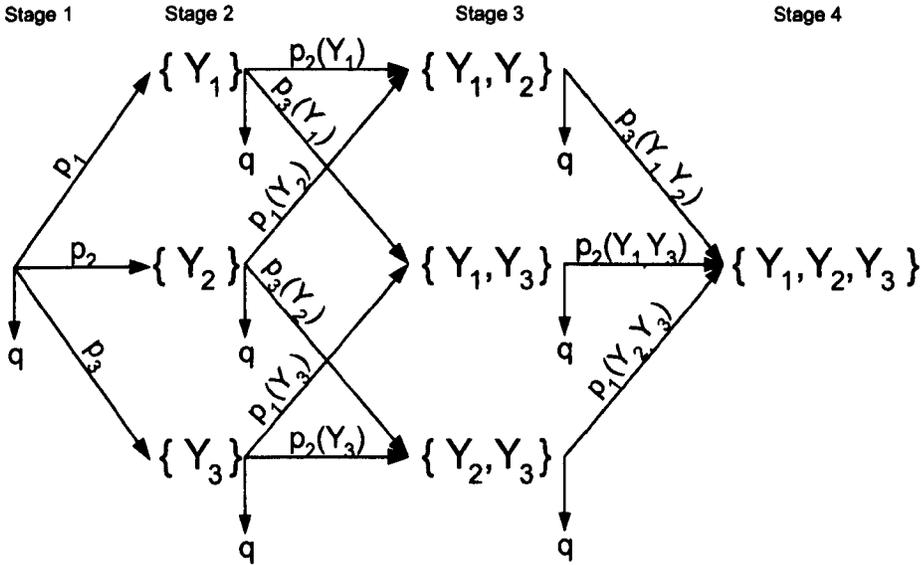


Figure 4.1: A Markov randomised monotone missingness process for $J = 3$. Dependence on \mathbf{X} is implicit.

4.2.1 Randomised monotone missingness (RMM) processes

We have seen here how the MAR assumption—expressed in the selection modelling framework—translates into the pattern-mixture language when the data are monotone (Molenberghs *et al.*, 1998). Such a translation does not exist in general for non-monotone missing data patterns.

The appropriateness of the MAR assumption for non-monotone missing data has been questioned (see Robins and Gill, 1997). These authors introduce a new mechanism, *randomised monotone missingness* (RMM)—a subset of MAR—and argue that this is the only plausible non-monotone MAR mechanism that is not MCAR. They show (in Gill and Robins, 1996) that there exist mechanisms that are MAR but not RMM, but that in order for a computer to generate data under such a mechanism, it requires knowledge of the unobserved data which is then ‘concealed’ later in the process. They call this phenomenon ‘MAR is more than it seems’ and say:

“We have been unable to conceive of a plausible social, economic, physical or biological process that would generate MAR processes that are not RMM representable, due to the subtle and precise manner in which the data must be ‘hidden’ to insure that the process is MAR. That is, we believe that natural missing data processes that are not representable as RMM processes will be [MNAR].”

Write the full data \mathbf{Z}_i for subject i as $\mathbf{Z}_i = (\mathbf{X}_i^T, \mathbf{Y}_i^T)^T$, where \mathbf{X}_i is observed with probability 1 and $\mathbf{Y}_i = (Y_{1,i}, Y_{2,i}, \dots, Y_{J,i})^T$ may be incompletely observed. The RMM process for subject i is described by Robins and Gill as follows. We start by observing \mathbf{X}_i . Then, we either observe one of $Y_{j,i}$ ($j = 1, \dots, J$) with probabilities $p_{j,i}(\mathbf{X}_i)$ ($j = 1, \dots, J$), respectively, or we quit, having only observed \mathbf{X}_i with probability $1 - \sum_{j=1}^J p_{j,i}(\mathbf{X}_i)$. Suppose we in fact observe $Y_{j_1,i}$. Now, at the next stage, we either observe one of $Y_{j,i}$ ($j = 1, \dots, j_1 - 1, j_1 + 1, \dots, J$) with probabilities $p_{j,i}(\mathbf{X}_i^T, Y_{j_1,i})$ ($j = 1, \dots, j_1 - 1, j_1 + 1, \dots, J$), respectively, or we quit, having only observed \mathbf{X}_i and $Y_{j_1,i}$ with probability $1 - \sum_{j \neq j_1} p_{j,i}(\mathbf{X}_i^T, Y_{j_1,i})$. Suppose that after m stages of this algorithm, we have observed $(\mathbf{X}_i^T, Y_{j_1,i}, Y_{j_2,i}, \dots, Y_{j_{m-1},i})$. At the next stage, we either observe one of $Y_{j,i}$ ($j \in \{1, \dots, J\} \setminus \{j_1, j_2, \dots, j_{m-1}\}$) with probabilities $p_{j,i}(\mathbf{X}_i^T, Y_{j_1,i}, \dots, Y_{j_{m-1},i})$ ($j \in \{1, \dots, J\} \setminus \{j_1, j_2, \dots, j_{m-1}\}$), respectively, or we quit with probability $1 - \sum_{j \neq j_1, \dots, j_{m-1}} p_{j,i}(\mathbf{X}_i^T, Y_{j_1,i}, \dots, Y_{j_{m-1},i})$.

Markov randomised monotone missingness (MRMM) is a special case of RMM in which the probability of observing a given variable conditional on the previous outcomes observed is independent of the order in which these variables were observed. Thus, for example, $p_{j,i}(\mathbf{X}_i^T, Y_{j_1,i}, Y_{j_2,i}) = p_{j,i}(\mathbf{X}_i^T, Y_{j_2,i}, Y_{j_1,i})$. Gill and Robins (1996) prove that any MAR mechanism representable as RMM is also representable as MRMM. The MRMM process when $J = 3$ is shown in Fig. 4.1.

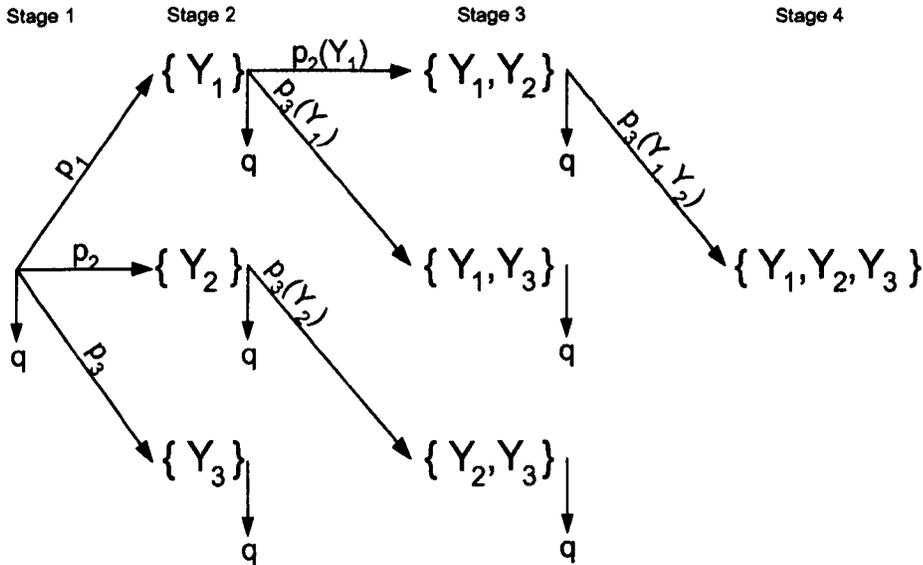


Figure 4.2: A Markov randomised monotone missingness process for longitudinal data

Notice that (omitting the subscript i), for example,

$$\begin{aligned} \mathbb{P}(Y_1, Y_2 \text{ both observed, } Y_3 \text{ missing}) &= [p_1(\mathbf{X}) p_2(\mathbf{X}^T, Y_1) + p_2(\mathbf{X}) p_1(\mathbf{X}^T, Y_2)] \\ &\quad \cdot [1 - p_3(\mathbf{X}^T, Y_1, Y_2)] \end{aligned}$$

where $p_1(\mathbf{X}) p_2(\mathbf{X}^T, Y_1)$ and $p_2(\mathbf{X}) p_1(\mathbf{X}^T, Y_2)$ are not constrained to be equal. Thus, the order in which the variables were observed is needed to estimate the probabilities $p_j(\cdot)$ —even in an MRMM process—but this order is never observed. Robins and Gill (1997) describe a method for estimating these probabilities, where the unobserved orderings are treated as missing data.

Intuitively, by allowing the partially-observed variables to be observed in a variety of different orderings, a non-monotone mechanism can be viewed (within a given ordering) as a dropout mechanism. In a situation (such as longitudinal repeated measures) when there exists only one plausible ordering (such as a temporal ordering), MRMM reduces to a very special case (shown in Fig. 4.2) where the probability of observing Y_3 , say, is dependent on Y_2 if and only if Y_2 has been observed. As Vansteelandt *et al.* (2007)

argue, it is implausible in most settings that the probability of observing Y_3 only depends on Y_2 if Y_2 happens to have been observed and that therefore, MAR is rarely a sensible assumption for non-monotone repeated measures. However, as a point of departure for sensitivity analyses it is useful to be aware of the form of this ignorable mechanism. Estimating the parameters of the mechanism shown in Fig. 4.2 is much more straightforward than in the general case (as shown in Fig. 4.1) as the order in which the variables were observed is always known.

Furthermore, Robins and Gill (1997) describe a statistical test of the hypothesis that, given that the mechanism is MAR, it is also RMM. They argue that if this hypothesis is not supported by the data, MAR should be ruled out even in situations when it might be deemed plausible *a priori*.

5

Simple methods

One of the reasons that missing data pose a problem is that they destroy the rectangular structure (or *balance*) of the dataset necessary for many statistical analyses (e.g. multiple linear regression, ANOVA). Because of this, most simple methods for handling missing data involve deriving a rectangular dataset from a non-rectangular one, either destructively or constructively, i.e. either by deleting incomplete lines of data or by imputing data in place of the missing values.

5.1 Complete case (CC) analysis

Removing all the subjects in the dataset except for those with complete data on all variables, and analysing the data that remain, is known as a *complete case* (CC) analysis. It is clearly inefficient, especially in a dataset with a large number of variables. Suppose, for example, that 20 variables are collected on 1,000 subjects, and that 5% of the data on each variable are missing. If the missing data are distributed uniformly across subjects and independently for each variable, even though only 5% of the data are missing, only $(.95)^{20} \cdot 100 = 36\%$ of the cases can be used in a CC analysis—less than 38% of the observed data.

Furthermore, except for the case when the missing data are MCAR, a CC analysis is biased. Assuming that the full-data analysis would have consisted of solving a score equation of the form

$$\sum_{i=1}^n \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \hat{\boldsymbol{\theta}}_n) = \mathbf{0}$$

then, the CC estimator of $\boldsymbol{\theta}$ is the one that solves

$$\sum_{i=1}^n \mathbf{1}(C_i = \infty) \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \hat{\boldsymbol{\theta}}_n^{\text{CC}}) = \mathbf{0} \quad (5.1.1)$$

The solution to (5.1.1) is consistent if and only if

$$\mathbf{E}[\mathbf{1}(C_i = \infty) \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0)] = \mathbf{0}$$

(Cox and Hinkley, 1974). Under the CCAR assumption,

$$\mathbf{E}[\mathbf{1}(C_i = \infty) \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0)] = \mathbf{E}\{\mathbf{E}[\mathbf{1}(C_i = \infty) \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) | \mathbf{Z}_i]\} \quad (5.1.2a)$$

$$= \mathbf{E}[\mathbf{P}(C_i = \infty | \mathbf{Z}_i) \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0)] \quad (5.1.2b)$$

$$= \mathbf{E}[\pi(\infty, \mathbf{Z}_i) \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0)] \quad (5.1.2c)$$

$$= \pi \mathbf{E}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0)] = \mathbf{0} \quad (5.1.2d)$$

since $\mathbf{E} [\mathbf{S}_{\theta}^F(\mathbf{Z}_i, \theta_0)] = \mathbf{0}$.

But the penultimate step relies on the CCAR assumption that $\pi(\infty, \mathbf{Z}_i) = \pi$, independently of \mathbf{Z}_i , which is not the case under CAR. Hence, CC estimators are inconsistent under CAR.

5.2 Inverse probability weighted complete case (IPWCC) analysis

The way in which the CC estimator is inconsistent motivates the *inverse probability weighted complete case* (IPWCC) estimator, the estimator which solves

$$\sum_{i=1}^n \frac{\mathbf{1}(C_i = \infty)}{\mathbb{P}(C_i = \infty | \mathbf{Z}_i)} \mathbf{S}_{\theta}^F(\mathbf{Z}_i, \hat{\theta}^{\text{IPWCC}}) = \mathbf{0}$$

First introduced by Horvitz and Thompson (1952), the idea of weighting each fully-observed subject by the probability of observing that subject is intuitively sensible. Informally, subjects who have only a probability of $\frac{1}{2}$ of being observed are weighted twice as much as those who are certain of being observed. We can think of these people as contributing twice: once for themselves, and once for their ‘twin’ who was not observed. This is analogous to using sampling weights in surveys with unequal sampling probabilities.

Following a similar argument to (5.1.2a-5.1.2d), we can show that

$$\mathbf{E} \left[\frac{\mathbf{1}(C_i = \infty)}{\mathbb{P}(C_i = \infty | \mathbf{Z}_i)} \mathbf{S}_{\theta}^F(\mathbf{Z}_i, \theta_0) \right] = \mathbf{0}$$

and the estimator is consistent under MAR, making IPWCC considerably better than CC estimator. The weighting has corrected the bias, but the inefficiency remains.

5.3 Single imputation methods

Since discarding incomplete lines of data results in a loss of precision, an alternative, popular approach is to impute data for the missing values and then analyse the augmented data as if they were the original fully-observed dataset.

The main single imputation procedures are:

- *Mean imputation*, where a missing $Z_{j,i}$ is replaced by the mean value of Z_j across all observed subjects, $(\sum_{i=1}^n R_{j,i} Z_{j,i}) / (\sum_{i=1}^n R_{j,i})$;
- *Regression imputation*, where a missing $Z_{j,i}$ is replaced by $\mathbb{E}(Z_{j,i} | \mathbf{Z}_i^{\text{obs}})$, where this conditional expectation is calculated using a regression model of Z_j on the variables included in $\mathbf{Z}_i^{\text{obs}}$, fitted to those subjects on whom all these variables were observed;
- *Stochastic regression imputation*, where a missing $Z_{j,i}$ is replaced by a random draw from the distribution of $Z_{j,i}$ given $\mathbf{Z}_i^{\text{obs}}$, estimated from the same regression model as above; and
- *Last observation carried forward*, used in repeated measures analyses. As the name suggests, LOCF simply replaces any missing value by the last observed value for that variable on that subject: each $Z_{t,i}$, for $t \geq D_i$ is replaced by $Z_{D_i-1,i}$.

Mean imputation can be biased—even under MCAR—depending on the analysis. For example, suppose that $\mathbf{Z}_i = (X_i, Y_i)^T$ and that the parameters of interest are those from the linear regression of Y on X . Suppose that there are missing data on Y alone, and that these data are missing completely at random. If we replace the missing Y -values by $(\sum_{i=1}^n R_i Y_i) / (\sum_{i=1}^n R_i)$ (which, under MCAR, has expectation equal to the marginal mean of Y), we shrink the estimate of the slope towards 0. Intuitively, this can be seen in Fig. 5.1.

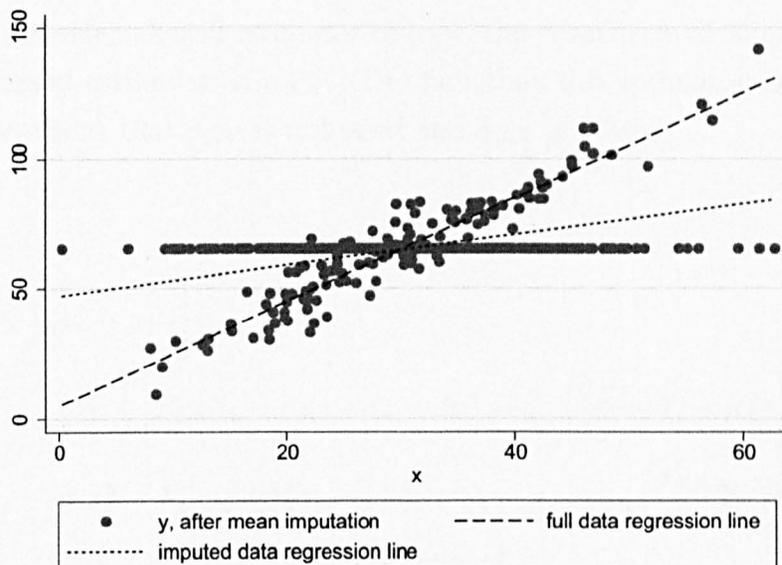


Figure 5.1: The bias introduced by mean imputation on the outcome in a linear regression model with outcomes missing completely at random

Regression imputation, too, can be biased under MCAR. Suppose that Y is fully-observed, X is missing completely at random for some subjects and we impute the missing X -values using their conditional expectation given Y as determined from the complete case regression of X on Y . We know that the parameters of this regression are consistently estimated in a complete case analysis since the missing data are MCAR. Thus, asymptotically, this is equivalent to setting \tilde{X}_i , the value of X in the imputed dataset equal to

$$R_i X_i + (1 - R_i) \mathbb{E}(X_i | Y_i) = R_i X_i + (1 - R_i) \left[\mu_X + \rho_{XY} \sqrt{\frac{\sigma_{XX}}{\sigma_{YY}}} (Y_i - \mu_Y) \right]$$

where μ_X , μ_Y , σ_{XX} , σ_{YY} are the true marginal means and variances of X and Y respectively, and ρ_{XY} is the true correlation between X and Y .

The slope parameter is estimated from the imputed dataset as

$$\frac{\hat{\sigma}_{XY}}{\hat{\sigma}_{XX}}$$

where $\hat{\sigma}_{XY}$ is a moment-based estimator of σ_{XY} , the covariance of X and Y , and $\hat{\sigma}_{XX}$ is a moment-based estimator of σ_{XX} . The fact that this estimator is biased follows from the observations that $\hat{\sigma}_{XY}$ is unbiased and $\hat{\sigma}_{XX}$ is biased:

$$\begin{aligned}
 \mathbb{E} \left[\left(\tilde{X}_i - \mu_X \right) (Y_i - \mu_Y) \right] &= \mathbb{E} \left\{ \left[R_i (X_i - \mu_X) + (1 - R_i) \rho_{XY} \sqrt{\frac{\sigma_{XX}}{\sigma_{YY}}} (Y_i - \mu_Y) \right] (Y_i - \mu_Y) \right\} \\
 &= \pi \mathbb{E} [(X_i - \mu_X) (Y_i - \mu_Y)] + (1 - \pi) \rho_{XY} \sqrt{\frac{\sigma_{XX}}{\sigma_{YY}}} \mathbb{E} [(Y_i - \mu_Y)^2] \\
 &= \pi \sigma_{XY} + (1 - \pi) \rho_{XY} \sqrt{\sigma_{XX} \sigma_{YY}} \\
 &= \sigma_{XY}
 \end{aligned}$$

where $\pi = \mathbb{E}(R_i | X_i, Y_i)$, which we assume to be independent of X_i and Y_i .

$$\begin{aligned}
 \mathbb{E} \left[\left(\tilde{X}_i - \mu_X \right)^2 \right] &= \mathbb{E} \left[R_i (X_i - \mu_X)^2 + (1 - R_i) \rho_{XY}^2 \frac{\sigma_{XX}}{\sigma_{YY}} (Y_i - \mu_Y)^2 \right] \\
 &= \pi \sigma_{XX} + (1 - \pi) \rho_{XY}^2 \sigma_{XX} \\
 &= \sigma_{XX} [1 - (1 - \pi) (1 - \rho_{XY}^2)] \\
 &\leq \sigma_{XX}
 \end{aligned}$$

with equality if and only if either $\pi = 1$ or $\rho_{XY}^2 = 1$, i.e. if and only if either there are no missing data or X and Y are either perfectly correlated or perfectly anti-correlated. Otherwise, σ_{XX} will be underestimated, and the slope parameter in the regression of Y on the imputed X -values will be overestimated.

This bias applies only to missing covariates. Regression imputation is consistent for

missing outcomes, although naïve estimates of precision from such an analysis will be biased.

LOCF is also, in general, biased even under MCAR (see Molenberghs *et al.*, 2004). The claim that LOCF always leads to conservative estimates is shown by these authors to be false. A paper by Shao and Zhong (2004) claims that

“... the LOCF one-way ANOVA test is actually asymptotically valid (that is, its asymptotic size is equal to the nominal size) in the special but important case where only two treatments are compared and the two treatment groups have the same number of patients, regardless of whether drop-out is informative or not.”

However, Carpenter *et al.* (2004) point out that the alleged validity applies *only* when the hypothesis being considered is that *at the last observed occasion*, the effect of both treatments is the same. It is not surprising that LOCF is valid in this scenario—indeed, as far as this hypothesis is concerned, no data are missing! Carpenter *et al.* (2004) argue that this hypothesis is of no clinical interest.

Furthermore, all naïve estimates of precision following non-stochastic imputation methods are biased, since the imputations are deterministic and are thus less variable than the (unobserved) observations they are replacing. What we have seen above is that this reduction in variability can also affect parameter estimates, depending on the rôle played by the imputed variable(s) in the analysis.

Stochastic regression imputation is an attempt both to correct the residual bias of regression imputation, and to correct the bias in the estimation of standard errors. It succeeds in the former but not the latter. There are two sources of uncertainty in a missing data imputation problem: that due to the original variance in the data, and that due to the missing data. Regression imputation takes into account neither of these (hence the shrinkage of σ_{XX} seen in the imputations above), and stochastic regression

imputation takes into account only the former, leading to consistent effect estimates, but biased estimates of precision. As we see in §6.7, it is possible to take both sources of uncertainty into account, and this can lead to valid inference.

6

Fully-parametric methods

In the next two chapters, we turn to more principled methods for handling coarsened or missing data. These methods are broadly divided into fully- and semiparametric methods. The choice between these can be based both on philosophical and practical considerations and is the subject of considerable debate in the current literature (Davidian *et al.*, 2005; Kang and Schafer, 2007).

6.1 The direct likelihood approach

Suppose we assume CAR and pose a parametric model for the coarsening mechanism as follows:

$$\mathbb{P}(C = c | \mathbf{Z}) = \pi(c, G_c(\mathbf{Z}), \boldsymbol{\xi})$$

where $\boldsymbol{\xi}$ is a set of parameters governing the coarsening mechanism. To make inference about $\boldsymbol{\theta}$ using the observed data, we must write the observed-data density $p_{C, G_c(\mathbf{Z})}(c, \mathbf{g}_c, \boldsymbol{\theta}, \boldsymbol{\xi})$ in terms of the full-data parameter $\boldsymbol{\theta}$.

$$\begin{aligned} p_{C, G_c(\mathbf{Z})}(c, \mathbf{g}_c, \boldsymbol{\theta}, \boldsymbol{\xi}) &= \int_{\{\mathbf{z}: G_c(\mathbf{z}) = \mathbf{g}_c\}} p_{C, \mathbf{Z}}(c, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\xi}) d\mathbf{z} \\ &= \int_{\{\mathbf{z}: G_c(\mathbf{z}) = \mathbf{g}_c\}} \mathbb{P}(C = c | \mathbf{Z} = \mathbf{z}, \boldsymbol{\xi}) p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z} \\ &= \pi(c, \mathbf{g}_c, \boldsymbol{\xi}) \int_{\{\mathbf{z}: G_c(\mathbf{z}) = \mathbf{g}_c\}} p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z} \end{aligned} \quad (6.1.1)$$

due to the CAR assumption. So we see that the CAR assumption means that the coarsening parameter $\boldsymbol{\xi}$ and the full-data parameter $\boldsymbol{\theta}$ separate, making likelihood methods simpler. If the function $\int_{\{\mathbf{z}: G_c(\mathbf{z}) = \mathbf{g}_c\}} p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z}$ could be maximised by a particular value $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}}$ would be a maximum-observed-likelihood estimator. In other words, when making inference about $\boldsymbol{\theta}$, provided that $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are *distinct* (i.e. that the parameter space of the full vector $(\boldsymbol{\theta}^T, \boldsymbol{\xi}^T)^T$ is the product of the individual parameter spaces), the part of the likelihood involving $\boldsymbol{\xi}$ can be ignored and the missingness mechanism need not be modelled. Within the likelihood framework, the CAR (or MAR) assumption, coupled with the distinct parameter assumption is called *ignorability*.

Although ignorability simplifies the task, it is still necessary to integrate (directly or indirectly) the full-data likelihood over the missing data. Anderson (1957) was the first to show how this could be achieved in the bivariate normal case. Suppose that only r out of n of the X -values are observed and that Y is fully-observed. By first noting that the parameters $(\mu_X, \mu_Y, \sigma_{XX}, \sigma_{XY}, \sigma_{YY})$ can be isomorphically mapped

onto $(\phi, \gamma, \sigma_{X|Y}, \mu_Y, \sigma_{YY})$, where ϕ , γ and $\sigma_{X|Y}$ are the parameters associated with the regression of X on Y , what Anderson showed was that in this case the observed-data likelihood factorises as:

$$p(Z_{\text{observed}}) = \left\{ \prod_{i=1}^n p(Y_i, \mu_Y, \sigma_{YY}) \right\} \left\{ \prod_{i=1}^r p(X_i|Y_i, \phi, \gamma, \sigma_{X|Y}) \right\} \quad (6.1.2)$$

The parameters separate, the two factors can be maximised separately, and a maximum likelihood solution is obtained. Furthermore, (6.1.2) can be generalised to any number of variables (see Little and Rubin, 2002, p. 145) if the missing data pattern is monotone. When the missing data pattern is non-monotone, an iterative algorithm of one sort or another is required. The Expectation-Maximisation (EM) algorithm is a popular method—not restricted to multivariate Gaussian data—and will be discussed in §6.2. For multivariate Gaussian data, by writing the model in terms of random effects (i.e. a *linear mixed model*), a method for maximising the observed-data likelihood under a flexible range of model restrictions is widely used and is discussed in §6.3. When this model is not expressed in terms of random effects, the same iterative estimation tools are used, and the model is known as the *multivariate linear model*.

Likelihood methods are not restricted to Gaussian data. Indeed, as (6.1.1) suggests, they can be conceived of whenever a parametric likelihood for the full data can be expressed. The *generalised linear mixed model* is an extension to non-Gaussian outcomes of the linear mixed model and will be discussed in §6.4. This is a *subject-specific* model. Other formulations include the Dale model (Dale, 1986) for categorical data, the Bahadur model (Bahadur, 1961) for binary data and the method proposed by Fitzmaurice and Laird (1993), also for binary data. These are examples of *marginal* models. Let us look more closely at this distinction.

Subject-specific versus marginal models for non-Gaussian data

We denote by $Y_{t,i}$ the t^{th} outcome on the i^{th} subject, where $1 \leq t \leq D_i - 1 \leq T$ and $1 \leq i \leq n$. Let $X_i, V_{1,i}, V_{2,i}, \dots, V_{p-1,i}$ be a set of p time-stationary covariates measured on subject i , where X_i is the covariate of interest (e.g. treatment group). If $T = 1$, then we fit an ordinary univariate regression model:

$$f[\mathbf{E}(Y_i | X_i, V_{1,i}, V_{2,i}, \dots, V_{p-1,i})] = \alpha + \beta X_i + \gamma_1 V_{1,i} + \dots + \gamma_{p-1} V_{p-1,i}$$

for some suitable link function $f(\cdot)$, such as the logit or probit function if Y is binary. However, when $T > 1$, we must take into account the fact that repeated measurements on the same subject are not independent. We can do this by introducing a random subject effect, U_i , into the linear predictor:

$$f[\mathbf{E}(Y_{t,i} | X_i, V_{1,i}, V_{2,i}, \dots, V_{p-1,i}, U_i)] = \alpha_t + \beta_t X_i + \gamma_{1,t} V_{1,i} + \dots + \gamma_{p-1,t} V_{p-1,i} + U_i$$

Often, interest lies in the treatment effect at the final timepoint. Looking at one specific subject, i , this is:

$$\begin{aligned} & \mathbf{E}(Y_{T,i} | X_i = 1, V_{1,i}, V_{2,i}, \dots, V_{p-1,i}, U_i) - \mathbf{E}(Y_{T,i} | X_i = 0, V_{1,i}, V_{2,i}, \dots, V_{p-1,i}, U_i) \\ &= f^{-1}(\alpha_T + \beta_T + \gamma_{1,T} V_{1,i} + \dots + \gamma_{p-1,T} V_{p-1,i} + U_i) \\ &\quad - f^{-1}(\alpha_T + \gamma_{1,T} V_{1,i} + \dots + \gamma_{p-1,T} V_{p-1,i} + U_i) \end{aligned} \quad (6.1.3)$$

We could fit this model to find an estimate of:

$$\begin{aligned} & f^{-1}(\alpha_T + \beta_T + \gamma_{1,T} V_{1,i} + \dots + \gamma_{p-1,T} V_{p-1,i} + U_i) \\ &\quad - f^{-1}(\alpha_T + \gamma_{1,T} V_{1,i} + \dots + \gamma_{p-1,T} V_{p-1,i} + U_i) \end{aligned}$$

and we would be doing a *subject-specific* analysis. In other words, we would be looking at the effect of X_i on $\mathbf{E}(Y_{T,i} | U_i)$, conditional on $V_{1,i}, \dots, V_{p-1,i}$. The key here is the fact that the expectation is conditional on the random effect U_i , and so the effect being measured is the effect on a particular subject, given his/her value of the random effect.

We might instead be interested in the effect of X_i on the marginal expectation $\mathbb{E}(Y_{T,i})$, conditional on $V_{1,i}, \dots, V_{p-1,i}$, in which case the analysis would be *marginal* or *population-averaged*. The equivalent of equation (6.1.3) is then:

$$\begin{aligned} \mathbb{E}_{U_i} [\mathbb{E}(Y_{T,i} | X_i = 1, V_{1,i}, V_{2,i}, \dots, V_{p-1,i}, U_i) - \mathbb{E}(Y_{T,i} | X_i = 0, V_{1,i}, V_{2,i}, \dots, V_{p-1,i}, U_i)] \\ = \mathbb{E}_{U_i} [f^{-1}(\alpha_T + \beta_T + \gamma_{1,T}V_{1,i} + \dots + \gamma_{p-1,T}V_{p-1,i} + U_i) \\ - f^{-1}(\alpha_T + \gamma_{1,T}V_{1,i} + \dots + \gamma_{p-1,T}V_{p-1,i} + U_i)] \quad (6.1.4) \end{aligned}$$

In general, (6.1.3) and (6.1.4) are not equal. For Gaussian data, $f(\cdot)$ is the identity function, the two U_i terms cancel, and this issue doesn't arise. Thus, for continuous, Gaussian outcomes $Y_{j,i}$, the effect of X_i on a particular subject, and the effect of X_i averaged over the whole population are equal. Another special case is that of Poisson data with Gaussian random effects when the log link is used. In this case the population-averaged and subject-specific effects differ only by a multiplicative offset (Young *et al.*, 2007). But in general we must decide which of the two types of effects is of interest.

In general, whether we choose to estimate marginal or subject-specific effects, additional computational methods are required to integrate and then to maximise (6.1.1), and it is to these methods that we turn our attention next. The approach can either be purely based on likelihood, or based on the likelihood via a Bayesian analysis. Frequentist methods (not based on likelihood) do not possess the ignorability property and thus must be adapted if they are to be valid under MAR. These are discussed further in Chapter 7.

6.2 Expectation-Maximisation algorithm

Under a general (non-monotone) missing data pattern, evaluating the integral in (6.1.1) analytically is often impossible. Orchard and Woodbury (1972) were the first to describe a method, which later became known as the Expectation-Maximisation (EM) algorithm, when the full data can be assumed to be from a multivariate normal distribution. The method has since been extended to a much wider range of full-data models (see Dempster *et al.*, 1977). The basic principle is as follows:

1. Start by computing an initial estimate $\hat{\boldsymbol{\theta}}^{(1)}$ of $\boldsymbol{\theta}$, e.g. from a complete case analysis, or one of the other simple methods already discussed. Within some loose regularity conditions, this can be both biased and inefficient, although the less biased and more efficient it is, the more quickly the algorithm is likely to converge.
2. **The E-step:** Using the current parameter estimate, $\hat{\boldsymbol{\theta}}^{(c)}$, calculate

$$q\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(c)}\right) = \mathbb{E} \left[\log p_{\mathbf{Z}}(\mathbf{z}) \mid \mathbf{z}^{\text{obs}}, \hat{\boldsymbol{\theta}}^{(c)} \right]$$

This is made easier if $\log p_{\mathbf{Z}}(\mathbf{z})$ is linear in \mathbf{z}^{mis} .

3. **The M-step:** Find $\hat{\boldsymbol{\theta}}^{(c+1)}$ that maximises $q\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(c)}\right)$ and return to step 2., iterating until convergence.

It has been shown (see Dempster *et al.*, 1977) that under certain regularity conditions, this algorithm always converges to the maximum observed-data likelihood estimate. The direct likelihood approach achieves this in one step: the observed-data likelihood is calculated, and then maximised. When using the EM algorithm, we use our current (incorrect) estimate of $\boldsymbol{\theta}$ to estimate the observed-data likelihood, and then we maximise this to obtain a better estimate of $\boldsymbol{\theta}$.

6.3 The linear mixed model and the multivariate normal linear model

In the multivariate normal case, assuming we're interested in the regression of the $(J \times 1)$ multivariate outcome \mathbf{Y}_i on a $(J \times p)$ matrix of covariates \mathbf{X}_i , we can write the full-data model as:

$$\mathbf{Y}_i \sim N_J(\mathbf{X}_i\boldsymbol{\alpha}, \mathbf{V})$$

independently and identically for each $i \in \{1, \dots, n\}$, with no restriction on \mathbf{V} . We call this the multivariate linear model (MLM).

Laird and Ware (1982) suggested the following formulation of a special case of the MLM:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \quad (6.3.1)$$

where $\boldsymbol{\beta}_i$ is a $(q \times 1)$ vector of *random effects* with distribution

$$\boldsymbol{\beta}_i \sim N_q(\mathbf{0}, \mathbf{G})$$

independently and identically for each i , with associated design matrix \mathbf{Z}_i , and

$$\boldsymbol{\varepsilon}_i \sim N_J(\mathbf{0}, \mathbf{R})$$

It follows from this definition that

$$\mathbf{Y}_i \sim N_J(\mathbf{X}_i\boldsymbol{\alpha}, \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \mathbf{R})$$

This is known as a *mixed model* since it contains a mixture of population fixed effects ($\boldsymbol{\alpha}$) and subject-specific random effects ($\boldsymbol{\beta}_i$). More specifically, since the expectation of the vector of repeated measurements is assumed to be a *linear* function of these parameters, the model is known as the *linear mixed model*. This model is discussed extensively by Verbeke and Molenberghs (1997), Diggle *et al.* (2002) and Brown and

Prescott (2006).

Often, the restriction $\mathbf{R} = \sigma^2 \mathbf{I}_J$ is imposed, i.e. that, conditional on β_i , the repeated observations on subject i are independent. If sufficiently many random effects are included however, it is possible not to impose any restriction on the implied structure for \mathbf{V} , and thus the linear mixed model formulation is equivalent to the multivariate linear model, with the sole restriction that the variance-covariance matrix is the same for each subject, independently of the covariates. This restriction can be relaxed to allow for different variance-covariance matrices in different groups of subjects (*complex variation*) but it is still the case that the matrix cannot vary continuously with continuous covariates. Of course, a less general set of random effects can be posited, whence emerges the flexibility of this formulation. Alternatively, varying the number and structure of the random effects can be viewed directly as imposing a structure on \mathbf{V} .

Even when the data are incomplete and non-monotone, the observed-data likelihood can be maximised either by maximum likelihood (ML) or *restricted* maximum likelihood (REML) (see Patterson and Thompson, 1971), using iterative Newton-Raphson procedures. Asymptotically, ML and REML estimates converge, but when p is not negligible compared with n , the estimates can disagree, the difference occurring in the estimates of variance and covariance parameters. For these parameters, the small-sample bias is, in general, smaller for REML estimates than for ML estimates (Verbeke and Molenberghs, 1997). However, since both ML and REML are based on the likelihood, ignorability under MAR applies under both procedures.

6.4 The generalised linear mixed model

For univariate data, linear models for Gaussian data were extended to *generalised linear models* for non-Gaussian data by Nelder and Wedderburn (1972) and McCullagh and Nelder (1989). In a generalised linear model, Y_i is assumed to come from a distribution

(not necessarily Gaussian) belonging to the exponential family of distributions, i.e. that the density function of Y_i satisfies

$$p_Y(y_i, \psi_i, \phi) = \exp \{a(\phi) [y_i \psi_i - g(\psi_i) + h(y_i)] + b(\phi, y_i)\}$$

where ψ_i is the parameter of interest, ϕ is a nuisance parameter and $a(\phi) > 0$. A *link function* $f(\cdot)$ relates ψ_i to the covariates such that

$$f(\psi_i) = \mathbf{X}_i \boldsymbol{\alpha}$$

The generalised linear model and the linear mixed model of §6.3 can be combined to form the *generalised linear mixed model* (GLMM), where, conditional on random effects $\boldsymbol{\beta}_i$ the elements $Y_{j,i}$ of \mathbf{Y}_i are independent observations from a density satisfying

$$p_Y(y_{j,i}, \psi_{j,i}, \phi) = \exp \{a(\phi) [y_{j,i} \psi_{j,i} - g(\psi_{j,i}) + h(y_{j,i})] + b(\phi, y_{j,i})\}$$

where

$$f(\psi_{j,i}) = \mathbf{X}_{j,i} \boldsymbol{\alpha} + \mathbf{Z}_{j,i} \boldsymbol{\beta}_i$$

and

$$\boldsymbol{\beta}_i \sim N_q(\mathbf{0}, \mathbf{G})$$

Unlike the linear mixed model, there is no general solution to the problem of integrating these densities over the distribution of the random effects. Thus, approximate techniques for finding maximum likelihood estimates are needed. For more details, see Breslow and Clayton (1993), Engel and Keen (1994), and Molenberghs and Verbeke (2005).

6.5 Hierarchical generalised linear models

An alternative approach, using Laplace approximations to avoid the integration problem, is a method based on the hierarchical likelihood proposed by Lee and Nelder (1996). In this method, the random effects are assumed to have exponential family distributions conjugate to those of the outcome variable (such as normal-normal, Poisson-gamma, binomial-beta). The authors propose an efficient algorithm for finding approximate solutions, making the method an attractive candidate for sensitivity analyses when the number of models to be fitted is higher than would be feasible under the more computationally-intensive GLMM approach (see Yun *et al.*, 2007).

6.6 Bayesian methods

Given on the one hand the attractive property of ignorability of the missingness model afforded by likelihood methods under MAR, coupled with the computational difficulty of obtaining such maximum likelihood estimates in practice on the other, Bayesian methods which approximate likelihood procedures are a useful tool. For example, Clayton (1996) and Zeger and Karim (1991) both suggest Bayesian solutions to the estimation of GLMMs. The distinct parameter condition now translates to the assumption that the prior distributions for the full-data parameters (θ) and the parameters of the coarsening mechanism (ξ) are independent.

Bayesian approaches are also useful in MNAR sensitivity analyses, where the MAR assumption is relaxed, and parameters governing the dependence of the selection model on the unobserved data are introduced. Given that information on the values of these parameters is not contained in the data, subjective priors for the sensitivity parameters are naturally introduced.

6.7 Multiple imputation

6.7.1 Motivation

Integrating the function $\int_{\{\mathbf{z}: G_c(\mathbf{z})=\mathbf{g}_c\}} p_{\mathbf{z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z}$ in (6.1.1) and then maximising with respect to $\boldsymbol{\theta}$, as we have noted already in this chapter, can be very difficult to do analytically. The EM algorithm and Bayesian methods can be used to overcome this problem; multiple imputation is another solution.

Based on the principle that full-data densities are easier to deal with than observed-data densities, multiple imputation (MI), (first suggested by Rubin, 1978) is a method in which (provided the CAR assumption holds) only the full-data model need be considered, but implemented in such a way that the inference about $\boldsymbol{\theta}$ is valid. It also extends naturally to CNAR mechanisms, a property which sets it apart from the EM algorithm, and which we explore in Chapter 11.

Suppose that the true value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$ were known to us, along with the true density of the full data given the observed data, $p_{\mathbf{z}|C, G_C(\mathbf{z})}(\mathbf{z}|c, \mathbf{g}_c, \boldsymbol{\theta}_0)$. Then, given the observed data $\{(C_i, G_{C_i}) : i = 1, \dots, n\}$, the full data $\{Z_i : i = 1, \dots, n\}$ could be generated using $p_{\mathbf{z}|C, G_C(\mathbf{z})}(\mathbf{z}|c, \mathbf{g}_c, \boldsymbol{\theta}_0)$.

Of course, $\boldsymbol{\theta}_0$ is not known to us. By definition, the observed data are generated using the true observed-data likelihood. The only difference is that, assuming that the posited density $p_{\mathbf{z}|C, G_C(\mathbf{z})}(\mathbf{z}|c, \mathbf{g}_c, \boldsymbol{\theta})$ be of the correct form, if we were to try to generate $\{Z_i : i = 1, \dots, n\}$ using $p_{\mathbf{z}|C, G_C(\mathbf{z})}(\mathbf{z}|c, \mathbf{g}_c, \boldsymbol{\theta})$, we would substitute an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ for $\boldsymbol{\theta}$, as opposed to substituting the true value, $\boldsymbol{\theta}_0$.

For this reason, imputed datasets cannot have a distribution identical to that of the full data. An adjustment is therefore required to any inference based on imputed data. One way of making this adjustment easier, is to impute the data, not once, but several times. Hence the name: *multiple* imputation.

6.7.2 A formal description

In order to be able to describe the work of Chapters 8 and 9, a rigorous formulation of MI is needed, such as that adopted by Tsiatis (2006), Wang and Robins (1998) and Robins and Wang (2000), where the frequentist properties of MI estimators are derived.

Given some initial estimator of θ , (more details to follow), for each observed-data point $[C_i, G_{C_i}(\mathbf{Z}_i)]$, we sample at random from the conditional distribution, $p_{\mathbf{Z}|C, G_C(Z)}(\mathbf{z}|c, \mathbf{g}_c)$, m times to obtain

$$\{\mathbf{Z}_{ij} : i = 1, \dots, n, j = 1, \dots, m\}$$

The j th estimator, $\hat{\theta}_{nj}^*$, is obtained by solving the full-data likelihood equation

$$\sum_{i=1}^n \mathbf{S}_{\theta}^F(\mathbf{Z}_{ij}, \hat{\theta}_{nj}^*) = \mathbf{0}$$

That is, we use the data from the j th imputed set and treat them as if they were full-data to obtain the full-data maximum-likelihood estimator, $\hat{\theta}_{nj}^*$. Then, the proposed multiple imputation estimator is

$$\hat{\theta}_n^* = m^{-1} \sum_{j=1}^m \hat{\theta}_{nj}^*$$

Rubin argues that, under appropriate conditions, this estimator is consistent and asymptotically normal. That is,

$$n^{\frac{1}{2}} (\hat{\theta}_n^* - \theta_0) \xrightarrow{\mathcal{D}} N(0, \Sigma^*)$$

Furthermore, he suggests that the asymptotic variance Σ^* be estimated by

$$\hat{\Sigma}^* = m^{-1} \sum_{j=1}^m \left[n^{-1} \sum_{i=1}^n -\frac{\partial \mathbf{S}_{\theta}^F(\mathbf{Z}_{ij}, \hat{\theta}_{nj}^*)}{\partial \theta^T} \right]^{-1} + \left(\frac{m+1}{m} \right) n \sum_{j=1}^m \frac{(\hat{\theta}_{nj}^* - \hat{\theta}_n^*)(\hat{\theta}_{nj}^* - \hat{\theta}_n^*)^T}{m-1} \quad (6.7.1)$$

That is, an average of the estimates of the full-data asymptotic variance, using the inverse of the full-data observed information matrix over the imputed full-data sets, plus the sample variance of the imputation estimators multiplied by a ‘finite m ’ correction factor, gives us the asymptotic variance of the MI estimator.

The issue remains of how to obtain the initial estimator of θ . There are two approaches:

1. **Improper imputation:** This is a frequentist approach in which an initial estimator $\hat{\theta}_n^I$ is obtained from the coarsened data, and the imputations \mathbf{Z}_{ij} ($\hat{\theta}_n^I$) are obtained by sampling from $p_{\mathbf{Z}|C, G_C(\mathbf{Z})}(\mathbf{z}|c, \mathbf{g}_c, \hat{\theta}_n^I)$.
2. **Proper imputation:** This is a Bayesian approach (and the one advocated by Rubin) in which the data are generated from the *predictive distribution*

$$p_{\mathbf{Z}|C, G_C(\mathbf{Z})}[\mathbf{z}|C_i, G_{C_i}(\mathbf{Z}_i)] = \int p_{\mathbf{Z}|C, G_C(\mathbf{Z})}(\mathbf{z}|C_i, G_{C_i}(\mathbf{Z}_i), \theta) p_{\theta|C, G_C(\mathbf{Z})}[\theta|C_i, G_{C_i}(\mathbf{Z}_i)] d\mu_{\theta}(\theta) \quad (6.7.2)$$

where $p_{\theta|C, G_C(\mathbf{Z})}[\theta|C_i, G_{C_i}(\mathbf{Z}_i)]$ is the Bayesian posterior distribution of θ given the observed data.

6.7.3 Improper vs. proper imputation

Improper imputation

We shall assume that the initial estimator is RAL and inefficient. That is,

$$n^{\frac{1}{2}} \left(\hat{\theta}_n^I - \theta_0 \right) = n^{-\frac{1}{2}} \sum_{i=1}^n q[C_i, G_{C_i}(\mathbf{Z}_i)] + o_p(1)$$

where $q[C_i, G_{C_i}(\mathbf{Z}_i)]$ is the i th influence function of the estimator $\hat{\theta}_n^I$.

It can be shown that

$$q [C_i, G_{C_i} (\mathbf{Z}_i)] = \varphi_{\text{eff}} [C_i, G_{C_i} (\mathbf{Z}_i)] + h [C_i, G_{C_i} (\mathbf{Z}_i)] \quad (6.7.3)$$

where $\varphi_{\text{eff}} [C_i, G_{C_i} (\mathbf{Z}_i)]$ is the efficient influence function defined by

$$\varphi_{\text{eff}} [C_i, G_{C_i} (\mathbf{Z}_i)] = [\mathbf{I}_{\theta\theta} (\boldsymbol{\theta}_0)]^{-1} \mathbf{S}_{\theta} [C_i, G_{C_i} (\mathbf{Z}_i)]$$

and

$$\mathbb{E} \{ \varphi_{\text{eff}} [C_i, G_{C_i} (\mathbf{Z}_i)] h^T [C_i, G_{C_i} (\mathbf{Z}_i)] \} = \mathbf{0}^{q \times q}$$

Tsiatis (2006) proves that

Theorem 6.1 (Variance of improper MI estimator).

$$n^{\frac{1}{2}} (\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \text{N} (0, \Sigma^*)$$

where

$$\begin{aligned} \Sigma^* = & [\mathbf{I}_{\theta\theta} (\boldsymbol{\theta}_0)]^{-1} + m^{-1} [\mathbf{I}_{\theta\theta}^F (\boldsymbol{\theta}_0)]^{-1} [\mathbf{I}_{\theta\theta}^F (\boldsymbol{\theta}_0) - \mathbf{I}_{\theta\theta} (\boldsymbol{\theta}_0)] [\mathbf{I}_{\theta\theta}^F (\boldsymbol{\theta}_0)]^{-1} \\ & + [\mathbf{I}_{\theta\theta}^F (\boldsymbol{\theta}_0)]^{-1} [\mathbf{I}_{\theta\theta}^F (\boldsymbol{\theta}_0) - \mathbf{I}_{\theta\theta} (\boldsymbol{\theta}_0)] \text{Var} \{ h [C_i, G_{C_i} (\mathbf{Z}_i)] \} \cdot \\ & [\mathbf{I}_{\theta\theta}^F (\boldsymbol{\theta}_0) - \mathbf{I}_{\theta\theta} (\boldsymbol{\theta}_0)] [\mathbf{I}_{\theta\theta}^F (\boldsymbol{\theta}_0)]^{-1} \quad (6.7.4) \end{aligned}$$

and $h [C_i, G_{C_i} (\mathbf{Z}_i)]$ is as given in (6.7.3).

Remark 6.1. The asymptotic variance of the most efficient estimator of $\boldsymbol{\theta}$ (i.e. the estimator that arises from maximising the observed-data likelihood) is $[\mathbf{I}_{\theta\theta} (\boldsymbol{\theta}_0)]^{-1}$. Clearly, there is a loss of efficiency with the use of multiple imputation.

Remark 6.2. We see from the second term in (6.7.4) that the asymptotic variance decreases as the number of imputations, m , increases.

Remark 6.3. If $\hat{\boldsymbol{\theta}}_n^I$ were an efficient estimator, i.e. if $q [C_i, G_{C_i} (\mathbf{Z}_i)] = \varphi_{\text{eff}} [C_i, G_{C_i} (\mathbf{Z}_i)]$, then we see that the third term in (6.7.4) vanishes. In this case, the second term represents the loss of efficiency due to multiple imputation. This vanishes as $m \rightarrow \infty$.

Remark 6.4. The variance of the initial estimator $\hat{\theta}_n^I$ is $[\mathbf{I}_{\theta\theta}(\theta_0)]^{-1} + \text{Var}\{h[C_i, G_{C_i}(\mathbf{Z}_i)]\}$. Comparing this with (6.7.4) as $m \rightarrow \infty$ yields a difference of

$$\text{Var}\{h[C_i, G_{C_i}(\mathbf{Z}_i)]\} - [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} [\mathbf{I}_{\theta\theta}^F(\theta_0) - \mathbf{I}_{\theta\theta}(\theta_0)] \text{Var}\{h[C_i, G_{C_i}(\mathbf{Z}_i)]\} \cdot [\mathbf{I}_{\theta\theta}^F(\theta_0) - \mathbf{I}_{\theta\theta}(\theta_0)] [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1}$$

which can be shown to be positive definite. Therefore, if the number of imputations is sufficiently large, the multiple imputation estimator will be more efficient than the initial estimator.

Proposition 6.2. *The expression on the RHS of (6.7.4) can be rewritten as*

$$[\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} + \left(\frac{m+1}{m}\right) [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} [\mathbf{I}_{\theta\theta}^F(\theta_0) - \mathbf{I}_{\theta\theta}(\theta_0)] [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} + [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} [\mathbf{I}_{\theta\theta}^F(\theta_0) - \mathbf{I}_{\theta\theta}(\theta_0)] \text{Var}\{q[C_i, G_{C_i}(\mathbf{Z}_i)]\} \cdot [\mathbf{I}_{\theta\theta}^F(\theta_0) - \mathbf{I}_{\theta\theta}(\theta_0)] [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} \quad (6.7.5)$$

Tsiatis (2006) then goes on to prove the following theorem:

Theorem 6.3 (Rubin's variance formula for improper MI). *The expression on the RHS of (6.7.1) is an asymptotically unbiased estimator of*

$$[\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} + \left(\frac{m+1}{m}\right) [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} [\mathbf{I}_{\theta\theta}^F(\theta_0) - \mathbf{I}_{\theta\theta}(\theta_0)] [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} \quad (6.7.6)$$

Remark 6.5. This is precisely the 1st two terms in (6.7.5) and hence Rubin's variance formula, when used with improper multiple imputation will tend to underestimate variances, leading to anti-conservative inferences.

Proper imputation

Let us assume that the sample size is large enough to approximate the posterior distribution of $\boldsymbol{\theta}$ by the sampling distribution of $\hat{\boldsymbol{\theta}}_n^I$, the initial estimator of $\boldsymbol{\theta}$ used for the improper imputation. Therefore, mimicking the Bayesian proper imputation, at the j th imputation, we sample $\boldsymbol{\theta}^{(j)}$ from

$$N\left(\hat{\boldsymbol{\theta}}_n^I, \frac{\hat{\text{Var}}\{q[C_i, G_{C_i}(\mathbf{Z}_i)]\}}{n}\right)$$

and then randomly choose \mathbf{Z}_{ij} from the conditional distribution with conditional density $p_{\mathbf{Z}|C, G_C}(\mathbf{z}|C_i, G_{C_i}(\mathbf{Z}_i), \boldsymbol{\theta}^{(j)})$. The j th imputed estimator is the solution to the equation

$$\sum_{i=1}^n \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_{ij}(\boldsymbol{\theta}^{(j)}), \hat{\boldsymbol{\theta}}_{nj}^*) = 0$$

and the final estimator is

$$\hat{\boldsymbol{\theta}}_n^* = m^{-1} \sum_{j=1}^m \hat{\boldsymbol{\theta}}_{nj}^*$$

Tsiatis (2006) proves the following results:

Theorem 6.4 (Variance of proper MI estimator).

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(0, \Sigma^*)$$

where

$$\begin{aligned} \Sigma^* &= [\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0)]^{-1} + \left(\frac{m+1}{m}\right) [\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0)]^{-1} [\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) - \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] [\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0)]^{-1} \\ &\quad + \left(\frac{m+1}{m}\right) [\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0)]^{-1} [\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) - \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \text{Var}\{q[C_i, G_{C_i}(\mathbf{Z}_i)]\} \\ &\quad \cdot [\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) - \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] [\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0)]^{-1} \quad (6.7.7) \end{aligned}$$

Theorem 6.5 (Rubin's variance formula for proper MI). *The expression on the RHS*

of (6.7.1) is an asymptotically unbiased estimator of

$$\begin{aligned} & [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} + \left(\frac{m+1}{m}\right) [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} [\mathbf{I}_{\theta\theta}^F(\theta_0) - \mathbf{I}_{\theta\theta}(\theta_0)] [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} \\ & \quad + \left(\frac{m+1}{m}\right) [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} [\mathbf{I}_{\theta\theta}^F(\theta_0) - \mathbf{I}_{\theta\theta}(\theta_0)] \\ & \quad \cdot \text{Var}\{q[C_i, G_{C_i}(\mathbf{Z}_i)]\} [\mathbf{I}_{\theta\theta}^F(\theta_0) - \mathbf{I}_{\theta\theta}(\theta_0)] [\mathbf{I}_{\theta\theta}^F(\theta_0)]^{-1} \end{aligned}$$

Remark 6.6. This is precisely (6.7.7) and hence Rubin's variance formula, when used with proper multiple imputation correctly estimates the asymptotic variance.

By comparing (6.7.5) with (6.7.7) we see that improper MI (where the initial estimator is fixed across imputations) results in a more efficient estimator than proper MI (where the initial estimator is sampled from $N\left(\hat{\theta}_n^I, \frac{\text{Var}\{q[C_i, G_{C_i}(\mathbf{Z}_i)]\}}{n}\right)$ for each imputation). As $m \rightarrow \infty$, however, the asymptotic variance of the proper MI estimator converges to that of the improper MI estimator.

Rubin's variance estimator underestimates the asymptotic variance when used with improper imputation but correctly estimates the asymptotic variance when used with proper imputation. By 'correctly estimates' we mean that the variance estimator converges in expectation to the asymptotic variance, i.e. that it is asymptotically unbiased for any fixed m . As $m \rightarrow \infty$, however, Rubin's estimator is also consistent for improper MI.

Consistent estimators can be derived in other ways for either of the MI estimators (and Tsiatis (2006) explicitly suggests some), but the advantage that Rubin's variance formula (used with proper MI) has over other estimators is its simplicity: it is easy to implement and can be applied to a large range of situations, without need for adaptation. This is reflected in the fact that routines for the implementation of MI are now available in most standard statistical software packages.

One feature of Tsiatis's formulation is that the initial estimator is always of θ , i.e. the

full parameter vector. When we use MI in practice, this is often not the case. For example, consider the bivariate normal regression example discussed in §5.3. In this example, supposing that the parameters of interest are those from the regression of Y on X : α , β and $\sigma_{Y|X}$, we would impute the missing X -values using the estimated parameters $\hat{\phi}$, $\hat{\gamma}$ and $\hat{\sigma}_{X|Y}$ from the regression of X on Y , and then use the imputed datasets to make inference about α , β and $\sigma_{Y|X}$. Of course, it is always possible to reformulate into Tsiatis's formulation, by using the relationships between the parameters. For example,

$$\hat{\alpha} = \hat{\mu}_Y - \frac{\hat{\gamma}\hat{\sigma}_{YY}}{\hat{\gamma}^2\hat{\sigma}_{YY} + \hat{\sigma}_{X|Y}} \left(\hat{\phi} + \hat{\gamma}\hat{\mu}_Y \right)$$

6.7.4 Multiple imputation using chained equations (MICE)

Thus far, in our discussion of multiple imputation, we have assumed that a parametric model for the joint distribution of the full data and the joint distribution of the coarsened data can be specified. There are many situations in which this might not be feasible, such as in a large dataset with a mixture of continuous, binary and categorical variables and/or when the missingness is non-monotone. An approach to multiple imputation which does not require specification of such joint distributions is *multiple imputation using chained equations* (MICE), first suggested by van Buuren *et al.* (1999). This approach works by first filling in the missing observations in an *ad hoc* fashion using randomly sampled observed values. Then, a univariate regression model is fitted to the first variable conditional on all the others, after discarding the *ad hoc* imputations for the first variable and the missing values are properly imputed based on this model. Next, the *ad hoc* imputations are discarded from the second variable and the second variable is regressed on all other variables. The missing values for the second variable are then imputed using this second regression model. The process continues until each variable in turn has had its *ad hoc* imputations replaced by the conditional regression imputations. This completes the first cycle. The process is repeated until a fixed number of cycles have been completed. In the second cycle, the imputed values from the first cycle are used instead of the *ad hoc* imputations, but

these too are discarded whenever the variable in question is the *outcome* variable in the regression, i.e. each univariate regression has on its left-hand side only observed values, but may include imputed values on the right-hand side. The imputed dataset at this final cycle constitutes the first imputed dataset in the MI procedure. A total of m imputed datasets are constructed in the same way, where for each imputation the process starts afresh with a new set of *ad hoc* imputations, and continues for the designated number of cycles.

This method is practically very attractive as it can deal with missing data on a large number of variables with different univariate distributions and without requiring that the missing data pattern be monotone. However, a full theoretical argument for the validity of this method has not been presented to date. Indeed, it is unlikely that such a proof exists except in the multivariate normal case, since a collection of univariate regression models—not all linear—cannot correspond to a well-defined joint distribution. The imputation model is inherently *uncongenial*, i.e. the stationary distribution to which the Gibbs sampler attempts to converge does not exist. However, simulation studies suggest that the bias caused by the uncongeniality is likely to be small in practice (Gelman and Raghunathan, 2001; van Buuren *et al.*, 2006).

7

Semiparametric methods

7.1 Introduction

All the methods described in the previous chapter are likelihood methods, valid under MAR, when the model is correctly specified. The computational complexity of the numerical algorithms needed to find approximations to these estimates is one disadvantage, but with modern processors, this problem is fast diminishing in many cases.

Another disadvantage, which cannot be overcome by increased computing power, is the necessity to specify the full-data likelihood fully-parametrically. For example, a linear mixed model assumes that the full data have a multivariate normal structure. Missing data issues aside, it is now well-known that many common statistical methods that assume normality are in fact extremely robust to violations of this assumption, e.g. the t -test (Rasch and Guiard, 2004). However, in missing data problems, the multivariate normality assumption plays a far greater role in the subsequent inference, because, not only is it assumed that each variable is marginally normally distributed, but also that the conditional distribution of each variable given any other variable is normally distributed. For example, in a repeated measures setting (such as the RECORD trial) with T timepoints, suppose that a patient drops out after just one post-baseline measurement of the outcome variable, then, at each subsequent timepoint, the linear mixed model implicitly assumes that the present observation, conditional on the past observations, is normally distributed. In an analysis based on the final timepoint, any effect due to the violation of the normality assumption is compounded $T - 1$ times for this patient.

Things are even worse for marginal discrete-data models, where there are problems in addition to the dependence on intractable modelling assumptions. The Bahadur model (Bahadur, 1961), for example, being based on correlations (pairwise and higher order) is easy to write down, but difficult to conceptualise, given our lack of intuition for (particularly higher order) correlations between binary variables. Furthermore, the model places heavy restrictions on the parameter space, and this problem intensifies as the number of timepoints increases. Intuition for whether these restrictions are plausible in practice is usually impossible to acquire in all but the simplest of settings, and in the presence of missing data the assumptions cannot be fully tested.

A less restrictive *semiparametric* approach is therefore an attractive alternative. The class of semiparametric models is vast and in §3.7 we defined two important subclasses, \mathcal{I} and \mathcal{R} . In §7.2 we give an example (the mean score method) of a method belonging to \mathcal{I} .

An important example of a method belonging to \mathcal{R} is Generalised Estimating Equations (GEE) (Liang and Zeger, 1986) and this is introduced in §7.3. However, since this is *not* a likelihood procedure, in general it is only valid under MCAR, and some further modelling of the missing value mechanism is necessary for the analysis to be valid under MAR. Many extensions of GEE for MAR mechanisms have been proposed, and we describe these in §7.4 and §7.5 of this chapter. In §7.6 we look at the efficiency of semiparametric methods under MAR according to the general framework proposed by James Robins and his colleagues. In particular, we describe the estimator similar to the GEE estimator proposed by Robins and Rotnitzky (1995), which is both consistent under MAR and semiparametric-efficient in a class which contains GEE and its variants.

Given our interest (in Chapters 12 and 13) in binary data, when discussing non-Gaussian outcomes, we will focus on binary outcomes, but the theory applies more generally.

7.2 Mean score method: an example of a method belonging to \mathcal{I}

Recall that the observed-data density (6.1.1) under the CAR assumption can be written as

$$p_{C, G_C(\mathbf{Z})}(c, \mathbf{g}_c, \boldsymbol{\theta}, \boldsymbol{\xi}) = \pi(c, \mathbf{g}_c, \boldsymbol{\xi}) \int_{\{\mathbf{z}: G_c(\mathbf{z}) = \mathbf{g}_c\}} p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z} \quad (7.2.1)$$

Thus the observed-data score function can be written as

$$\begin{aligned} \mathbf{S}_{\boldsymbol{\theta}}[C, G_C(\mathbf{Z})] &= \frac{\partial}{\partial \boldsymbol{\theta}} \log [p_{C, G_C(\mathbf{Z})}(c, \mathbf{g}_c, \boldsymbol{\theta}, \boldsymbol{\xi})] \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \log \left[\pi(c, \mathbf{g}_c, \boldsymbol{\xi}) \int_{\{\mathbf{z}: G_c(\mathbf{z}) = \mathbf{g}_c\}} p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z} \right] \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \log [\pi(c, \mathbf{g}_c, \boldsymbol{\xi})] + \frac{\partial}{\partial \boldsymbol{\theta}} \log \left[\int_{\{\mathbf{z}: G_c(\mathbf{z}) = \mathbf{g}_c\}} p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z} \right] \end{aligned}$$

$$\begin{aligned}
&= 0 + \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \int_{\{\mathbf{z}: G_c(\mathbf{z})=\mathbf{g}_c\}} p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z}}{\int_{\{\mathbf{z}: G_c(\mathbf{z})=\mathbf{g}_c\}} p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z}} \\
&= \frac{\int_{\{\mathbf{z}: G_c(\mathbf{z})=\mathbf{g}_c\}} \frac{\partial}{\partial \boldsymbol{\theta}} p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z}}{\int_{\{\mathbf{z}: G_c(\mathbf{z})=\mathbf{g}_c\}} p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z}} \\
&= \frac{\int_{\{\mathbf{z}: G_c(\mathbf{z})=\mathbf{g}_c\}} \frac{\partial}{\partial \boldsymbol{\theta}} \log [p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta})] p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z}}{\int_{\{\mathbf{z}: G_c(\mathbf{z})=\mathbf{g}_c\}} p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z}} \\
&= \frac{\mathbb{E} [\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}, \boldsymbol{\theta}) | C, G_C(\mathbf{Z})]}{\int_{\{\mathbf{z}: G_c(\mathbf{z})=\mathbf{g}_c\}} p_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z}}
\end{aligned}$$

Thus the observed-data score equation is

$$\sum_{i=1}^n \frac{\mathbb{E} [\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}) | C_i, G_{C_i}(\mathbf{Z}_i)]}{\int_{\{\mathbf{z}: G_{c_i}(\mathbf{z})=\mathbf{g}_{c_i}\}} p_{\mathbf{Z}_i}(\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z}} = \mathbf{0}$$

or, equivalently,

$$\sum_{i=1}^n \mathbb{E} [\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}) | C_i, G_{C_i}(\mathbf{Z}_i)] = \mathbf{0}$$

For missing data problems, this can be written as

$$\sum_{i=1}^n \left\{ \mathbb{1}(C_i = \infty) \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}) + \mathbb{1}(C_i < \infty) \mathbb{E} [\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}) | \mathbf{Z}_i^{\text{obs}}] \right\} = \mathbf{0} \quad (7.2.2)$$

Solving this observed-data score equation to find estimates of $\boldsymbol{\theta}$ is exactly equivalent to maximising (7.2.1) and hence is a fully-parametric approach. But this requires that $\mathbb{E} [\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}) | \mathbf{Z}_i^{\text{obs}}]$ be determined analytically as a function of $\mathbf{Z}_i^{\text{obs}}$ and $\boldsymbol{\theta}$.

Suppose instead that we solve

$$\sum_{i=1}^n \left\{ \mathbb{1}(C_i = \infty) \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}) + \mathbb{1}(C_i < \infty) \hat{\mathbb{E}} [\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}) | \mathbf{Z}_i^{\text{obs}}] \right\} = \mathbf{0} \quad (7.2.3)$$

where $\hat{\mathbb{E}}[\mathbf{S}_\theta^F(\mathbf{Z}_i, \theta) | \mathbf{Z}_i^{\text{obs}}]$ is estimated semiparametrically (or non-parametrically), then the resulting estimator belongs to \mathcal{I} .

Reilly and Pepe (1995) suggested that when \mathbf{Z}_i consists entirely of discrete random variables, $\hat{\mathbb{E}}[\mathbf{S}_\theta^F(\mathbf{Z}_i, \theta) | \mathbf{Z}_i^{\text{obs}}]$ should be estimated as the mean of $\mathbf{S}_\theta^F(\mathbf{Z}_i, \theta)$ for subjects who share the same values of $\mathbf{Z}_i^{\text{obs}}$ as the subject for whom $\mathbf{S}_\theta^F(\mathbf{Z}_i, \theta)$ is missing. This is known as the *mean score method*.

7.3 Generalised estimating equations (GEE): an example of a method belonging to \mathcal{R}

Suppose that $(Y_{1,i}, Y_{2,i}, \dots, Y_{D_i-1,i})$ are $D_i - 1 \leq T$ correlated outcomes measured on each of $i = 1, \dots, n$ subjects with fully-observed time-stationary covariates $(X_{1,i}, \dots, X_{p,i})$. Let \mathbf{X}_i be a $[(D_i - 1) \times (p + 1)T]$ covariate matrix for subject i , constructed as follows:

- Row 1 is the row vector $(1, X_{1,i}, \dots, X_{p,i})$ followed by $(p + 1)(T - 1)$ zeros
- Row 2 starts with $p + 1$ zeros, then the row vector $(1, X_{1,i}, \dots, X_{p,i})$ followed by $(p + 1)(T - 2)$ zeros
- ...
- Row $D_i - 1$ starts with $(p + 1)(D_i - 2)$ zeros, then the row vector $(1, X_{1,i}, \dots, X_{p,i})$ followed by $(p + 1)(T - D_i + 1)$ zeros

Let $\beta_j = (\beta_{j,0}, \beta_{j,1}, \dots, \beta_{j,p})^T$ be a $[(p + 1) \times 1]$ parameter vector for time j , such that $\mathbb{E}(Y_{j,i} | X_{1,i}, \dots, X_{p,i}) = \mu_{j,i} = f^{-1}(\eta_{j,i}) = f^{-1}(\beta_{j,0} + \beta_{j,1}X_{1,i} + \dots + \beta_{j,p}X_{p,i})$, and let $\beta = (\beta_1^T, \dots, \beta_T^T)^T$. Let \mathbf{D}_i be a $[(D_i - 1) \times (D_i - 1)]$ diagonal matrix with $(k, k)^{\text{th}}$ element

$$\frac{\partial \mu_{k,i}}{\partial \eta_{k,i}}$$

The reader is asked to note the distinction between \mathbf{D}_i , the matrix of partial derivatives, and D_i , the dropout indicator. The convention of using a boldface \mathbf{D}_i for the former and not for the latter will be maintained throughout. Let \mathbf{C}_i be the $[(D_i - 1) \times (D_i - 1)]$ correlation matrix of $(Y_{1,i}, Y_{2,i}, \dots, Y_{D_i-1,i})$,

$$\mathbf{V}_i = \text{diag} [\text{Var} (Y_{1,i}), \dots, \text{Var} (Y_{D_i-1,i})],$$

$$\mathbf{W}_i = \mathbf{V}_i^{\frac{1}{2}} \mathbf{C}_i \mathbf{V}_i^{\frac{1}{2}},$$

$$\mathbf{Y}_i = \begin{pmatrix} Y_{1,i} \\ Y_{2,i} \\ \vdots \\ Y_{D_i-1,i} \end{pmatrix}$$

and

$$\boldsymbol{\mu}_i = \mathbb{E} (\mathbf{Y}_i | X_{1,i}, X_{2,i}, \dots, X_{p,i})$$

Then the GEE estimate of $\boldsymbol{\beta}$ is the solution to:

$$\sum_{i=1}^n \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \quad (7.3.1)$$

For multivariate Gaussian outcomes, the multivariate distribution of \mathbf{Y}_i is entirely specified by the 1st- and 2nd-order moments, $f(\cdot)$ is the identity and (7.3.1) becomes

$$\sum_{i=1}^n \mathbf{X}_i^T \mathbf{W}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

which can be shown to be equal to the observed-data score equation and thus (assuming that the model is correctly specified and the parameters of the correlation matrix consistently estimated), these estimates are fully-efficient.

However, for non-Gaussian data, this is not true. Some information about the joint distribution of the outcomes is contained in the 3rd and higher order moments, which are not included in the model and therefore (7.3.1) is not a score equation, and the procedure is not a maximum likelihood procedure. This means that estimates are not fully-efficient. However, it can be easily shown (Liang and Zeger, 1986) that GEE estimates are consistent (asymptotically unbiased) as long as the missing data are MCAR, by showing that the summand in the left hand side of (7.3.1) has zero expectation. The GEE procedure is viewed favourably in the trade off between efficiency, practicality and the robustness which comes with the reduction from parametric to semiparametric modelling assumptions.

7.3.1 Working correlation structure

Another way in which non-Gaussian outcomes differ from Gaussian outcomes is that their variance is a function of their mean. This is why \mathbf{W}_i , the variance-covariance matrix, is split into two components — \mathbf{V}_i , which is a function of the mean vector $\boldsymbol{\mu}_i$, and \mathbf{C}_i the correlation matrix, which is functionally independent of $\boldsymbol{\mu}_i$ — as follows:

$$\mathbf{W}_i = \mathbf{V}_i^{\frac{1}{2}} \mathbf{C}_i \mathbf{V}_i^{\frac{1}{2}}$$

We are rarely directly interested in \mathbf{C}_i , and it is thus common to assume a ‘working’ structure for this matrix, which may or may not be correct. This is legitimate since it can be shown that our estimates of $\boldsymbol{\mu}_i$ are consistent under MCAR even when \mathbf{C}_i is misspecified, although correctly specifying \mathbf{C}_i leads to greater efficiency (Liang and Zeger, 1986). Thus GEE belongs to the subclass \mathcal{R} of semiparametric models.

Common choices for \mathbf{C}_i are:

- Independence: \mathbf{C}_i is the identity matrix. GEE with the independence working correlation matrix is often called *Independence estimating equations* or IEE.

- Exchangeable: $\mathbf{C}_i = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & & \ddots & & \vdots \\ \rho & \cdots & \rho & 1 & \rho \\ \rho & \cdots & \rho & \rho & 1 \end{pmatrix}$
- First-order auto-regressive: $\mathbf{C}_i = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{D_i-2} \\ \rho & 1 & \rho & \cdots & \rho^{D_i-3} \\ \vdots & & \ddots & & \vdots \\ \rho^{D_i-3} & \cdots & \rho & 1 & \rho \\ \rho^{D_i-2} & \cdots & \rho^2 & \rho & 1 \end{pmatrix}$
- Unstructured: $\mathbf{C}_i = \begin{pmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,D_i-1} \\ \rho_{1,2} & 1 & \rho_{2,3} & \cdots & \rho_{2,D_i-1} \\ \vdots & & \ddots & & \vdots \\ \rho_{1,D_i-2} & \cdots & \rho_{D_i-3,D_i-2} & 1 & \rho_{D_i-2,D_i-1} \\ \rho_{1,D_i-1} & \cdots & \rho_{D_i-3,D_i-1} & \rho_{D_i-2,D_i-1} & 1 \end{pmatrix}$

Once a structure has been chosen, the parameters of \mathbf{C}_i are estimated. The original method proposed by Liang and Zeger (1986) uses the Pearson residuals

$$\hat{r}_{j,i} = \frac{Y_{j,i} - \hat{\mu}_{j,i}}{\sqrt{\mathbf{V}_i^{(j,j)}}}$$

where $\mathbf{V}_i^{(j,j)}$ is the $(j, j)^{\text{th}}$ element of \mathbf{V}_i . For example, in the exchangeable structure, $\hat{\rho}$ is estimated by:

$$\hat{\rho} = \frac{\sum_{i=1}^n \sum_{t>t'} \hat{r}_t \hat{r}_{t'}}{\sum_{i=1}^n \frac{1}{2} (D_i - 1) (D_i - 2) - p + 1}$$

Smith and Kenward (2000) and Lipsitz *et al.* (2000) argue against this so-called available pairs method of estimating \mathbf{C}_i when data are missing, and suggest a more prin-

cipled alternative, based on quadratic estimation (Crowder, 1985, 1992). Using the notation of Smith and Kenward (2000), if $\mathbf{r}_i = (r_{1,i}, \dots, r_{D_i-1,i})^T$ is the vector of Pearson residuals for the i^{th} subject, then these are approximately normally distributed with zero mean and variance-covariance matrix \mathbf{C}_i , a function of the parameters $\boldsymbol{\rho} = (\rho_1, \dots, \rho_r)^T$. $\hat{\boldsymbol{\rho}}$ is obtained from the estimating equations

$$\sum_{i=1}^n \frac{\partial \mathbf{C}_i^{-1}}{\partial \rho_k} (\mathbf{C}_i - \mathbf{r}_i \mathbf{r}_i^T), \quad k = 1, \dots, r$$

7.3.2 Algorithm for fitting GEE

Assuming a working structure for \mathbf{C}_i , in order to solve (7.3.1), we expand the summand in a Taylor series about $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, where $\hat{\boldsymbol{\beta}}$ is the estimate that satisfies (7.3.1) and $\boldsymbol{\beta}^*$ is close to $\hat{\boldsymbol{\beta}}$. This gives rise to the following Fisher scoring procedure for solving (7.3.1):

1. Choose an initial estimate $\boldsymbol{\beta}^{(1)}$, for example, by fitting a least squares regression of the observed Y on \mathbf{X} independently for each timepoint.
2. Calculate $\boldsymbol{\mu}_i^{(1)} = f^{-1}(\mathbf{X}_i \boldsymbol{\beta}^{(1)})$ and \mathbf{C}_i for each i , using one of the methods outlined in §7.3.1.
3. Calculate $\boldsymbol{\beta}^{(2)}$ as

$$\boldsymbol{\beta}^{(2)} = \boldsymbol{\beta}^{(1)} + \left\{ \sum_{i=1}^n \mathbf{X}_i^T \mathbf{D}_i [\boldsymbol{\beta}^{(1)}] \mathbf{W}_i^{-1} [\boldsymbol{\beta}^{(1)}] \mathbf{D}_i [\boldsymbol{\beta}^{(1)}] \mathbf{X}_i \right\}^{-1} \cdot \sum_{i=1}^n \mathbf{X}_i^T \mathbf{D}_i [\boldsymbol{\beta}^{(1)}] \mathbf{W}_i^{-1} [\boldsymbol{\beta}^{(1)}] \{ \mathbf{Y}_i - \boldsymbol{\mu}_i [\boldsymbol{\beta}^{(1)}] \}$$

4. Substitute $\boldsymbol{\beta}^{(2)}$ for $\boldsymbol{\beta}^{(1)}$.
5. Repeat steps 2.–4. until the absolute difference between $\boldsymbol{\beta}^{(2)}$ and $\boldsymbol{\beta}^{(1)}$ is smaller than some pre-specified tolerance.

7.3.3 Estimating precision using the sandwich estimator

If the structure of \mathbf{C}_i is correctly specified, and its parameters consistently estimated, it can be shown that the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is given by:

$$\left(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right)^{-1}$$

However, as we are unlikely ever to be certain that the structure of \mathbf{C}_i is correctly specified, Liang and Zeger (1986) suggest using the following *sandwich* estimator of variance, where the residuals are used to correct for any misspecification of \mathbf{C}_i :

$$\left(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right)^{-1} \left[\sum_{i=1}^n \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i) (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)^T \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right] \cdot \left(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right)^{-1} \quad (7.3.2)$$

7.4 Weighted GEE

As we have already mentioned, GEE as described above is only valid in general under MCAR. One method for eliminating the asymptotic bias of GEE under MAR is inverse-probability weighting as introduced in §5.2. The weighting can be done either at the subject level or at the observation level.

Consider the following modification of (7.3.1):

$$\sum_{i=1}^n \frac{1}{\mathbf{P}(D_i = d_i | \mathbf{X}_i, \mathbf{Y}_i)} \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \quad (7.4.1)$$

This formulation was suggested by Fitzmaurice *et al.* (1995). Using a method by

Rotnitzky and Wypij (1994) for calculating asymptotic bias, Fitzmaurice *et al.* argue that their method (known as *cluster-weighted GEE*) is asymptotically unbiased under MAR when the inverse-probability-of-dropout weights are consistently estimated.

In order to apply this method we first estimate the inverse-probability-of-dropout weights from a series of logistic regression models as follows:

- The first logistic regression model estimates $\mathbb{P}(D_i = 1 | \mathbf{X}_i) =: p_{1,i}^C$. (This is often assumed to be zero for all subjects).
- The second logistic regression model estimates $\mathbb{P}(D_i = 2 | \mathbf{X}_i, Y_{1,i}, D_i > 1) =: p_{2,i}^C$.
- ...
- The k^{th} logistic regression model estimates

$$\mathbb{P}(D_i = k | \mathbf{X}_i, Y_{1,i}, \dots, Y_{k-1,i}, D_i > k - 1) =: p_{k,i}^C$$

- ...
- The T^{th} logistic regression model estimates

$$\mathbb{P}(D_i = T | \mathbf{X}_i, Y_{1,i}, \dots, Y_{T-1,i}, D_i > T - 1) =: p_{T,i}^C$$

- $p_{T+1,i}^C = 1$ by definition.
- Then, we put these together to get the marginal probabilities as follows:

$$p_{1,i}^M := p_{1,i}^C$$

and for $k > 1$,

$$\begin{aligned} p_{k,i}^M &:= \mathbb{P}(D_i = k | \mathbf{X}_i, Y_{1,i}, \dots, Y_{k-1,i}) \\ &= \mathbb{P}(D_i = k | \mathbf{X}_i, Y_{1,i}, \dots, Y_{k-1,i}, D_i > k - 1) \\ &\quad \cdot \mathbb{P}(D_i > k - 1 | \mathbf{X}_i, Y_{1,i}, \dots, Y_{k-1,i}) \end{aligned}$$

$$\begin{aligned}
&= \mathbf{P}(D_i = k | \mathbf{X}_i, Y_{1,i}, \dots, Y_{k-1,i}, D_i > k - 1) \\
&\quad \cdot [1 - \mathbf{P}(D_i \leq k - 1 | \mathbf{X}_i, Y_{1,i}, \dots, Y_{k-1,i})] \\
&= p_{k,i}^C (1 - p_{1,i}^M - p_{2,i}^M - \dots - p_{k-1,i}^M)
\end{aligned}$$

where throughout we are making the MAR assumption.

We add a cluster-level weight variable to the dataset, where the weight for subject i is

$$\frac{1}{p_{d,i}^M}$$

and the parameters of the model are estimated using the algorithm described in §7.3.2.

Now let $\pi_{t,i} = \mathbf{P}(R_{t,i} = 1 | Y_{1,i}, \dots, Y_{t-1,i}, \mathbf{X}_i)$ and

$$\Phi_i = \text{diag} \left(\frac{R_{1,i}}{\pi_{1,i}}, \dots, \frac{R_{T,i}}{\pi_{T,i}} \right)$$

and consider the following modification of (7.3.1):

$$\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \Phi_i (\tilde{\mathbf{Y}}_i - \tilde{\boldsymbol{\mu}}_i) = \mathbf{0} \quad (7.4.2)$$

where $\tilde{\mathbf{X}}_i$, $\tilde{\mathbf{D}}_i$, $\tilde{\mathbf{W}}_i$, $\tilde{\mathbf{Y}}_i$ and $\tilde{\boldsymbol{\mu}}_i$ are the counterfactual T -dimensional versions of \mathbf{X}_i , \mathbf{D}_i , \mathbf{W}_i , \mathbf{Y}_i and $\boldsymbol{\mu}_i$, which would have been used in a GEE with all the data fully-observed. Note that $\tilde{\mathbf{Y}}_i$ is the only one of these which involves unobserved data, since the other quantities are all functions only of $(X_{1,i}, \dots, X_{p,i})$ which are always observed. Furthermore, any unobserved element of $\tilde{\mathbf{Y}}_i$ corresponds to a column of zeros in Φ_i , hence the left-hand side of (7.4.2) is a function only of the observed data and is thus well-defined.

This is known as *observation-weighted GEE* and belongs to the class of estimators proposed by Robins *et al.* (1995) (although they advocate the use of a more efficient

estimator in the same class).

The weights are calculated by taking the inverse of the following estimated probabilities:

$$\begin{aligned}
 p_{j,i}^{\text{M-O}} &:= \mathbf{P}(R_{j,i} = 1 | \mathbf{X}_i, Y_{1,i}, \dots, Y_{j-1,i}) \\
 &= \mathbf{P}(D_i > j | \mathbf{X}_i, Y_{1,i}, \dots, Y_{j-1,i}) \\
 &= 1 - \mathbf{P}(D_i \leq j | \mathbf{X}_i, Y_{1,i}, \dots, Y_{j-1,i}) \\
 &= 1 - p_{1,i}^{\text{M}} - p_{2,i}^{\text{M}} - \dots - p_{j,i}^{\text{M}}
 \end{aligned}$$

where the $p_{j,i}^{\text{M}}$ are as defined in the cluster-level weighting procedure described above.

The equation (7.4.2) now represents a full-data GEE as far as estimation is concerned, since all the matrices and vectors are of full-data dimension. We have re-weighted the values of $(\mathbf{Y}_i - \boldsymbol{\mu}_i)$, and “padded out” the vector with zeros so that it is of full-data dimension.

If the means model is saturated (see Definition 3.2), the estimates obtained will be the same, irrespective of the choice of \mathbf{C}_i (O’Brien *et al.*, 2006), and thus we can carry out the analysis using the independence correlation matrix. The independence assumption means that data from previous timepoints are not involved in the estimation of the effect of interest at the final timepoint and observation-level weighted GEE is thus (in this case, when the means model is saturated) equivalent to a weighted univariate logistic regression at the final timepoint, using the observation-level weights.

7.4.1 High variability in the weights

If our estimate of the probability of dropout (for cluster-weighted GEE) or the probability of being observed (for observation-weighted GEE) is 0, then by definition, there are no examples in the dataset of such a subject dropping out or being observed, respectively, at that time. This means that at no point is there a practical problem with

infinite weights. However, if our model predicts values of the probabilities which are very close to 0, then some observations may have extremely large weights. This can lead to efficiency problems, with a few observations dominating the analysis, effectively reducing the sample size. All weighting procedures work best if the weights are moderate and not too variable (see Kang and Schafer, 2007, and the contribution to the discussion by Robins *et al.*).

7.5 MI-GEE

We noted in §6.7 that MI is a flexible method applicable in a wide variety of settings. MI-GEE (Paik, 1997) is one such setting. Under monotonicity, the imputation step is done sequentially: for each occasion we can impute the current outcome based on the values of the outcome on all previous occasions and the covariates. Then we fit a GEE to each completed dataset and combine the results using Rubin's rules. When applying MI to repeated binary data in this way there is one theoretical issue which we now consider.

Implicit in §6.7 is that the imputation and substantive models are both correctly specified. Otherwise, the MI estimates would be inconsistent. In the case of MI-GEE, a logistic regression is an obvious candidate for the imputation model, and also (marginally at each timepoint), for the substantive model. Note that the substantive model is marginal: it is the logistic regression of the outcome at each timepoint conditional only on the covariates, whereas the imputation model is a series of logistic regressions in which the outcome is always viewed conditionally on the covariate *and* all previous outcomes.

If one replaced 'logistic' with 'linear' in the above paragraph, there would be no problem, because if a collection of variables has a multivariate Gaussian distribution, then each of the variables has a marginal univariate Gaussian distribution, and the distribution of any one of the variables conditional on any selection of the others is also

univariate Gaussian. However, the equivalent does *not* hold for binary data under logistic regression. Indeed, if each outcome variable marginally obeys the assumptions of a univariate logistic regression, then it is impossible for the assumptions of the logistic regressions to hold conditionally on previous outcomes and vice versa. However, evidence suggests that only in extreme cases does this *uncongeniality* lead to a noticeable bias in practice. For more on this issue see Meng (1994), where it is shown that having as full and rich an imputation model as possible helps to protect against the possible biases introduced by uncongeniality.

Once the imputations have been drawn, the subsequent GEE analyses are performed on complete datasets, and therefore, if the means model is saturated, the choice of covariance structure is irrelevant, and MI-GEE is equivalent to MI-IEE.

7.6 Improved efficiency and double robustness

7.6.1 Augmented inverse probability weighted (AIPW) estimator (\mathcal{I} -type)

Recall that in §5.2, we derived the IPWCC estimator, the estimator that solves

$$\sum_{i=1}^n \frac{\mathbf{1}(C_i = \infty)}{\mathbb{P}(C_i = \infty | \mathbf{Z}_i)} \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \hat{\boldsymbol{\theta}}^{\text{IPWCC}}) = \mathbf{0}$$

We commented on its consistency, but also its inefficiency. Robins, Rotnitzky and their co-workers have published many papers on how to improve the efficiency of this estimator by using information on the incomplete cases to *augment* the IPWCC estimator. The resulting estimator is known as the *Augmented Inverse Probability Weighted* (AIPW) estimator. See for example Robins and Rotnitzky (1992), Robins *et al.* (1994) and Tsiatis (2006).

Consider the alternative estimating equation

$$\sum_{i=1}^n \left\{ \frac{\mathbf{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty | \mathbf{Z}_i)} \mathbf{S}_{\theta}^F(\mathbf{Z}_i, \hat{\theta}^{\text{AIPW}}) + \left[1 - \frac{\mathbf{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty | \mathbf{Z}_i)} \right] \phi[C_i, G_{C_i}(\mathbf{Z}_i), \hat{\theta}^{\text{AIPW}}] \right\} = \mathbf{0} \quad (7.6.1)$$

where $\phi[C_i, G_{C_i}(\mathbf{Z}_i), \theta]$ is—for the time being—just an arbitrary function. This is a sensible equation to consider because

$$\begin{aligned} & \mathbf{E} \left\{ \left[1 - \frac{\mathbf{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty | \mathbf{Z}_i)} \right] \phi[C_i, G_{C_i}(\mathbf{Z}_i), \theta] \right\} \\ &= \mathbf{E} \left(\mathbf{E} \left\{ \left[1 - \frac{\mathbf{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty | \mathbf{Z}_i)} \right] \phi[C_i, G_{C_i}(\mathbf{Z}_i), \theta] \middle| C_i, G_{C_i}(\mathbf{Z}_i) \right\} \right) \\ &= \mathbf{E} \left(\mathbf{E} \left\{ \left[1 - \frac{\mathbf{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty | \mathbf{Z}_i)} \right] \middle| C_i, G_{C_i}(\mathbf{Z}_i) \right\} \phi[C_i, G_{C_i}(\mathbf{Z}_i), \theta] \right) = \mathbf{0} \end{aligned}$$

under the CAR assumption, and therefore the consistency of the IPWCC estimator is preserved.

The Hilbert Space/Influence Function theory that underpins most of the work carried out by Robins *et al.* in this field can be used to exhibit the optimum (i.e. most efficient) choice of $\phi[C_i, G_{C_i}(\mathbf{Z}_i), \theta]$. Loosely speaking, a Hilbert space is an extension of Euclidean space that allows for potentially infinite dimensions, and the set of influence functions for RAL estimators forms a Hilbert space. The advantage of thinking about estimators in this way is that the length of influence functions as defined by the distance metric in this Hilbert space is related to the variance of the associated estimator, and therefore the search for an efficient estimator can be translated into a geometry problem and the extensive theory of Hilbert spaces (for example, the Projection Theorem) can be exploited to find answers to our RAL estimator problems.

This theory tells us that the most efficient estimator of this form is found by choosing

$$\phi[C_i, G_{C_i}(\mathbf{Z}_i), \theta] = \mathbf{E}[\mathbf{S}_{\theta}^F(\mathbf{Z}_i, \theta) | C_i, G_{C_i}(\mathbf{Z}_i)] \quad (7.6.2)$$

Therefore, the AIPW estimator is the solution to:

$$\sum_{i=1}^n \left\{ \frac{\mathbb{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty | \mathbf{Z}_i)} \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \hat{\boldsymbol{\theta}}^{\text{AIPW}}) + \left[1 - \frac{\mathbb{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty | \mathbf{Z}_i)} \right] \mathbf{E} \left[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \hat{\boldsymbol{\theta}}^{\text{AIPW}}) \middle| C_i, G_{C_i}(\mathbf{Z}_i) \right] \right\} = \mathbf{0} \quad (7.6.3)$$

If $\mathbf{E}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}) | C_i, G_{C_i}(\mathbf{Z}_i)]$ is estimated semiparametrically or non-parametrically, then this estimator is semiparametric and belongs to the subclass \mathcal{I} of semiparametric models.

7.6.2 Double robustness

In order to obtain an estimate of $\hat{\boldsymbol{\theta}}^{\text{AIPW}}$, we must first obtain estimates of

$$\mathbf{P}(C_i = \infty | \mathbf{Z}_i) \quad (7.6.4)$$

and of

$$\mathbf{E}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}) | C_i, G_{C_i}(\mathbf{Z}_i)] \quad (7.6.5)$$

Robins and his colleagues have shown that the AIPW estimator has a property known as *double robustness*:

Theorem 7.1. *If either the model that gives (7.6.4) or the model that gives (7.6.5) (but not both) is incorrectly specified, then the AIPW estimator remains consistent.*

Proof. Let $\pi[C_i, G_{C_i}(\mathbf{Z}_i)]$ be the true value of $\mathbf{P}(C_i = \infty | \mathbf{Z}_i)$ and let $\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]$ be its estimate under the model for (7.6.4). Similarly, let $\hat{\mathbf{E}}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \hat{\boldsymbol{\theta}}^{\text{AIPW}}) | C_i, G_{C_i}(\mathbf{Z}_i)]$ be the estimate of $\mathbf{E}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \hat{\boldsymbol{\theta}}^{\text{AIPW}}) | C_i, G_{C_i}(\mathbf{Z}_i)]$.

For consistency, we need that

$$\mathbf{E} \left(\frac{\mathbf{1}(C_i = \infty)}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) + \left\{ 1 - \frac{\mathbf{1}(C_i = \infty)}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \right\} \hat{\mathbf{E}}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \right) = \mathbf{0}$$

Assuming only that the model for (7.6.4) is correct,

$$\begin{aligned} & \mathbf{E} \left(\frac{\mathbf{1}(C_i = \infty)}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) + \left\{ 1 - \frac{\mathbf{1}(C_i = \infty)}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \right\} \hat{\mathbf{E}}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \right) \\ &= \mathbf{E} \left[\mathbf{E} \left(\frac{\mathbf{1}(C_i = \infty)}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) + \left\{ 1 - \frac{\mathbf{1}(C_i = \infty)}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \right\} \hat{\mathbf{E}}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \middle| C_i, G_{C_i}(\mathbf{Z}_i) \right) \right] \\ &= \mathbf{E} \left(\frac{\pi[C_i, G_{C_i}(\mathbf{Z}_i)]}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \mathbf{E}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) | C_i, G_{C_i}(\mathbf{Z}_i)] + \left\{ 1 - \frac{\pi[C_i, G_{C_i}(\mathbf{Z}_i)]}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \right\} \hat{\mathbf{E}}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \right) \\ &= \mathbf{E} \left\{ \mathbf{E}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) | C_i, G_{C_i}(\mathbf{Z}_i)] + (1 - 1) \hat{\mathbf{E}}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \right\} \\ &= \mathbf{E} \left\{ \mathbf{E}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \right\} = \mathbf{E}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0)] = \mathbf{0} \end{aligned}$$

Assuming only that the model for (7.6.5) is correct,

$$\begin{aligned} & \mathbf{E} \left(\frac{\mathbf{1}(C_i = \infty)}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) + \left\{ 1 - \frac{\mathbf{1}(C_i = \infty)}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \right\} \hat{\mathbf{E}}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \right) \\ &= \mathbf{E} \left[\mathbf{E} \left(\frac{\mathbf{1}(C_i = \infty)}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) + \left\{ 1 - \frac{\mathbf{1}(C_i = \infty)}{\hat{\pi}[C_i, G_{C_i}(\mathbf{Z}_i)]} \right\} \hat{\mathbf{E}}[\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \middle| C_i, G_{C_i}(\mathbf{Z}_i) \right) \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbf{E} \left(\frac{\pi [C_i, G_{C_i}(\mathbf{Z}_i)]}{\hat{\pi} [C_i, G_{C_i}(\mathbf{Z}_i)]} \mathbf{E} [\mathbf{S}_{\theta}^F(\mathbf{Z}_i, \theta_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \right. \\
&\quad \left. + \left\{ 1 - \frac{\pi [C_i, G_{C_i}(\mathbf{Z}_i)]}{\hat{\pi} [C_i, G_{C_i}(\mathbf{Z}_i)]} \right\} \hat{\mathbf{E}} [\mathbf{S}_{\theta}^F(\mathbf{Z}_i, \theta_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \right) \\
&= \mathbf{E} \left(\mathbf{E} [\mathbf{S}_{\theta}^F(\mathbf{Z}_i, \theta_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \right. \\
&\quad \left. + \left\{ \frac{\pi [C_i, G_{C_i}(\mathbf{Z}_i)]}{\hat{\pi} [C_i, G_{C_i}(\mathbf{Z}_i)]} - \frac{\pi [C_i, G_{C_i}(\mathbf{Z}_i)]}{\hat{\pi} [C_i, G_{C_i}(\mathbf{Z}_i)]} \right\} \mathbf{E} [\mathbf{S}_{\theta}^F(\mathbf{Z}_i, \theta_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \right) \\
&= \mathbf{E} \left\{ \mathbf{E} [\mathbf{S}_{\theta}^F(\mathbf{Z}_i, \theta_0) | C_i, G_{C_i}(\mathbf{Z}_i)] \right\} \\
&= \mathbf{E} [\mathbf{S}_{\theta}^F(\mathbf{Z}_i, \theta_0)] \\
&= \mathbf{0}
\end{aligned}$$

□

By comparing (7.6.3) with (7.2.2), we see how double robustness is gained at the expense of efficiency. If the model for estimating (7.6.5) is correct, then the asymptotic efficiency of (7.2.2) is the same as the asymptotic efficiency of the corresponding maximum likelihood estimate, but if the model is incorrect, (7.2.2) is inconsistent. When both models are correct, (7.6.3) is less efficient than (7.2.2), but if the model for estimating (7.6.5) is incorrect, then, as long as the model for estimating (7.6.4) is correct, (7.6.3) is consistent. If the model for (7.6.5) is incorrect and the model for (7.6.4) is correct, then the efficiency of (7.6.3) decreases relative to (7.2.2), but (7.6.3) remains consistent.

What happens to (7.6.3) when both models are incorrect remains a contentious issue in the literature, with Kang and Schafer (2007) claiming that

“... at least in some settings, two wrong models are not better than one.”

For more on double robustness, see the article (and discussion, in particular the contributions by Tsiatis and Davidian and by Robins *et al.*) by Kang and Schafer (2007).

7.6.3 AIPW estimator (\mathcal{R} -type)

In §7.6.1, we started by assuming that had we observed the full data we would estimate θ by solving the full-data score equation

$$\sum_{i=1}^n \mathbf{S}_{\theta}^F(\mathbf{Z}_i, \hat{\theta}) = \mathbf{0}$$

However, supposing that we wish to relax these assumptions about the full-data density, we could instead propose

$$\sum_{i=1}^n \mathbf{U}_{\theta}(\mathbf{Z}_i, \hat{\theta}) = \mathbf{0}$$

where $\mathbf{U}_{\theta}(\cdot)$ is any function satisfying

$$\mathbf{E}[\mathbf{U}_{\theta}(\mathbf{Z}_i, \theta_0)] = \mathbf{0} \tag{7.6.6}$$

It is trivial that such a function exists. Suppose that $\mathbf{E}(\mathbf{Z}_i) = \psi$, then ψ must be a function of θ , otherwise θ would not fully describe the distribution of \mathbf{Z}_i , and thus

$$\mathbf{U}_{\theta}(\mathbf{Z}_i, \hat{\theta}) = \mathbf{Z}_i - \hat{\psi}$$

is one possible (non-parametric) choice.

Whatever the choice of $\mathbf{U}_{\theta}(\cdot)$ (as long as it satisfies (7.6.6)), it follows that

$$\sum_{i=1}^n \frac{\mathbf{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty)} \mathbf{U}_{\theta}(\mathbf{Z}_i, \hat{\theta}^{\text{IPWCC}}) = \mathbf{0}$$

leads to consistent estimates under CAR.

It also follows that, by considering estimating equations of the form

$$\sum_{i=1}^n \left\{ \frac{\mathbf{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty | \mathbf{Z}_i)} \mathbf{U}_\theta(\mathbf{Z}_i, \hat{\theta}^{\text{AIPW}}) + \left[1 - \frac{\mathbf{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty | \mathbf{Z}_i)} \right] \phi[C_i, G_{C_i}(\mathbf{Z}_i), \hat{\theta}^{\text{AIPW}}] \right\} = \mathbf{0} \quad (7.6.7)$$

the efficiency could be increased.

Robins *et al.* show that, for a particular choice of $\mathbf{U}_\theta(\mathbf{Z}_i, \theta)$,

$$\phi[C_i, G_{C_i}(\mathbf{Z}_i), \theta] = \mathbf{E}[\mathbf{U}_\theta(\mathbf{Z}_i, \theta) | C_i, G_{C_i}(\mathbf{Z}_i)] \quad (7.6.8)$$

is the optimal choice for $\phi[C_i, G_{C_i}(\mathbf{Z}_i), \theta]$, and hence, the more general AIPW estimating equation is given by

$$\sum_{i=1}^n \left\{ \frac{\mathbf{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty | \mathbf{Z}_i)} \mathbf{U}_\theta(\mathbf{Z}_i, \hat{\theta}^{\text{AIPW}}) + \left[1 - \frac{\mathbf{1}(C_i = \infty)}{\mathbf{P}(C_i = \infty | \mathbf{Z}_i)} \right] \mathbf{E}[\mathbf{U}_\theta(\mathbf{Z}_i, \hat{\theta}^{\text{AIPW}}) | C_i, G_{C_i}(\mathbf{Z}_i)] \right\} = \mathbf{0} \quad (7.6.9)$$

The double robustness property carries through, the only difference now being that when we compare (7.6.9) with (7.2.2), the comparative efficiency of (7.6.9) is lower than when we compared (7.6.3) with (7.2.2), since not only are we introducing inverse probability weights, but $\mathbf{U}_\theta(\cdot)$ is suboptimal in terms of efficiency. Again, however there is a trade-off between dependence on modelling assumptions and efficiency.

If $\mathbf{U}_\theta(\mathbf{Z}_i, \theta)$ poses parametric restrictions only on $\mathbf{E}(\mathbf{Z}_i)$ (and no higher moments), then the semiparametric estimator belongs to \mathcal{R} . Note that (7.6.9) is only the most efficient estimator of the form given in (7.6.7), i.e. for a particular choice of $\mathbf{U}_\theta(\mathbf{Z}_i, \theta)$. Rotnitzky and Robins (1997) show how to choose $\mathbf{U}_\theta(\mathbf{Z}_i, \theta)$ to achieve the most efficient estimator in a given semiparametric class (such as \mathcal{R}). In general, the optimal $\mathbf{U}_\theta(\mathbf{Z}_i, \theta)$ is non-obvious, in the sense that it doesn't correspond to an estimating

equation (such as GEE) in common use. Moreover, in many situations the optimal $U_{\theta}(\mathbf{Z}_i, \theta)$ does not have a closed-form representation and an iterative algorithm is required to achieve the semiparametric efficiency bound (see, for example Tsiatis, 2006, ch. 10). In many situations, the difference in efficiency between the IPWCC and a sensibly-chosen AIPW estimator is far greater than the difference in efficiency between this AIPW estimator and the semiparametric-efficient estimator. This, coupled with the often intractable form of the semiparametric-efficient estimator means that these optimal estimators are rarely used in practice.

7.6.4 Regression formulation from Bang and Robins

Although AIPW methods have very attractive properties, one feature that has probably severely restricted their use in practice is that no general method exists for their derivation. The paper by Bang and Robins (2005), where a reasonably general method is described for three situations (cross-sectional univariate missing data, longitudinal data with monotone dropout and marginal structural models) is therefore a very important addition to the literature on this topic. First, we describe the Bang and Robins approach in the univariate cross-sectional setting before describing the extension to longitudinal data with monotone dropout.

7.6.4.1 Cross-sectional univariate missing data

Let the full data $\mathbf{Z}_i = (\mathbf{X}_i^T, Y_i)^T$ for subject $i \in \{1, \dots, n\}$ be a fully-observed vector of covariates \mathbf{X}_i and a scalar outcome Y_i which could be missing ($R_i = 0$) or observed ($R_i = 1$) and interest lies in estimating $\mu = \mathbf{E}(Y_i)$.

Under the MAR assumption, consistent estimators of μ could in theory be obtained in two ways. First, an IPWCC estimator

$$\sum_{i=1}^n \frac{R_i}{\mathbb{P}(R_i = 1 | \mathbf{X}_i)} (Y_i - \mu) = 0$$

and second, a regression estimator

$$\sum_{i=1}^n [\mathbb{E}(Y_i | \mathbf{X}_i) - \mu] = 0$$

The first would require a model for the inverse probability weights and the second would require a model for Y_i conditional on \mathbf{X}_i .

Bang and Robins (2005) suggest combining these two ideas as follows. First a suitable regression model (such as logistic regression) is chosen for R conditional on \mathbf{X} —we call this the π -model. Let $\hat{\boldsymbol{\alpha}}$ be the parameter estimates from this regression and let $\hat{\pi}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}})$ be the predicted probabilities (that $R_i = 1$) from this model. Then we fit a generalised linear model for Y conditional on \mathbf{X} and $\hat{\pi}^{-1}$ (i.e. with the inverse probability weights included as a covariate in the linear predictor) to those subjects who have complete data. We call the corresponding model *without* the inverse probability weights, i.e.

$$\mathbb{E}(Y_i | \mathbf{X}_i, R_i = 1) = \Psi \left[s(\mathbf{X}_i, \hat{\boldsymbol{\beta}}) \right]$$

the y -model, where Ψ is the canonical link function from an appropriate GLM and $s(\mathbf{X}, \boldsymbol{\beta})$ is a known function. We call

$$\mathbb{E} \left[Y_i | \mathbf{X}_i, \hat{\pi}^{-1}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}), R_i = 1 \right] = \Psi \left[s(\mathbf{X}_i, \hat{\boldsymbol{\beta}}) + \hat{\phi} \hat{\pi}^{-1}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}) \right]$$

the extended y -model.

Let

$$\hat{e}(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}) = \Psi \left[s(\mathbf{X}_i, \hat{\boldsymbol{\beta}}) + \hat{\phi} \hat{\pi}^{-1}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}) \right]$$

be the predictions from the extended y -model. Note that although the extended y -model was fitted to the complete cases only, $\hat{e}(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi})$ can be calculated for all subjects. Finally, the proposed estimator is the solution $\hat{\mu}_{\text{DR}}$ to

$$\sum_{i=1}^n \left[\hat{e}(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}) - \mu_{\text{DR}} \right] = 0$$

Theorem 7.2 (DR cross-sectional estimator). *The estimator $\hat{\mu}_{\text{DR}}$ is doubly robust. That is, if either the π -model or the y -model is incorrectly specified, but not both, $\hat{\mu}_{\text{DR}}$ is a consistent estimator of μ . Furthermore, its asymptotic efficiency is optimal amongst estimators which put no parametric restriction on the distribution of Y .*

Proof. Let us write \hat{e}_i for $\hat{e}(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi})$ and $\hat{\pi}_i$ for $\hat{\pi}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}})$. The DR estimating equation

$$\sum_{i=1}^n (\hat{e}_i - \mu_{\text{DR}}) = 0 \tag{7.6.10}$$

can be rewritten as

$$\sum_{i=1}^n [\hat{e}_i - \mu_{\text{DR}} + R_i \hat{\pi}_i^{-1} (Y_i - \hat{e}_i)] = 0 \tag{7.6.11}$$

This follows from the fact that $\sum_{i=1}^n R_i \hat{\pi}_i^{-1} (Y_i - \hat{e}_i)$ is numerically zero since we included $\hat{\pi}_i^{-1}$ in our GLM for the extended y -model.

But we can rewrite (7.6.11) as

$$\sum_{i=1}^n [R_i \hat{\pi}_i^{-1} (Y_i - \mu_{\text{DR}}) + (1 - R_i \hat{\pi}_i^{-1}) (\hat{e}_i - \mu_{\text{DR}})] = 0$$

which we immediately recognise as being of the same form as (7.6.9) and thus must be consistent when the π -model is correctly specified.

Showing that $\hat{\mu}_{\text{DR}}$ is a consistent estimator of μ when the y -model is correctly specified is straightforward. We must show that the expectation of the summand in the LHS of (7.6.10), at the true parameter-values, is zero. But when the y -model is correctly

specified, the true value of ϕ is 0, and e_i reduces to (the unextended) $\Psi[s(\mathbf{X}_i, \boldsymbol{\beta})]$. Thus, by the MAR assumption, and the consistency of a correctly specified GLM, $\hat{\mu}_{\text{DR}}$ is a consistent estimator of μ when the y -model is correctly specified.

Asymptotic efficiency follows from the fact that \hat{e}_i is a consistent estimator of $\mathbf{E}(Y_i | \mathbf{X}_i)$, the orthogonal and optimal choice defined in (7.6.8). \square

7.6.4.2 Longitudinal data with monotone missingness

Let us now suppose that the full data $\mathbf{Z}_i = (\mathbf{X}_i^T, \mathbf{Y}_i^T)^T$ for subject $i \in \{1, \dots, n\}$ consist of a fully-observed vector of covariates \mathbf{X}_i and a vector of repeated measures $\mathbf{Y}_i = (Y_{1,i}, \dots, Y_{T,i})^T$ subject to monotone dropout and that interest lies in estimating $\mu = \mathbf{E}(Y_{i,T})$. Let $\mathbf{R}_i = (R_{1,i}, \dots, R_{T,i})^T$ be the vector of missingness indicators with $R_{t,i} = \mathbf{1}(Y_{t,i} \text{ is observed})$, and let D_i —the earliest t for which $R_{t,i} = 0$ —be the dropout indicator.

The IPWCC estimator is the solution to:

$$\sum_{i=1}^n \frac{R_{T,i}}{\mathbf{P}(R_{T,i} = 1 | \mathbf{X}_i, Y_{1,i}, \dots, Y_{T-1,i})} (Y_{T,i} - \mu) = 0$$

This represents one way in which we might obtain a consistent (albeit inefficient) estimator of μ under MAR. We would calculate $\mathbf{P}(R_{T,i} = 1 | \mathbf{X}_i, Y_{1,i}, \dots, Y_{T-1,i})$ under the MAR assumption, by fitting a series of models to estimate

$$\lambda(t | \mathbf{X}_i, Y_{1,i}, \dots, Y_{t-1,i}) = \mathbf{P}(D_i = t | D_i \geq t, \mathbf{X}_i, Y_{1,i}, \dots, Y_{t-1,i})$$

Then, the marginal probabilities $\pi_{t,i} = \mathbf{P}(Y_{t,i} \text{ is observed} | \mathbf{X}_i, Y_{1,i}, \dots, Y_{t-1,i}) =$

$\mathbf{P}(D_i > t | \mathbf{X}_i, Y_{1,i}, \dots, Y_{t-1,i})$ are estimated as

$$\hat{\pi}_{t,i} = \prod_{j=1}^t [1 - \lambda(j | \mathbf{X}_i, Y_{1,i}, \dots, Y_{j-1,i})]$$

However, there is an alternative regression estimator, analogous to the cross-sectional case, which—in the longitudinal setting—is best described recursively. Let $H_{T,i} = Y_{T,i}$ if $R_{T,i} = 1$ ($H_{T,i}$ is not defined for $R_{T,i} = 0$) and, for $t < T$, $H_{t,i} = \mathbf{E}(H_{t+1,i} | \mathbf{X}_i, Y_{1,i}, \dots, Y_{t,i}, D_i \geq t+2)$ (again, with $H_{t,i}$ undefined if $D_i \leq t+1$). Upon quick inspection of these recursive functions, we see that, under MAR, $\mathbf{E}(H_{1,i}) = \mu, \forall i$. This leads to the following alternative estimating equation for μ :

$$\sum_{i=1}^n (\hat{H}_{1,i} - \mu) = 0$$

In the first representation, models (such as logistic regression) must be chosen for estimating each

$$\lambda(t | \mathbf{X}_i, Y_{1,i}, \dots, Y_{t-1,i}) = \mathbf{P}(D_i = t | D_i \geq t, \mathbf{X}_i, Y_{1,i}, \dots, Y_{t-1,i})$$

and in the second representation models (such as linear regression if $Y_{t,i}$ is continuous, logistic regression if $Y_{t,i}$ is binary) must be chosen for estimating each

$$H_{t,i} = \mathbf{E}(H_{t+1,i} | \mathbf{X}_i, Y_{1,i}, \dots, Y_{t,i}, D_i \geq t+2)$$

Bang and Robins (2005) go on to show how a doubly robust estimator may be derived by combining these two representations. The algorithm is as follows:

1. Fit a series of parametric regression models to estimate $\lambda(t | \mathbf{X}_i, Y_{1,i}, \dots, Y_{t-1,i})$,

from the observed data, and from these obtain estimates

$$\hat{\pi}_{t,i} = \prod_{j=2}^t \left[1 - \hat{\lambda}(j | \mathbf{X}_i, Y_{1,i}, \dots, Y_{j-1,i}) \right]$$

2. Let $H_{T,i} = Y_{T,i}$.
3. For $t = T - 1, T - 2, \dots, 2$,
 - (a) For subjects with $D_i \geq t + 1$, fit a parametric regression model with $\hat{H}_{t,i}$ as the outcome and $\mathbf{X}_i, Y_{1,i}, Y_{2,i}, \dots, Y_{t-1,i}$ and $\hat{\pi}_{t-1,i}^{-1}$ as predictors.
 - (b) For subjects with $D_i \geq t$, let $\hat{H}_{t-1,i}$ be the predicted values from the regression in (a).
4. The doubly robust estimator of μ is given by $n^{-1} \sum_{i=1}^n \hat{H}_{1,i}$.

The proof that this estimator is both doubly robust and asymptotically optimally efficient (amongst estimators which pose no parametric restriction on the distribution of \mathbf{Y}) is similar to the proof given above for the cross-sectional case (see the appendix of Bang and Robins (2005)).

No variance estimator exists for these regression-formulated DR estimators and Bang and Robins (2005) suggest using the bootstrap to obtain variance estimates.

7.6.5 A semiparametric-efficient GEE-type estimator

We return to the repeated binary outcome case described in §7.3, and to the problem of finding a consistent (under MAR) estimator which is more efficient than weighted-GEE but with the same semiparametric restrictions. More formally, we use a class of weighted estimating equations to which GEE belongs, as described by Robins *et al.* (1995). This class of estimating equations is given by

$$\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i^* \tilde{\mathbf{W}}_i^{-1} \tilde{\Phi}_i (\tilde{\mathbf{Y}}_i - \tilde{\boldsymbol{\mu}}_i) = \mathbf{0} \quad (7.6.12)$$

where each quantity is defined as in (7.4.2), except for $\tilde{\mathbf{D}}_i^*$, which is *any* $([p+1]T \times T)$ matrix of functions of \mathbf{X}_i and $\tilde{\boldsymbol{\beta}}$. Although $\tilde{\Phi}_i$ is thought of as the matrix of observation-level weights, by setting all the weights (wrongly) to 1 and setting $\tilde{\mathbf{D}}_i^* = \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i$, we see that ordinary unweighted GEE belongs to this class as well as observation-level weighted GEE. Furthermore, by setting all the weights for subject i to be $[\mathbb{P}(D_i = d_i)]^{-1}$, we see that cluster-level weighted GEE also belongs to this class.

Many references to Robins *et al.* (1995) and Robins and Rotnitzky (1995) in the literature fail to recognise the distinction between (7.4.2) and (7.6.12), and Robins *et al.* are often wrongly claimed to advocate the use of (7.4.2). Although (7.4.2) belongs to the class described by (7.6.12), it is not the most efficient estimator in this class.

Robins and Rotnitzky (1995) derive the most efficient estimator in the (7.6.12) class and prove that its efficiency attains the semiparametric efficiency bound for this class. They describe an adaptive procedure for its estimation, which we now describe.

In addition to the notation introduced in §7.3 and §7.4, let

$$G_{j,t,i} = \mathbb{E}(Y_{j,i} - \mu_{j,i} | R_{t-1,i} = 1, \bar{Y}_{t-1,i}, \mathbf{X}_i)$$

Then the procedure is as follows:

1. Calculate $\hat{\boldsymbol{\beta}}$, an initial (inefficient) estimator of $\boldsymbol{\beta}$ e.g. from a suitable non-augmented IPWCC estimating equation.

2. Specify a regression model (such as logistic) for

$$\mathbb{P}(R_{j,i} = 1 \mid R_{t-1,i} = 1, \bar{Y}_{t-1,i}, \mathbf{X}_i)$$

and estimate its parameters using maximum likelihood. Let $\hat{\lambda}_{j,t,i}$ be the estimate of $\mathbb{P}(R_{j,i} = 1 \mid R_{t-1,i} = 1, \bar{Y}_{t-1,i}, \mathbf{X}_i)$ from this model. Monotonicity dictates that $\hat{\lambda}_{j,t,i} = 1$ if $j < t$.

3. Use the estimates $\hat{\lambda}_{j,t,i}$ to calculate estimates

$$\hat{\pi}_{t,i} = \prod_{k=2}^t \hat{\lambda}_{k,k-1,i}$$

of $\mathbb{P}(R_{t,i} = 1 \mid \bar{Y}_{t-1,i}, \mathbf{X}_i)$.

4. Let $K_{j,t,i} = \hat{\pi}_{t,i} \hat{\pi}_{j,i}^{-1} (Y_{j,i} - \mu_{j,i})$ and specify a regression model for

$$\mathbb{E}(K_{j,t-1,i} \mid R_{j,i} = 1, \bar{Y}_{t-1,i}, \mathbf{X}_i)$$

Let $\hat{\kappa}_{j,t-1,i}$ be the estimate of $\mathbb{E}(K_{j,t-1,i} \mid R_{j,i} = 1, \bar{Y}_{t-1,i}, \mathbf{X}_i)$ from this model.

5. Let $\hat{G}_{j,t,i} = \hat{\lambda}_{j,t,i} \hat{\kappa}_{j,t-1,i}$ be an estimate of $G_{j,t,i}$.
6. Let $\hat{\mathbf{Q}}_{t,i}$ be a column vector with j^{th} element $\hat{\pi}_{t,i}^{-1} \hat{G}_{j,t,i}$ if $j \geq t$ and 0 otherwise.
7. Let $\hat{\mathbf{P}}_i = \sum_{t=1}^T (R_{t,i} - \hat{\lambda}_{t,t-1,i} R_{t-1,i}) \hat{\mathbf{Q}}_{t,i}$.
8. Let $\mathbf{U}_i = \hat{\Phi}_i (\tilde{\mathbf{Y}}_i - \hat{\boldsymbol{\mu}}_i)$ where $\hat{\Phi}_i$ is calculated using $\hat{\pi}_{t,i}$ and $\hat{\boldsymbol{\mu}}_i$ is calculated using $\hat{\boldsymbol{\beta}}$.
9. Estimate

$$\hat{\Lambda} = \mathbb{E} \left[(\mathbf{U}_i - \mathbf{P}_i) (\mathbf{U}_i - \mathbf{P}_i)^T \mid \mathbf{X}_i \right]$$

by multivariate least squares.

10. Let $\hat{\mathbf{D}}_i^*$ be the estimate of $\tilde{\mathbf{D}}_i^* = \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i$ obtained when substituting $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$.
11. Let $\hat{\mathbf{S}}_{t,i} = \hat{\mathbf{D}}_i^* \hat{\Lambda} \hat{\mathbf{Q}}_{t,i}$.

12. Find $\hat{\omega}$, the partial maximum likelihood estimate of ω in

$$\text{logit}\lambda_{t,t-1,i}^{(S)}(\omega) = \text{logit}\lambda_{t,t-1,i} + \delta^T \hat{\mathbf{S}}_{t,i} \quad (7.6.13)$$

13. Update $\hat{\Phi}_i$ using $\lambda_{t,t-1,i}^{(S)}(\hat{\omega})$.

14. Finally, solve

$$\sum_{i=1}^n \hat{\mathbf{D}}_i^* \hat{\Lambda} \hat{\Phi}_i (\tilde{\mathbf{Y}}_i - \hat{\boldsymbol{\mu}}_i) = \mathbf{0}$$

to find β_{adap} , the adaptive semiparametric-efficient estimator of β .

The result as it appears in Robins and Rotnitzky (1995) is more general, since it relaxes the time-stationary constraint on the covariates and includes an additional set of covariates (also time-updated) $V_{1,i}, \dots, V_{D_i-1,i}$ for each subject, where the substantive model is the regression of \mathbf{Y} on \mathbf{X} , without conditioning on these additional covariates. The authors in Robins *et al.* (1995) derive a consistent sandwich estimator for the variance of β_{adap} .

Note that this estimator is not the augmented version of weighted-GEE. Such an estimator would improve the efficiency of weighted-GEE, but is not optimally efficient amongst estimators in the class defined above. Robins and Rotnitzky (1995) have chosen the optimal semiparametric estimating function $\mathbf{U}(\cdot)$, which is not the GEE estimating function, and the estimator they derive is its augmented counterpart.

This estimator differs from the estimator we would obtain from the Bang and Robins (2005) procedure described in §7.6.4.2 since the former imposes a parametric restriction on the marginal means of the outcomes whereas the latter is in this sense non-parametric. In the special case where the marginal means model in the former is saturated, we would expect the two methods to converge.

Part III

Multiple imputation for doubly robust estimation

8

Doubly robust multiple imputation

8.1 Motivation

As we noted in §7.6.1, a consistent, efficient and doubly robust estimator is given by the solution to the following estimating equation:

$$\sum_{i=1}^n \left\{ \frac{R_i}{\mathbb{P}(R_i = 1|Z_i)} \mathbf{S}_\theta^F(Z_i, \hat{\theta}^{\text{AIPW}}) + \left[1 - \frac{R_i}{\mathbb{P}(R_i = 1|Z_i)} \right] \mathbb{E} \left[\mathbf{S}_\theta^F(Z_i, \hat{\theta}^{\text{AIPW}}) \mid R_i, G_{R_i}(Z_i) \right] \right\} = 0 \quad (8.1.1)$$

Other than in a few simple situations, however, calculating the conditional expectation in the second term analytically is difficult. Doubly Robust Multiple Imputation (DRMI) is a novel method which tries to overcome the difficulty associated with this step using MI.

8.2 Description of the method

We start by describing the method in the special case where the full data $\mathbf{Z}_i = (V_i, \mathbf{U}_i^T)^T$ for subject i consist of a $(d+1)$ vector \mathbf{U}_i which is always observed and a univariate V_i which is observed only for n_c of the n subjects. As usual, $R_i = 1$ if V_i is observed, and $R_i = 0$ otherwise. Let

$$\mathbf{Z}^c = (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_{n_c})^T$$

be (after re-ordering) the data for the complete cases and let

$$\mathbf{Z}^- = (\mathbf{Z}_1^-, \mathbf{Z}_2^-, \dots, \mathbf{Z}_n^-)^T$$

where $\mathbf{Z}_i^- = (\cdot, \mathbf{U}_i^T)^T$ and \cdot denotes a missing value. Then define an augmented data matrix

$$\mathbf{Z}^* = (\mathbf{Z}^{c^T}, \mathbf{Z}^{-T})^T$$

Thus \mathbf{Z}^* has $n + n_c$ rows, n_c of which are complete. Let \mathbf{V}^* be the first column of \mathbf{Z}^* and \mathbf{U}^* the remaining d columns.

Let $\pi_i = \mathbb{P}(R_i = 1)$ and define weights for the augmented matrix \mathbf{Z}^* as follows:

$$\begin{aligned} W_i^c &= \frac{1}{\pi_i} \\ W_i^- &= R_i \left(1 - \frac{1}{\pi_i}\right) + (1 - R_i) \\ \mathbf{W}^c &= (W_1^c, W_2^c, \dots, W_{n_c}^c)^T \\ \\ \mathbf{W}^- &= (W_1^-, W_2^-, \dots, W_n^-)^T \\ \mathbf{W}^* &= (\mathbf{W}^{c^T}, \mathbf{W}^{-T})^T \end{aligned}$$

The missingness indicator for the augmented dataset is

$$R_i^* = \begin{cases} 1 & \text{if } V_i^* \text{ is observed} \\ 0 & \text{if } V_i^* \text{ is missing} \end{cases}$$

Our proposed method uses multiple imputation to make inference about $\boldsymbol{\theta}$ in a weighted analysis on the augmented data, \mathbf{Z}^* , weighted by \mathbf{W}^* , where $\boldsymbol{\theta}$ is the parameter of interest, governing the distribution of the original full data, \mathbf{Z} . We assume for now that the probabilities π_i are known. The relationship between the original and augmented datasets is illustrated in Fig. 8.1.

As described in §6.7, in multiple imputation, an estimator $\hat{\boldsymbol{\theta}}_j^*$ is obtained from each of m imputed datasets and the MI estimator is given by $\hat{\boldsymbol{\theta}}^{MI} = m^{-1} \sum_{j=1}^m \hat{\boldsymbol{\theta}}_j^*$. In the situation described above, each $\hat{\boldsymbol{\theta}}_j^*$ is the solution to

$$\sum_{i=1}^n W_i^* \mathbf{S}_{\boldsymbol{\theta}}^F \left\{ \mathbf{Z}_{ij}^* \left[\hat{\boldsymbol{\theta}}_I^{(j)} \right], \hat{\boldsymbol{\theta}}_j^* \right\} = \mathbf{0}, \quad (8.2.1)$$

where $\hat{\boldsymbol{\theta}}_I^{(j)}$ is some initial estimator of $\boldsymbol{\theta}$ (which could be different for each j , depending on whether the imputations are proper or improper), and $\mathbf{Z}_{ij}^* \left[\hat{\boldsymbol{\theta}}_I^{(j)} \right]$ is equal to \mathbf{Z}_i^* if

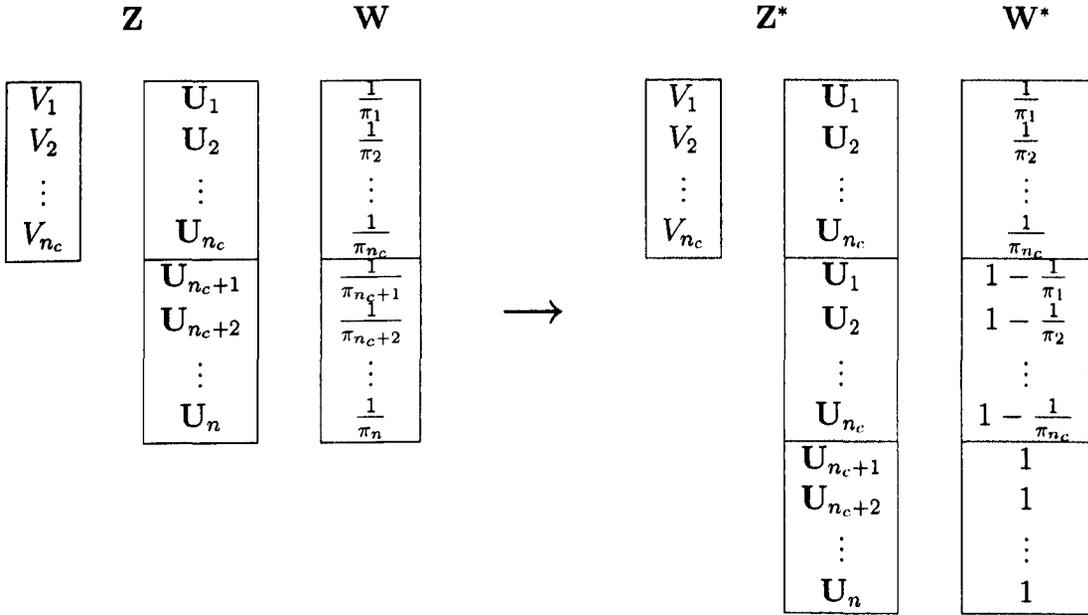


Figure 8.1: A diagrammatic representation of the robust MI formulation

$R_i^* = 1$ and $\tilde{\mathbf{Z}}_{ij}^* = (\tilde{V}_{ij}^*, \mathbf{U}_i)$ if $R_i^* = 0$, where \tilde{V}_{ij}^* is the imputed value of the missing V_i^* as imputed in the j th of the m datasets.

Equation (8.2.1) can be rewritten as

$$\sum_{i=1}^n \left(\frac{R_i}{\pi_i} \mathbf{S}_{\theta}^F(\mathbf{Z}_i^*, \hat{\theta}_j^*) + \left(1 - \frac{R_i}{\pi_i}\right) \mathbf{S}_{\theta}^F\left\{\mathbf{Z}_{ij}^* \left[\hat{\theta}_I^{(j)}\right], \hat{\theta}_j^*\right\}\right) = \mathbf{0}, \quad (8.2.2)$$

which is very similar to (8.1.1). Recall (from §6.7) that $\tilde{V}_{ij}^* = \mathbb{E}\left[V_i \mid \mathbf{U}_i, \hat{\theta}_I^{(j)}\right] + \varepsilon_{ij}$, where ε_{ij} has zero expectation. Thus the main difference between (8.1.1) and (8.2.2) is the fact that the expectation in (8.1.1) has been taken inside the score function in (8.2.2). This is analogous to the difference between the observed-data score equation (7.2.2):

$$\sum_{i=1}^n \left\{ R_i \mathbf{S}_{\theta}^F(\mathbf{Z}_i, \theta) + (1 - R_i) \mathbb{E}\left[\mathbf{S}_{\theta}^F(\mathbf{Z}_i, \theta) \mid \mathbf{Z}_i^{\text{obs}}\right] \right\} = \mathbf{0}$$

and the ordinary multiple imputation estimating equation (for imputed data set j):

$$\sum_{i=1}^n \left[R_i \mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta}) + (1 - R_i) \mathbf{S}_{\boldsymbol{\theta}}^F(\tilde{\mathbf{Z}}_i, \boldsymbol{\theta}) \right] = \mathbf{0}$$

Because of this, we wouldn't expect robust MI to perform as well in finite samples as analytically-derived doubly-robust estimating procedures. It should, however, be much easier to implement, especially in complex situations, and the effect of taking the expectation inside the score function diminishes as the sample size increases for the same reasons (see Wang and Robins, 1998; Robins and Wang, 2000; Tsiatis, 2006) that ordinary multiple imputation estimators are consistent.

We carried out a simulation study (results not shown) to assess the properties of this proposed robust MI estimator. To facilitate comparison, we used the simple bivariate normal example, where the AIPW estimator can be analytically derived.

In this example, we let

$$\mathbf{Z}_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix} \stackrel{\text{iid}}{\sim} N_2 \left[\begin{pmatrix} 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 0.5 & 1 \\ 1 & 2.5 \end{pmatrix} \right]$$

and looked at the parameters the regression of Y on X , with Y fully-observed for all individuals, but X missing for some individuals.

In simulations, the robust MI procedure performed very well and almost as well as the doubly-robust estimator in terms of bias and precision. However, Rubin's variance formula for ordinary MI, when applied to robust MI is considerably biased even with a sample size of 10,000. It is not surprising that Rubin's variance formula fails: the data are far from being i.i.d. and a model that would generate the 'full' data in our augmented dataset is inconceivable.

It is clear that a better variance formula is needed if this procedure is to be of use in

practice. This requires a careful derivation of the true variance of the robust multiple imputation estimator.

8.3 Variance estimation

This section is closely based on Tsiatis (2006), Chapter 14, but what Tsiatis shows for ordinary multiple imputation is adapted here to robust multiple imputation. To simplify the derivations, we assume for the remainder of this chapter that the missingness probability π_i is a known, fixed value, specific to each subject, and not a function of the data.

Lemma 8.1. *If $\hat{\theta}^{\text{rob}}$ is the robust MI estimator of θ , then*

$$\begin{aligned} \therefore n^{\frac{1}{2}} \left(\hat{\theta}^{\text{rob}} - \theta_0 \right) &= n^{-\frac{1}{2}} \sum_{i=1}^n \left[I_{\theta\theta}^F(\theta_0) \right]^{-1} \left[m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\mathbf{Z}_{ij}^*(\theta_0), \theta_0 \right] \right. \\ &+ \left(\frac{1 - \pi_i}{\pi_i} \right) \left\{ m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\mathbf{Z}_{ij}^*(\theta_0), \theta_0 \right] - m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\tilde{\mathbf{Z}}_{ij}^*(\theta_0), \theta_0 \right] \right\} \\ &+ m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left\{ \mathbf{Z}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} \\ &- m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\mathbf{Z}_{ij}^*(\theta_0), \theta_0 \right] \\ &\left. - \left(\frac{1 - \pi_i}{\pi_i} \right) \left(m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left\{ \tilde{\mathbf{Z}}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} - m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\tilde{\mathbf{Z}}_{ij}^*(\theta_0), \theta_0 \right] \right) \right] + o_p(1) \end{aligned}$$

The proof of this Lemma is given in Appendix A.1.

Lemma 8.2 (Influence function for improper robust multiple imputation). *The i th*

influence function for the improper robust improper imputation estimator is

$$\begin{aligned}
& [I_{\theta\theta}^F(\boldsymbol{\theta}_0)]^{-1} \left\{ \frac{1}{\pi_i} m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F[\mathbf{Z}_{ij}^*(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0] - \left(\frac{1 - \pi_i}{\pi_i} \right) m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F[\tilde{\mathbf{Z}}_{ij}^*(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0] \right. \\
& \left. + [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}(\boldsymbol{\theta}_0)] q[R_i, G_{R_i}(\mathbf{Z}_i)] - \overline{\left(\frac{1 - \pi_i}{\pi_i} \right)} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}^U(\boldsymbol{\theta}_0)] q[R_i, G_{R_i}(\mathbf{Z}_i)] \right\}
\end{aligned} \tag{8.3.1}$$

The proof of this Lemma is given in Appendix A.2.

Lemma 8.3 (Variance of the i th influence function for improper robust multiple imputation). *The variance of the i th influence function (8.3.1) is given by*

$$\begin{aligned}
& [I_{\theta\theta}^F(\boldsymbol{\theta}_0)]^{-1} \left(\frac{1}{\pi_i^2} \{ m^{-1} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}(\boldsymbol{\theta}_0)] + I_{\theta\theta}(\boldsymbol{\theta}_0) \} \right. \\
& + \left(\frac{1 - \pi_i}{\pi_i} \right)^2 \{ m^{-1} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}^U(\boldsymbol{\theta}_0)] + I_{\theta\theta}^U(\boldsymbol{\theta}_0) \} \\
& + [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}(\boldsymbol{\theta}_0)] \text{Var} \{ q[R_i, G_{R_i}(\mathbf{Z}_i)] \} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}(\boldsymbol{\theta}_0)] \\
& + \overline{\left(\frac{1 - \pi_i}{\pi_i} \right)}^2 [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}^U(\boldsymbol{\theta}_0)] \text{Var} \{ q[R_i, G_{R_i}(\mathbf{Z}_i)] \} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}^U(\boldsymbol{\theta}_0)] \\
& - 2 \left(\frac{1 - \pi_i}{\pi_i^2} \right) \{ m^{-1} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}(\boldsymbol{\theta}_0)] + I_{\theta\theta}^U(\boldsymbol{\theta}_0) \} + \frac{2}{\pi_i} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}(\boldsymbol{\theta}_0)] \\
& - \frac{2}{\pi_i} \overline{\left(\frac{1 - \pi_i}{\pi_i} \right)} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}^U(\boldsymbol{\theta}_0)] - \frac{1 - \pi_i}{\pi_i} I_{\theta\theta}^U(\boldsymbol{\theta}_0) [I_{\theta\theta}^F(\boldsymbol{\theta}_0)]^{-1} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}(\boldsymbol{\theta}_0)] \\
& - \frac{1 - \pi_i}{\pi_i} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}(\boldsymbol{\theta}_0)] [I_{\theta\theta}^F(\boldsymbol{\theta}_0)]^{-1} I_{\theta\theta}^U(\boldsymbol{\theta}_0) \\
& + 2 \left(\frac{1 - \pi_i}{\pi_i} \right) \overline{\left(\frac{1 - \pi_i}{\pi_i} \right)} I_{\theta\theta}^U(\boldsymbol{\theta}_0) [I_{\theta\theta}^F(\boldsymbol{\theta}_0)]^{-1} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}^U(\boldsymbol{\theta}_0)] \\
& - \overline{\left(\frac{1 - \pi_i}{\pi_i} \right)} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}(\boldsymbol{\theta}_0)] \text{Var} \{ q[R_i, G_{R_i}(\mathbf{Z}_i)] \} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}^U(\boldsymbol{\theta}_0)] \\
& \left. - \overline{\left(\frac{1 - \pi_i}{\pi_i} \right)} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}^U(\boldsymbol{\theta}_0)] \text{Var} \{ q[R_i, G_{R_i}(\mathbf{Z}_i)] \} [I_{\theta\theta}^F(\boldsymbol{\theta}_0) - I_{\theta\theta}(\boldsymbol{\theta}_0)] \right) [I_{\theta\theta}^F(\boldsymbol{\theta}_0)]^{-1}
\end{aligned} \tag{8.3.2}$$

The proof of this Lemma is given in Appendix A.3.

Claim 8.4. *The mean of the variances of the n influence function is asymptotically equal to the variance of $n^{\frac{1}{2}} (\hat{\boldsymbol{\theta}}^{*\text{rob}} - \boldsymbol{\theta}_0)$.*

Proof. From Definition 3.5, we have that

$$n^{\frac{1}{2}} (\hat{\boldsymbol{\theta}}^{*\text{rob}} - \boldsymbol{\theta}_0) = n^{-\frac{1}{2}} \sum_{i=1}^n \boldsymbol{\varphi}(\mathbf{Z}_i) + o_p(1)$$

where $\boldsymbol{\varphi}(\mathbf{Z}_i)$ is the i th influence function of $\hat{\boldsymbol{\theta}}^{*\text{rob}}$.

Usually we require that $\{\boldsymbol{\varphi}(\mathbf{Z}_i) : i = 1, \dots, n\}$ be i.i.d. but in the way we have constructed $\boldsymbol{\varphi}(\mathbf{Z}_i)$, they are independent but *not* identically distributed.

Thus,

$$\text{Var} \left[n^{\frac{1}{2}} (\hat{\boldsymbol{\theta}}^{*\text{rob}} - \boldsymbol{\theta}_0) \right] \rightarrow n^{-1} \sum_{i=1}^n \text{Var} [\boldsymbol{\varphi}(\mathbf{Z}_i)]$$

□

This means that if we can evaluate (8.3.2) from the data, then we know the asymptotic variance of the improper robust MI estimator. In order to do this, we must estimate $\text{Var} \{q[R_i, G_{R_i}(\mathbf{Z}_i)]\}$, $I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0)$, $I_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0)$ and $I_{\boldsymbol{\theta}\boldsymbol{\theta}}^U(\boldsymbol{\theta}_0)$.

Seeing as $\hat{\boldsymbol{\theta}}_I$ comes from a simple analysis on the complete cases, an estimate $\hat{\text{Var}} \{q[R_i, G_{R_i}(\mathbf{Z}_i)]\}$ of $\text{Var} \{q[R_i, G_{R_i}(\mathbf{Z}_i)]\}$ should be readily available. Also, Tsiatis (2006) shows that we can use

$$\hat{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) = -m^{-1} \sum_{j=1}^m \left\{ n^{-1} \sum_{i=1}^n \frac{\partial \mathbf{S}_{\boldsymbol{\theta}}^F \left[\mathbf{Z}_{ij}^* (\hat{\boldsymbol{\theta}}_I), \hat{\boldsymbol{\theta}}_j^* \right]}{\partial \boldsymbol{\theta}^T} \right\}$$

to estimate $I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0)$.

As for $I_{\theta\theta}(\theta_0)$, he suggests using

$$n^{-1} \sum_{i=1}^n (m-1)^{-1} \cdot \sum_{j=1}^m \left\{ \mathbf{S}_{\theta}^F \left[\mathbf{Z}_{ij}^* (\hat{\theta}_I), \hat{\theta}_j^* \right] - \bar{S}_{\theta_i}^F (\hat{\theta}^*) \right\} \left\{ \mathbf{S}_{\theta}^F \left[\mathbf{Z}_{ij}^* (\hat{\theta}_I), \hat{\theta}_j^* \right] - \bar{S}_{\theta_i}^F (\hat{\theta}^*) \right\}^T$$

to estimate $I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)$.

Analogously, we suggest using

$$n^{-1} \sum_{i=1}^n (m-1)^{-1} \cdot \sum_{j=1}^m \left\{ \mathbf{S}_{\theta}^F \left[\tilde{\mathbf{Z}}_{ij}^* (\hat{\theta}_I), \hat{\theta}_j^* \right] - \bar{S}_{\theta_i}^F (\hat{\theta}^*) \right\} \left\{ \mathbf{S}_{\theta}^F \left[\tilde{\mathbf{Z}}_{ij}^* (\hat{\theta}_I), \hat{\theta}_j^* \right] - \bar{S}_{\theta_i}^F (\hat{\theta}^*) \right\}^T$$

to estimate $I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}^U(\theta_0)$.

In the case where $\hat{\theta}_I^{(j)}$ is sampled from the posterior distribution $p[\theta | R_i, G_{R_i}(\mathbf{Z}_i)]$. We will assume that the sample is large enough for $\hat{\theta}_I^{(j)}$ to be from

$$N \left(\hat{\theta}_I^{\text{improper}}, \frac{\hat{\text{Var}} \{q[R_i, G_{R_i}(\mathbf{Z}_i)]\}}{n} \right)$$

Lemma 8.5 (Variance of the proper robust multiple imputation estimator). *The variance of $n^{\frac{1}{2}} (\hat{\theta}^{*\text{rob}} - \theta_0)$ for robust proper MI is given by*

$$n^{-\frac{1}{2}} \sum_{i=1}^n [I_{\theta\theta}^F(\theta_0)]^{-1} \left(\frac{1}{\pi_i^2} \{m^{-1} [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)] + I_{\theta\theta}(\theta_0)\} \right. \\ \left. + \left(\frac{1 - \pi_i}{\pi_i} \right)^2 \{m^{-1} [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}^U(\theta_0)] + I_{\theta\theta}^U(\theta_0)\} \right)$$

P.T.O.

$$\begin{aligned}
& + \left(\frac{m+1}{m} \right) [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)] \text{Var} \{q[R_i, G_{R_i}(\mathbf{Z}_i)]\} [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)] \\
& + \left(\frac{m+1}{m} \right) \left(\frac{1-\pi_i}{\pi_i} \right)^2 [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}^U(\theta_0)] \text{Var} \{q[R_i, G_{R_i}(\mathbf{Z}_i)]\} [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}^U(\theta_0)] \\
& - 2 \left(\frac{1-\pi_i}{\pi_i^2} \right) \left\{ m^{-1} [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)] + I_{\theta\theta}^U(\theta_0) \right\} + \frac{2}{\pi_i} [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)] \\
& - \frac{2}{\pi_i} \left(\frac{1-\pi_i}{\pi_i} \right) [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}^U(\theta_0)] - \frac{1-\pi_i}{\pi_i} I_{\theta\theta}^U(\theta_0) [I_{\theta\theta}^F(\theta_0)]^{-1} [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)] \\
& - \frac{1-\pi_i}{\pi_i} [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)] [I_{\theta\theta}^F(\theta_0)]^{-1} I_{\theta\theta}^U(\theta_0) \\
& + 2 \left(\frac{1-\pi_i}{\pi_i} \right) \left(\frac{1-\pi_i}{\pi_i} \right) I_{\theta\theta}^U(\theta_0) [I_{\theta\theta}^F(\theta_0)]^{-1} [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}^U(\theta_0)] \\
& - \left(\frac{m+1}{m} \right) \left(\frac{1-\pi_i}{\pi_i} \right) [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)] \text{Var} \{q[R_i, G_{R_i}(\mathbf{Z}_i)]\} [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}^U(\theta_0)] \\
& - \left(\frac{m+1}{m} \right) \left(\frac{1-\pi_i}{\pi_i} \right) [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}^U(\theta_0)] \text{Var} \{q[R_i, G_{R_i}(\mathbf{Z}_i)]\} [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)] \\
& \cdot [I_{\theta\theta}^F(\theta_0)]^{-1}
\end{aligned}$$

The proof of this Lemma is given in Appendix A.4.

8.4 Discussion

The basic idea of using multiple imputation in the way described above to obtain approximate doubly-robust estimates is very appealing and simulations (not shown) demonstrate that the bias and precision of the estimates compare well with the true doubly-robust procedure.

We have failed, however, to derive a Rubin-type variance estimator, i.e. a variance estimator similar to (6.7.1), even for proper imputation. The fact that (8.3.2) is so complicated and would take many steps for the user to calculate, probably means that

the method as it stands is not of much use in practice.

It should be possible to obtain valid estimates of variance using the bootstrap, but combining bootstrapping with multiple imputation would be computationally very intensive. As will be seen in the next chapter, going down this road is not necessary since by re-formulating doubly-robust estimation using the regression representation proposed by Bang and Robins (2005), multiple imputation for doubly-robust estimation can be much more successful.

9

Robust multiple imputation: an alternative formulation

9.1 Introduction

In §7.6.4, we described the method proposed by Bang and Robins (2005) for constructing doubly robust (DR) estimators. One limitation of their approach is that the bootstrap is required to obtain estimates of variance. Another limitation is that

the method does not extend to non-monotone missingness patterns. One further (and smaller) limitation is that when the number of partially-observed variables is greater than one and these partially observed variables are binary or categorical, their claim that the method can be applied using ‘off-the-shelf regression software’ is not quite true. We return to this point in Chapters 12 and 13 on binary data. Finally, for longitudinal monotone patterns, we discuss an important feature not made explicit by Bang and Robins in their paper, namely that their formulation requires the specification of suitable regressions for later outcome variables, conditional on some earlier outcomes, marginalised over intermediate values of the outcome. We consider this to be unnatural and potentially difficult when the form of the conditional distribution of the later variable given all previous outcomes is a non-linear function of the intermediate variables.

In the previous chapter, we suggested using multiple imputation to facilitate the approximation of doubly robust estimators, but the method failed to be practical because of the intractable form of the variance of this estimator. In this chapter we propose an alternative formulation of doubly robust MI—based on the Bang and Robins formulation—with the aim of overcoming all the limitations listed above. The proposed method can be implemented using existing MI software and is very flexible. We start by describing doubly robust multiple imputation (DRMI) in the cases described by Bang and Robins (2005) before going on to describe DRMI in broader settings. Finally, we confirm the theoretical properties of our estimator using simulation studies.

9.2 The proposed method

9.2.1 Univariate ignorable missing data

Let the full data $\mathbf{Z}_i = (\mathbf{X}_i^T, Y_i)^T$ for subject $i \in \{1, \dots, n\}$ be a fully-observed vector of covariates \mathbf{X}_i and a scalar outcome Y_i which could be missing ($R_i = 0$) or observed ($R_i = 1$) and interest lies in estimating $\mu = \mathbb{E}(Y_i)$.

Following the same idea as proposed by Bang and Robins (2005), first a suitable regression model (such as logistic regression) is chosen for R conditional on \mathbf{X} —the π -model. Let $\hat{\boldsymbol{\alpha}}$ be the parameter estimates from this regression and let $\hat{\pi}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}})$ be the predicted probabilities (that $R_i = 1$) from this model.

Next, we fit a suitable regression model for Y conditional on \mathbf{X} and $\hat{\pi}^{-1}$ to those subjects who have complete data. We call the corresponding model *without* the inverse probability weights, i.e.

$$\mathbb{E}(Y_i | \mathbf{X}_i, R_i = 1) = \Psi[s(\mathbf{X}_i, \boldsymbol{\beta})] \quad (9.2.1)$$

the y -model, where $\Psi^{-1}(\cdot)$ is the canonical link function from an appropriate GLM and $s(\mathbf{X}, \boldsymbol{\beta})$ is a known function of $\boldsymbol{\beta}$ and \mathbf{X} . We call

$$\mathbb{E}[Y_i | \mathbf{X}_i, \hat{\pi}^{-1}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}), R_i = 1] = \Psi[s(\mathbf{X}_i, \boldsymbol{\beta}) + \phi \hat{\pi}^{-1}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}})]$$

the extended y -model.

Let us write $\hat{\pi}_i$ for $\hat{\pi}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}})$ and let

$$\hat{e}(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}) = \Psi[s(\mathbf{X}_i, \hat{\boldsymbol{\beta}}) + \hat{\phi} \hat{\pi}_i^{-1}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}})]$$

be the predictions from the extended y -model.

Now we draw $m > 1$ imputations for each of the missing values of Y based on the extended y -model. For example, if Y is continuous, and the y -model a linear regression, let $\hat{V}_{(\hat{\boldsymbol{\beta}}, \hat{\phi})}$ be the estimated variance-covariance matrix of $(\hat{\boldsymbol{\beta}}, \hat{\phi})$ and $\hat{V}_{Y|\mathbf{X}, \hat{\pi}^{-1}}$ be the estimator from the extended y -model of

$$\text{Var}[Y_i | \mathbf{X}_i, \hat{\pi}^{-1}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}), R_i = 1]$$

We draw m times from the large-sample approximation to the posterior distribution

of $(\hat{\beta}, \hat{\phi})$:

$$[\beta^{(j)}, \phi^{(j)}] \stackrel{\text{i.i.d.}}{\sim} N \left[(\hat{\beta}, \hat{\phi}), \hat{V}_{(\hat{\beta}, \hat{\phi})} \right], \quad j = 1, \dots, m$$

and m times from the large-sample approximation to the posterior distribution of $\hat{V}_{Y|\mathbf{X}, \hat{\pi}^{-1}}$:

$$V_{Y|\mathbf{X}, \hat{\pi}^{-1}}^{(j)} \stackrel{\text{i.i.d.}}{\sim} \hat{V}_{Y|\mathbf{X}, \hat{\pi}^{-1}} \frac{1}{n_c - p} \chi_{n_c - p}^{2-1}, \quad j = 1, \dots, m$$

where $n_c = \sum_{i=1}^n R_i$ and p is the number of parameters estimated in the extended y -model.

m imputed datasets are then generated with $\tilde{Y}_i^{(j)}$ replacing Y in the j^{th} dataset where

$$\tilde{Y}_i^{(j)} = R_i Y_i + (1 - R_i) \left\{ \hat{e} [\mathbf{X}_i^T, \beta^{(j)}, \phi^{(j)}, \hat{\pi}_i^{-1}] + \varepsilon_i^{(j)} \right\}$$

and $\varepsilon_i^{(j)} \stackrel{\text{i.i.d.}}{\sim} N \left[0, V_{Y|\mathbf{X}, \hat{\pi}^{-1}}^{(j)} \right]$.

When the y -model is not a linear regression model, the imputations are drawn *properly* according to the appropriate imputation distribution.

Finally, our proposed estimator is the solution $\hat{\mu}_{\text{DRMI}}$ to

$$\sum_{j=1}^m \sum_{i=1}^n \left[\tilde{Y}_i^{(j)} - \mu_{\text{DRMI}} \right] = 0$$

Theorem 9.1 (Multiply imputed DR univariate estimator). *The estimator $\hat{\mu}_{\text{DRMI}}$ is doubly robust. That is, if at least one of the two models (the π -model and the y -model) is correctly specified (but not necessarily both), $\hat{\mu}_{\text{DRMI}}$ is a consistent estimator of μ .*

Sketch proof. The consistency of $\hat{\mu}_{\text{DRMI}}$ when the y -model is correctly specified follows (as in the proof of Theorem 7.2) from the fact that the true value of ϕ is zero. If the π -model is correctly specified, but not the y -model, it is slightly less evident that $\hat{\mu}_{\text{DRMI}}$ remains consistent.

We continue to write $\hat{\pi}_i$ for $\hat{\pi}(\mathbf{X}_i, \hat{\alpha})$. The DRMI estimating equation

$$\sum_{j=1}^m \sum_{i=1}^n \left[R_i Y_i + (1 - R_i) \tilde{Y}_i^{(j)} - \mu_{\text{DRMI}} \right] = 0$$

can be rewritten as

$$\begin{aligned} & \sum_{j=1}^m \sum_{i=1}^n \left\{ R_i \left[Y_i - \hat{e} \left(\mathbf{X}_i^T, \hat{\beta}, \hat{\phi}, \hat{\pi}_i^{-1} \right) \right] + R_i \hat{\pi}_i^{-1} \left[Y_i - \hat{e} \left(\mathbf{X}_i^T, \hat{\beta}, \hat{\phi}, \hat{\pi}_i^{-1} \right) \right] \right. \\ & \left. + (1 - R_i) \left[\tilde{Y}_i^{(j)} - \hat{e} \left(\mathbf{X}_i^T, \hat{\beta}, \hat{\phi}, \hat{\pi}_i^{-1} \right) \right] + \hat{e} \left(\mathbf{X}_i^T, \hat{\beta}, \hat{\phi}, \hat{\pi}_i^{-1} \right) - \mu_{\text{DRMI}} \right\} = 0 \quad (9.2.2) \end{aligned}$$

This follows from the fact that $\sum_{j=1}^m \sum_{i=1}^n R_i \hat{\pi}_i^{-1} (Y_i - e_i)$ is numerically zero since we included $\hat{\pi}_i^{-1}$ in our extended y -model GLM.

$\sum_{j=1}^m \sum_{i=1}^n R_i \left[Y_i - \hat{e} \left(\mathbf{X}_i^T, \hat{\beta}, \hat{\phi}, \hat{\pi}_i^{-1} \right) \right]$ is also numerically zero, assuming that a constant term is included in our GLM. Furthermore, $(1 - R_i) \left[\tilde{Y}_i^{(j)} - \hat{e} \left(\mathbf{X}_i^T, \hat{\beta}, \hat{\phi}, \hat{\pi}_i^{-1} \right) \right]$ has zero expectation, since the proper imputations have been drawn from the posterior predictive distribution with mean $\hat{e} \left(\mathbf{X}_i^T, \hat{\beta}, \hat{\phi}, \hat{\pi}_i^{-1} \right)$ (see (6.7.2)). Thus we can rewrite (9.2.2) as

$$\begin{aligned} & \sum_{j=1}^m \sum_{i=1}^n \left\{ R_i \hat{\pi}_i^{-1} (Y_i - \mu_{\text{DRMI}}) + (1 - R_i \hat{\pi}_i^{-1}) \left[\hat{e} \left(\mathbf{X}_i^T, \hat{\beta}, \hat{\phi}, \hat{\pi}_i^{-1} \right) - \mu_{\text{DRMI}} \right] \right. \\ & \left. + (1 - R_i) \left[\tilde{Y}_i^{(j)} - \hat{e} \left(\mathbf{X}_i^T, \hat{\beta}, \hat{\phi}, \hat{\pi}_i^{-1} \right) \right] \right\} = 0 \end{aligned}$$

which we immediately recognise as being of the same form as (7.6.9) with the added term $(1 - R_i) \left[\tilde{Y}_i^{(j)} - \hat{e} \left(\mathbf{X}_i^T, \hat{\beta}, \hat{\phi}, \hat{\pi}_i^{-1} \right) \right]$ which has zero expectation, even when the y -model is incorrect. Thus, $\hat{\mu}_{\text{DRMI}}$ is consistent whenever the π -model is correctly specified. \square

We propose that $\text{Var}(\hat{\mu}_{\text{DRMI}})$ be estimated using Rubin's variance formula (see §6.7) as

$$\frac{1}{m} \sum_{j=1}^m \sum_{i=1}^n \frac{[\tilde{Y}_i^{(j)} - \bar{Y}^{(j)}]^2}{n-1} + \frac{m+1}{m} \sum_{j=1}^m \frac{[\bar{Y}^{(j)} - \bar{Y}]^2}{m-1} \quad (9.2.3)$$

where $\bar{Y}^{(j)} = \frac{1}{n} \sum_{i=1}^n \tilde{Y}_i^{(j)}$ and $\bar{Y} = \frac{1}{m} \sum_{j=1}^m \bar{Y}^{(j)}$.

However, this variance estimator has two important drawbacks:

1. It treats the weights as just another covariate in the imputation model. Thus the variance estimator is conditional on $\hat{\pi}^{-1}(X_i, \hat{\alpha})$ and ignores the fact that these weights are estimated from the data.
2. Putting this problem to one side, when the y -model is correctly specified, the fact that the weights are treated as just another covariate justifies the use of Rubin's variance formula. In other words, if the weights were not estimated, the correct specification of the y -model would render (9.2.3) a consistent estimator of the variance of $\hat{\mu}_{\text{DRMI}}$, by the standard argument for the consistency of Rubin's variance formula in (non-DR) ordinary multiple imputation. However, if the y -model is misspecified, but the π -model correctly specified, there is no reason to suppose that (9.2.3) remains consistent. Hence, our proposed variance formula is (ignoring the added problem noted in 1.) singly robust, but does not inherit the DR property of the estimator itself.

9.2.2 Longitudinal ignorable missing data

The same idea can be extended to the case of multivariate missing data, and—unlike the Bang and Robins (2005) approach—the pattern need not be monotone.

Let the full data $\mathbf{Z}_i = (\mathbf{X}_i^T, \mathbf{Y}_i^T)^T$ for subject $i \in \{1, \dots, n\}$ consist of a fully-observed vector of covariates \mathbf{X}_i and a vector of partially-observed outcome variables

$\mathbf{Y}_i = (Y_{1,i}, \dots, Y_{T,i})^T$ and that interest lies in estimating $\mu = \mathbb{E}(Y_{i,T})$. Let $\mathbf{R}_i = (R_{1,i}, \dots, R_{T,i})^T$ be the vector of missingness indicators with $R_{t,i} = \mathbf{1}(Y_{t,i} \text{ is observed})$.

We first describe the DRMI method for monotone longitudinal data before moving to the case of non-monotone longitudinal data in §9.2.2.2.

9.2.2.1 Monotone longitudinal data

When the missingness pattern is monotone, we can easily estimate $\hat{\pi}_{t,i} = \mathbb{P}(R_{t,i} = 1 | \mathbf{Z}_i) = \mathbb{P}(R_{t,i} = 1 | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i})$ at each time t , as described in §7.6.4.2, for example by fitting a logistic regression model to R_t conditional on $\mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}$ to those subjects with $R_{t-1} = 1$. The marginal probabilities $\hat{\pi}_{t,i}$ are then obtained as a product of these conditional probabilities, as described in §7.6.4.2.

We proceed by fitting the model using MI. The y -model is postulated sequentially by first specifying a model for Y_1 given \mathbf{X} , and then a model for Y_2 given Y_1 and \mathbf{X} etc. To construct an extended y -model, for each $t \in \{1, \dots, T\}$, $\hat{\pi}_{t,i}^{-1} = \mathbb{P}(R_{t,i} = 1 | \mathbf{X}_i, Y_{1,i}, \dots, Y_{t-1,i})^{-1}$ is included as an additional covariate, additional to \mathbf{X} and $\bar{\mathbf{Y}}_{t-1}$, in the model for $Y_{t,i}$. Starting with Y_1 , any missing values in Y_1 are multiply imputed, with the imputations drawn from the extended y -model for Y_1 conditional on \mathbf{X} and $\hat{\pi}_1^{-1}$. Next, any missing values in Y_2 are multiply imputed, with the imputations drawn from the extended y -model for Y_2 conditional on Y_1, \mathbf{X} and $\hat{\pi}_2^{-1}$; for subjects with Y_1 also missing, the imputed value of Y_1 from the j th imputed dataset is used to impute Y_2 in the j th imputed dataset, and so on.

By starting with Y_1 and working upwards in this way, we encounter a problem which does not arise in the method proposed by Bang and Robins (2005), which starts with Y_T and works downwards. The problem is that $\hat{\pi}_{t,i}$ can only be calculated for subjects who have $Y_{t-1,i}$ observed, but (unlike Bang and Robins (2005)), we require that $\hat{\pi}_{t,i}$ be known for all subjects.

Suppose a particular subject, i_1 , drops out after being observed at time $t - 2$. At time $t - 1$, in the j th imputed dataset, a value $\tilde{Y}_{t-1,i_1}^{(j)}$ of Y_{t-1,i_1} is imputed, based on \mathbf{X}_{i_1} , $\bar{\mathbf{Y}}_{t-2,i_1}$, and $\hat{\pi}_{t-1,i_1}$, which are all observed. But at the next timepoint, t , we would like to impute the missing Y_{t,i_1} using \mathbf{X}_{i_1} , $\bar{\mathbf{Y}}_{t-2,i_1}$, $\tilde{Y}_{t-1,i_1}^{(j)}$, and $\hat{\pi}_{t,i_1}$. The marginal probability $\hat{\pi}_{t,i_1}$ is the product of $\hat{\pi}_{t-1,i_1}$ and $\hat{\lambda}(t | \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t-1,i_1})$, the estimate of the conditional probability that $R_{t,i_1} = 1$, conditional on \mathbf{X}_{i_1} , $\bar{\mathbf{Y}}_{t-1,i_1}$, and $R_{t-1,i_1} = 1$, as defined on page 108. It is this latter conditional probability which cannot be estimated directly for this subject. However, as a function of the missing Y_{t-1,i_1} , it is known. Thus our proposed method works by imputing a value for $\hat{\pi}_{t,i_1}$, based on $\hat{\pi}_{t-1,i_1}$, $\hat{\lambda}(t | \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t-1,i_1})$ and $\tilde{Y}_{t-1,i_1}^{(j)}$ as follows:

$$\hat{\pi}_{t,i_1}^{(j)} = \hat{\pi}_{t-1,i_1} \hat{\lambda}(t | \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t-2,i_1}, \tilde{Y}_{t-1,i_1}^{(j)})$$

In other words, no additional model is fitted to obtain the imputation $\hat{\pi}_{t,i_1}^{(j)}$, and no additional draws (for $\hat{\pi}_{t,i_1}^{(j)}$), nor additional draws from the Bayesian posterior distribution of any additional parameters are made. Rather, $\hat{\pi}_{t,i_1}^{(j)}$ is imputed as a deterministic function of $\hat{\pi}_{t-1,i_1}$ and $\hat{\lambda}(t | \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t-2,i_1}, \tilde{Y}_{t-1,i_1}^{(j)})$, which, as function of \mathbf{X}_i and $\bar{\mathbf{Y}}_{t-1,i}$, is estimated using subjects who have Y_{t-1} observed, as previously. This deterministic imputation is analogous to the way in which quadratic functions of covariates, say, are dealt with in ordinary multiple imputation. If X and X^2 are both covariates in the analysis model, multiple imputations $X_i^{(j)}$ of any missing X_i are obtained in the ordinary way, but then the imputed value of X_i^2 is simply $[X_i^{(j)}]^2$, the square of the imputation.

Similarly, for subject i_1 at time $t + 1$, our proposed method works by first imputing a value for $\hat{\pi}_{t+1,i_1}$, based on $\hat{\pi}_{t,i_1}^{(j)}$, $\hat{\lambda}(t + 1 | \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t,i_1})$, $\tilde{Y}_{t-1,i_1}^{(j)}$ and $\tilde{Y}_{t,i_1}^{(j)}$ as follows:

$$\hat{\pi}_{t+1,i_1}^{(j)} = \hat{\pi}_{t,i_1}^{(j)} \hat{\lambda}(t + 1 | \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t-2,i_1}, \tilde{Y}_{t-1,i_1}^{(j)}, \tilde{Y}_{t,i_1}^{(j)})$$

and then Y_{t+1,i_1} is imputed using \mathbf{X}_{i_1} , $\bar{\mathbf{Y}}_{t-2,i_1}$, $\tilde{Y}_{t-1,i_1}^{(j)}$, $\tilde{Y}_{t,i_1}^{(j)}$ and $\hat{\pi}_{t+1,i_1}^{(j)}$.

Finally, $\hat{\mu}_{\text{DRMI}}$ can be calculated as the solution to

$$\sum_{j=1}^m \sum_{i=1}^n \left[\tilde{Y}_{T,i}^{(j)} - \mu_{\text{DRMI}} \right] = 0 \quad (9.2.4)$$

and a variance estimate analogous to (9.2.3) obtained using Rubin's variance formula. The same caveats that this variance estimator does not acknowledge the uncertainty due to the fact that the weights have been estimated, and (even ignoring this problem) is only singly robust, applies equally here as in the univariate case.

Let $H_t \left(\mathbf{X}_i^T, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}} \right)$ be the predictions from the Bang and Robins procedure for longitudinal monotone data (as described in §7.6.4.2) after $T - t$ iterations of step 3(a). Let $\hat{\mathbb{E}} \left(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i} \right)$ be the mean of the distribution from which the DRMI imputations for $Y_{T,i}$, for a subject who drops out after time t , are drawn.

Lemma 9.2.

$$\mathbb{E} \left[H_t \left(\mathbf{X}_i^T, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i}^{-1}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}} \right) \right] = \mathbb{E} \left[\hat{\mathbb{E}} \left(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i} \right) \right]$$

where expectations are taken with respect to the true distribution of $(\mathbf{X}_i, \bar{\mathbf{Y}}_{T,i})$.

Sketch proof. That Lemma 9.2 is true is immediate if the y -model is correct, since both $H_t \left(\mathbf{X}_i^T, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i}^{-1}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}} \right)$ and $\hat{\mathbb{E}} \left(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i} \right)$ are consistent estimators of $\mathbb{E} \left(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i} \right)$. However, the argument (see Tsiatis, 2006, ch. 14) showing that multiple imputation recovers the full-data distribution when the imputation model is correctly specified can also be used to show that when it is incorrectly specified, the incorrect distribution it recovers is equivalent to the hypothetical full-data distribution implied by that incorrectly specified imputation distribution. \square

Theorem 9.3 (Multiply imputed DR monotone longitudinal estimator). *The estimator $\hat{\mu}_{\text{DRMI}}$ is doubly robust. That is, if at least one of the two models (the π -model and the y -model) is correctly specified (but not necessarily both), $\hat{\mu}_{\text{DRMI}}$ is a consistent estimator of μ .*

Proof. As for the univariate case, that $\hat{\mu}_{\text{DRMI}}$ is consistent when only the π -model is misspecified is intuitively obvious. We therefore concentrate on the consistency of $\hat{\mu}_{\text{DRMI}}$ when only the y -model is misspecified.

Assuming that Y_1 is always observed, that D_i is the dropout indicator (as defined in Definition 3.16), and that $\bar{\mathbf{Z}}_{t,i}$ denotes the history of \mathbf{Z}_i up to and including t (as defined in Definition 3.17), the general form of the AIPW estimating equation (as described by Tsiatis, 2006, p.208) can be written as

$$\sum_{i=1}^n \left\{ \frac{\mathbb{1}(D_i = T + 1)}{\mathbb{P}(D_i = T + 1 | \mathbf{X}_i, \bar{\mathbf{Y}}_{T,i})} (Y_{T,i} - \mu_{\text{AIPW}}) + \sum_{t=1}^T \mathbb{1}(D_i \geq t) [\mathbb{1}(D_i = t) - \mathbb{P}(D_i = t | D_i \geq t, \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i})] h_t(\mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \mu_{\text{AIPW}}) \right\} = 0 \quad (9.2.5)$$

and the optimal choice of the functions $h_t(\cdot)$ is given by

$$h_t(\mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \mu_{\text{AIPW}}) = \frac{\mathbb{E}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}) - \mu_{\text{AIPW}}}{\mathbb{P}(R_{t,i} = 1 | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i})}$$

This is not shown here but can be found both in Tsiatis (2006) and in Robins (1999). In our notation, (9.2.5) can be rewritten as

$$\sum_{i=1}^n \left[\frac{R_{T,i}}{\hat{\pi}_{T,i}} (Y_{T,i} - \mu_{\text{AIPW}}) + \sum_{t=1}^T R_{t-1,i} \left(\frac{\hat{\pi}_{t,i}}{\hat{\pi}_{t-1,i}} - R_{t,i} \right) \frac{\mathbb{E}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}) - \mu_{\text{AIPW}}}{\hat{\pi}_{t,i}} \right] = 0 \quad (9.2.6)$$

which is equivalent to

$$\sum_{i=1}^n \left\{ \mathbb{E}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{1,i}) - \mu_{\text{AIPW}} + \sum_{t=1}^T \frac{R_{t,i}}{\hat{\pi}_{t,i}} [\mathbb{E}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}) - \mathbb{E}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i})] \right\} = 0 \quad (9.2.7)$$

Our estimator (9.2.4) can be rewritten as

$$\begin{aligned}
& \sum_{j=1}^m \sum_{i=1}^n \left\{ R_{T,i} (Y_{T,i} - \mu_{\text{DRMI}}) + R_{T-1,i} (1 - R_{T,i}) \left[\hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\pi}_{t-1,i}) - \mu_{\text{DRMI}} \right] \right. \\
& + \cdots + R_{1,i} (1 - R_{2,i}) \left[\hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{1,i}) - \mu_{\text{DRMI}} \right] + (1 - R_{1,i}) \left[\hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i) - \mu_{\text{DRMI}} \right] \\
& \quad + R_{T-1,i} (1 - R_{T,i}) \left[\tilde{Y}_{T,i}^{(j)} - \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\pi}_{t-1,i}) \right] + \cdots \\
& \quad \left. + R_{1,i} (1 - R_{2,i}) \left[\tilde{Y}_{T,i}^{(j)} - \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{1,i}) \right] (1 - R_{1,i}) \left[\tilde{Y}_{T,i}^{(j)} - \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i) \right] \right\} = 0
\end{aligned}$$

and this is equivalent to

$$\begin{aligned}
& \sum_{i=1}^n \sum_{j=1}^m \left\{ \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i) - \mu_{\text{DRMI}} \right. \\
& \quad + \sum_{t=1}^T R_{t,i} \left[\hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i}) - \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\pi}_{t-1,i}) \right] \\
& \quad \left. + \sum_{t=1}^T R_{t-1,i} (1 - R_{t,i}) \left[\tilde{Y}_{T,i}^{(j)} - \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\pi}_{t-1,i}) \right] \right\} = 0
\end{aligned}$$

To show that $\hat{\mu}_{\text{DRMI}}$ is a doubly-robust estimator of μ , we must show that

$$\begin{aligned}
& \mathbb{E} \left(\sum_{j=1}^m \left\{ \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i) - \mu_{\text{DRMI}} \right. \right. \\
& \quad + \sum_{t=1}^T R_{t,i} \left[\hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i}) - \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\pi}_{t-1,i}) \right] \\
& \quad \left. \left. + \sum_{t=1}^T R_{t-1,i} (1 - R_{t,i}) \left[\tilde{Y}_{T,i}^{(j)} - \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\pi}_{t-1,i}) \right] \right\} \right) = 0
\end{aligned}$$

when at least one of the y - and π -models is correctly specified, where the outer expectation is with respect to the true distribution of $\mathbf{X}_i, \bar{\mathbf{Y}}_{T,i}$.

The final term is zero (by the definition of $\hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i})$ as the mean of the distribution from which $\tilde{Y}_{T,i}^{(j)}$ is drawn) and thus our requirement becomes that

$$\mathbb{E} \left(\sum_{j=1}^m \left\{ \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i) - \mu_{\text{DRMI}} + \sum_{t=1}^T R_{t,i} \left[\hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i}) - \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\pi}_{t-1,i}) \right] \right\} \right) = 0$$

when at least one of the y - and π -models is correctly specified, or, equivalently:

$$\mathbb{E} \left\{ \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i) - \mu_{\text{DRMI}} + \sum_{t=1}^T R_{t,i} \left[\hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i}) - \hat{\mathbb{E}}(Y_{T,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\pi}_{t-1,i}) \right] \right\} = 0$$

By Lemma 9.2, this can be rewritten as

$$\mathbb{E} \left\{ H_0(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}) - \mu_{\text{DRMI}} + \sum_{t=1}^T R_{t,i} \left[H_t(\mathbf{X}_i^T, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}) - H_{t-1}(\mathbf{X}_i^T, \bar{\mathbf{Y}}_{t-1,i}, \hat{\pi}_{t-1,i}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}) \right] \right\} = 0 \quad (9.2.8)$$

which is the same as

$$\mathbb{E} \left\{ H_0(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}) - \mu_{\text{DRMI}} + \sum_{t=1}^T \frac{R_{t,i}}{\hat{\pi}_{t,i}} \left[H_t(\mathbf{X}_i^T, \bar{\mathbf{Y}}_{t,i}, \hat{\pi}_{t,i}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}) - H_{t-1}(\mathbf{X}_i^T, \bar{\mathbf{Y}}_{t-1,i}, \hat{\pi}_{t-1,i}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}) \right] \right\} = 0 \quad (9.2.9)$$

since both the second term in (9.2.8) and (9.2.9) are numerically zero (assuming that a constant term was included in the extended y -model).

Then we are done, since the expression inside the expectation in (9.2.9) is the same as the summand in (9.2.7). In other words, that the equality (9.2.9) holds whenever at least one of the y - and π -models is correctly specified follows from the double robustness of $\hat{\mu}_{\text{AIPW}}$. \square

9.2.2.2 Non-monotone longitudinal data

For non-monotone missingness patterns, we recommend first testing the hypothesis that the missing data mechanism belongs to the randomised monotone missingness (RMM) sub-class described in §4.2.1 using the test described by Robins and Gill (1997). If the data do not support this hypothesis, then MAR should be rejected as implausible; even in this case, however, an analysis which assumes ignorability might be required as a point of departure for subsequent sensitivity analyses.

Under the assumption that the data are RMM, the parameters shown in Fig. 4.2 (or the appropriate extension thereof to more timepoints) can be easily estimated. In this example (where there are three outcome variables, but the argument easily extends to any number of outcome variables) we start by defining a ‘stage 2’ variable, $S_{2,i}$ taking the value $s_{2,i}$ where

$$s_{2,i} = \inf \{1, 2, 3 : Y_{s_{2,i}} \text{ is observed}\}$$

or the value 0 if none of $\{Y_{1,i}, Y_{2,i}, Y_{3,i}\}$ is observed. A multinomial logit model is fitted to $S_{2,i}$, conditional on the covariates, and the probabilities $p_1(\mathbf{X}_i)$, $p_2(\mathbf{X}_i)$, and $p_3(\mathbf{X}_i)$ (as shown in Fig. 4.2) are estimated. Then, a ‘stage 3’ variable, $S_{3,i}$, is defined to take the value $s_{3,i} - 1$ where

$$s_{3,i} = \inf \{2, 3 : Y_{s_{3,i}} \text{ is observed and } Y_{k,i} \text{ is observed, where } k < s_{3,i}\}$$

or the value 0 if only one of $\{Y_{1,i}, Y_{2,i}, Y_{3,i}\}$ is observed. For each level $s_{2,i}$ of $S_{2,i}$, a multinomial logit model is fitted to $S_{3,i}$ conditional on $Y_{s_{2,i},i}$ and the covariates. The probabilities $p_2(\mathbf{X}_i, Y_{1,i})$, $p_3(\mathbf{X}_i, Y_{1,i})$, and $p_3(\mathbf{X}_i, Y_{2,i})$ (as shown in Fig. 4.2) are estimated. The models are fitted using only the subjects for whom $S_{2,i} = s_{2,i}$.

Finally, a ‘stage 4’ variable, $S_{4,i}$, taking the value $s_{4,i}$ where

$$s_{4,i} = \begin{cases} 1 & \text{if } Y_{1,i}, Y_{2,i}, Y_{3,i} \text{ are all observed} \\ 0 & \text{otherwise} \end{cases}$$

is defined and, for each pair $\{s_{2,i}, s_{3,i}\}$, a logistic regression is fitted to $S_{4,i}$ conditional on $Y_{s_{2,i},i}, Y_{s_{3,i},i}$ and the covariates. The probabilities $p_3(\mathbf{X}_i, Y_{1,i}, Y_{2,i})$ are estimated. These models are fitted using only the subjects for whom $\{S_{2,i}, S_{3,i}\} = \{s_{2,i}, s_{3,i}\}$.

From these estimated probabilities, we would like to estimate each of

$$\mathbb{P}(R_{1,i} = 1 | \mathbf{X}_i) = p_1(\mathbf{X}_i) \quad (9.2.10)$$

$$\mathbb{P}(R_{2,i} = 1 | \mathbf{X}_i, Y_{1,i}) = p_1(\mathbf{X}_i) p_2(\mathbf{X}_i, Y_{1,i}) + p_2(\mathbf{X}_i) \quad (9.2.11)$$

$$\begin{aligned} \mathbb{P}(R_{3,i} = 1 | \mathbf{X}_i, Y_{1,i}, Y_{2,i}) &= p_1(\mathbf{X}_i) p_2(\mathbf{X}_i, Y_{1,i}) p_3(\mathbf{X}_i, Y_{1,i}, Y_{2,i}) + p_1(\mathbf{X}_i) p_3(\mathbf{X}_i, Y_{1,i}) \\ &\quad + p_2(\mathbf{X}_i) p_3(\mathbf{X}_i, Y_{2,i}) + p_3(\mathbf{X}_i) \end{aligned} \quad (9.2.12)$$

Note that even in this non-monotone setting, since the data are longitudinal, it remains the case that $\hat{\pi}_{t,i} = \mathbb{P}(R_{t,i} = 1 | \mathbf{Z}_i) = \mathbb{P}(R_{t,i} = 1 | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i})$, i.e. that the missingness probabilities at each timepoint depend only on past measurements of Y .

There is no problem with (9.2.10) but (9.2.11) and (9.2.12) are undefined for some subjects. For example, if subject i has only Y_2 observed then $p_2(\mathbf{X}_i, Y_{1,i})$ cannot be calculated. Upto a function of the unknown $Y_{1,i}$, it can, however, be specified and in such cases (9.2.11) and (9.2.12) are specified as known functions of the unknown $Y_{1,i}$ or $Y_{2,i}$. This completes the description of the π -model.

We proceed by fitting the model using MI, and to cope with the non-monotone pattern, MI using chained equations (MICE) as described in §6.7.4 is used. As with the monotone case, for each $t \in \{1, \dots, T\}$, $\hat{\pi}_{t,i}^{-1}$ is included as an additional covariate (additional to the specified y -model) when imputing $Y_{t,i}$. As we noted above, $\hat{\pi}_{t,i}^{-1}$ itself, in general, is missing for some subjects, and is therefore imputed (deterministically) as

a function of the (possibly imputed) $\tilde{Y}_{1,i}, \dots, \tilde{Y}_{t-1,i}$.

Although when generating such data, we would only need to consider the distribution of each outcome variable Y_t conditional on the covariates and the previous $t - 1$ outcome variables (since the future cannot determine the past), for the analysis model (the y -model), it will be necessary—in this non-monotone case—to postulate the implied models for Y_t given all future outcome variables as well, and the future outcome variables must be included in the imputation models, e.g. Y_2 must be included in the imputation model for Y_1 . Thus, the extended y -model in the non-monotone case differs from that of the monotone case, since the imputation model for Y_t conditions on all past and future values of Y , as well as \mathbf{X} and $\hat{\pi}_t^{-1}$.

Finally, $\hat{\mu}_{\text{DRMI}}$ is again calculated as the solution to

$$\sum_{j=1}^m \sum_{i=1}^n \left[\tilde{Y}_{T,i}^{(j)} - \mu_{\text{DRMI}} \right] = 0 \quad (9.2.13)$$

and a variance estimate (subject to the same caveats as above) obtained using Rubin's variance formula.

Conjecture 9.4 (Multiply imputed DR non-monotone longitudinal estimator). *The estimator $\hat{\mu}_{\text{DRMI}}$ is doubly robust. That is, if at least one of the two models (the π -model and the y -model) is correctly specified (but not necessarily both), $\hat{\mu}_{\text{DRMI}}$ is a consistent estimator of μ .*

Sketch proof. The general form of the AIPW estimating equation (Tsiatis, 2006, p.173)

for non-monotone data can be written as

$$\sum_{i=1}^n \left\{ \frac{R_{1,i} \cdots R_{T,i}}{\mathbb{P}(R_{1,i} \cdots R_{T,i} = 1 | \mathbf{X}_i, \bar{\mathbf{Y}}_{T,i})} (Y_{T,i} - \mu_{\text{AIPW}}) + \sum_{r_1 \cdots r_T \neq 1} \left[\mathbb{1}(R_{1,i} = r_1, \dots, R_{T,i} = r_T) - \frac{R_{1,i} \cdots R_{T,i} \mathbb{P}(R_{1,i} = r_1, \dots, R_{T,i} = r_T)}{\mathbb{P}(R_{1,i} \cdots R_{T,i} = 1 | \mathbf{X}_i, \bar{\mathbf{Y}}_{T,i})} \right] \cdot g[r_1, \dots, r_T, G_{r_1, \dots, r_T}(\mathbf{X}_i, \bar{\mathbf{Y}}_{T,i}), \mu_{\text{AIPW}}] \right\} = 0 \quad (9.2.14)$$

where the functions $g(\cdot)$ could be any functions of the observed data.

As we have already noted, although in the general formula given in (9.2.14) the response probabilities are conditional on all outcomes, since we are restricting our consideration to non-monotone longitudinal data under a RMM mechanism, it remains in our case that

$$\mathbb{P}(R_{t,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{T,i}) = \mathbb{P}(R_{t,i} | \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i})$$

For ease of writing, let us drop the subscript i and consider a simple example with only three timepoints and no \mathbf{X} . Suppose that Y_1 is always observed, but that Y_2 and Y_3 are both subject to missingness, in a non-monotone pattern.

Also for ease of writing, let W_2 be the inverse of the probability that Y_2 is observed, conditional on Y_1 ; let W_3 be the inverse of the probability that Y_3 is observed, conditional on Y_1 and Y_2 ; and let W_{23} be the inverse of the probability that both Y_2 and Y_3 are observed, conditional on Y_1 and Y_2 .

Consider the subjects with intermittent missingness, i.e. the subjects who have Y_3 observed but Y_2 missing. Consider the hypothetical dataset which consists of all the observed data together with the true unobserved values of Y_2 for these subjects with intermittent missingness. The pattern of missingness in this hypothetical dataset is clearly monotone. Let H_t be the hypothetical predictions from the Bang

and Robins procedure for longitudinal monotone data (as described in §7.6.4.2) after $T - t$ iterations of step 3(a), applied to this hypothetical monotone dataset. That is $H_3(Y_3) = Y_3$, $H_2(Y_1, Y_2, W_3) = \hat{\mathbb{E}}(H_3 | Y_1, Y_2, W_3)$ and $H_1(Y_1, W_2) = \hat{\mathbb{E}}(H_2 | Y_1, W_2)$. Let $\tilde{H}_2^{(j)} = H_2(Y_1, \tilde{Y}_2^{(j)}, W_3(Y_1, \tilde{Y}_2^{(j)}))$. Finally, let E_2 be the mean (over the imputation distribution) of $\tilde{H}_2^{(j)}$.

We can rewrite (9.2.14) as

$$\begin{aligned} \sum \left\{ R_2 R_3 W_{23} (Y_3 - \mu) + \left(R_2 - \frac{R_2 R_3 W_{23}}{W_2} \right) [W_2 (H_2 - H_1) + H_1 - \mu] \right. \\ \left. + \left(R_3 - \frac{R_2 R_3 W_{23}}{W_3} \right) [W_3 (Y_3 - E_2) + H_1 - \mu] \right. \\ \left. + \left(1 - R_2 - R_3 - R_2 R_3 W_{23} + \frac{R_2 R_3 W_{23}}{W_2} + \frac{R_2 R_3 W_{23}}{W_3} \right) (H_1 - \mu) \right\} = 0 \quad (9.2.15) \end{aligned}$$

since $W_2 (H_2 - H_1) + H_1$, $W_3 (Y_3 - E_2) + H_1$ and H_1 are functions only of the observed data for subjects with only (Y_1, Y_2) , (Y_1, Y_3) and Y_1 observed, respectively.

Using similar arguments to those already used, for example that

$$\sum R_2 W_2 (H_2 - H_1)$$

is numerically zero, we can show that the expectation of the summand in (9.2.13) is equal to the expectation of the summand in (9.2.15). The DR property of the solution to the latter therefore implies the DR property of $\hat{\mu}_{\text{DRMI}}$, which completes the (sketch) proof. \square

9.2.3 Non-monotone cross-sectional ignorable missing data

The arguments above could be extended to the case where the data are not constrained to be longitudinal, but this would require a method for estimating the weights when the

order in which the variables were observed is not known. Although Robins and Gill (1997) propose a method for calculating the complete-case weights in a randomised monotone missingness setting using an EM algorithm with the path followed by a particular subject through Fig. 4.1 treated as a missing value, they also prove that the same method *cannot* be used to identify the individual path probabilities, suggesting that the timepoint-specific missingness probabilities cannot be determined either.

9.2.4 A closer look at Bang and Robins for longitudinal data

In section 3 of their paper, when describing the algorithm for constructing the DR estimator for longitudinal data, Bang and Robins write

For subjects with $C \geq m$, specify and fit by IRLS a parametric regression model $e_{m-1}(\bar{\mathbf{L}}_{m-1}; \boldsymbol{\beta}_{m-1}, \phi_{m-1}) = \Psi[s_{m-1}(\bar{\mathbf{L}}_{m-1}; \boldsymbol{\beta}_{m-1}) + \phi \hat{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})]$ for the conditional expectation $\mathbb{E}[\hat{H}_m(\mu) | C \geq m, \bar{\mathbf{L}}_{m-1}]$.

Allowing for slight differences in notation, this corresponds to step 3(a) in our description in §7.6.4.2. The focus for us is the function $s_{m-1}(\cdot)$ which specifies the functional form of the linear predictor for the chosen regression. In section 3.1 they describe a simulation study and write

Let $\mathbf{L} = (\mathbf{L}_1^T, L_2, L_3)^T$ represent the full data with $\mathbf{L}_1 = (V_{11}, V_{12}, V_{13})^T$ and $L_3 = Y$. So the censoring variable C takes a value in $\{1, 2, 3\}$. V_{1i} ($i = 1, 2, 3$) were generated independently from a standard normal, L_2 from $N[s_1(\mathbf{L}_1; \boldsymbol{\beta}), 1]$, and Y from $N[s_2(\bar{\mathbf{L}}_2; \boldsymbol{\beta}), 1]$ as presented in Table 1C.

Referring to Table 1C, we see that the functions $s_1(\cdot)$ and $s_2(\cdot)$ are defined as follows:

$$s_1(\mathbf{L}_1; \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot (1, V_{11}, V_{11}V_{13})^T, \quad \boldsymbol{\beta} = (0, 3, -2)$$

$$s_2(\bar{\mathbf{L}}_2; \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot (1, V_{11}^2, V_{12}, V_2^2, V_{12}V_2)^T, \boldsymbol{\beta} = (0, -3, 3, 1, -2)$$

where $V_2 = L_2$.

Although not stated in so many words, the implication here is that the functions $s(\cdot)$ used to generate the data are the exact same functions $s(\cdot)$ used in step 3(a) as the linear predictor for the chosen GLMs. Further thought, however, reveals that (except for the regression of H_T , i.e. the regression which uses $s_2(\cdot)$ in this example) the two sets of functions do not, in general, coincide and that the example chosen by the authors for their simulation study is an example in which the $s_1(\cdot)$ needed for the analysis is quite different from the $\tilde{s}_1(\cdot)$ (relabelled $\tilde{s}_1(\cdot)$ to differentiate it from $s_1(\cdot)$) used to generate the data. In this example, where every variable is normally distributed, it is relatively straightforward to derive the function needed for the analysis as we now show. It should come as no surprise that $s_1(\cdot)$ is not equal to $\tilde{s}_1(\cdot)$ in general, since the former is $\mathbf{E}(Y | \mathbf{L}_1)$ and the latter is $\mathbf{E}(L_2 | \mathbf{L}_1)$.

The conditional distribution of $L_2 | \mathbf{L}_1$ is

$$N(3V_{11} - 2V_{11}V_{13}, 1)$$

and the conditional distribution of $Y | \bar{\mathbf{L}}_2$, is

$$N(-3V_{11}^2 + 3V_{12} + L_2^2 - 2V_{12}L_2, 1)$$

Thus the conditional expectation of $Y | \mathbf{L}_1$, is

$$\begin{aligned} & -3V_{11}^2 + 3V_{12} + 1 + (3V_{11} - 2V_{11}V_{13})^2 - 2V_{12}(3V_{11} - 2V_{11}V_{13}) \\ & = 1 + 3V_{12} + 6V_{11}^2 - 6V_{11}V_{12} - 12V_{11}^2V_{13} + 4V_{11}V_{12}V_{13} + 4V_{11}^2V_{13}^2 \end{aligned}$$

Thus, when carrying out the simulation study under the ‘both models correct’ scenario, the authors must have used $1, V_{12}, V_{11}^2, V_{11}V_{12}, V_{11}^2V_{13}, V_{11}V_{12}V_{13}, V_{11}^2V_{13}^2$ as the covariates for the second linear regression stage, as opposed to $1, V_{11}, V_{11}V_{13}$ as the paper strongly

suggests. In the Gaussian case, this additional step is straightforward, but when the data are non-Gaussian, a suitable function $s_1(\cdot)$ could be difficult (or even impossible) to derive even if functions $s_2(\cdot)$ and $\tilde{s}_1(\cdot)$ could be easily postulated.

We note that the corresponding issue does not apply in DRMI, since our imputation model is formulated for each variable individually conditional on the other variables. The expectation of Y given \mathbf{L}_1 for a subject with L_2 and Y missing is calculated sequentially by first estimating the distribution of L_2 given \mathbf{L}_1 and then the distribution of Y given \mathbf{L}_1 and the imputed value (conditional on \mathbf{L}_1) of L_2 . We have already noted this feature of DRMI, since it gives rise to the need to impute the marginal missingness probability at time t for subjects who dropped out before time $t - 1$. Thus, we now see that the feature which, earlier in our description of the method, seemed to be a disadvantage when compared with the method of Bang and Robins (2005), the same feature also offers an advantage here, namely that it is not necessary to postulate models for Y given \mathbf{L}_1 marginalised over L_2 .

9.3 Simulation studies

9.3.1 Univariate ignorable missing data

First we repeat the first simulation study carried out by Bang and Robins (2005), adding our DRMI estimator as a fourth estimator to be compared with the IPWCC estimator, the outcome regression (OR) estimator and the Bang and Robins doubly robust (DR) estimator.

The OR estimator is the solution to

$$\sum_{i=1}^n \left[e(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}) - \mu_{\text{OR}} \right] = 0$$

where $e(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}})$ are the predictions from the (non-extended) y -model (9.2.1). This is equivalent to a maximum likelihood analysis.

In this simulation study, $\mathbf{X} = (X_1, X_2, X_3)$ is always fully-observed and generated from a trivariate normal distribution with mean $(0, 0, 0)$ and variance-covariance matrix equal to the (3×3) identity matrix. Y is normally distributed with mean $s_{\text{true}}(\mathbf{X}, \boldsymbol{\beta})$ and unit variance, where $s_{\text{true}}(\mathbf{X}, \boldsymbol{\beta}) = \boldsymbol{\beta}(1, X_1^2, X_2, X_2 X_3)^T$ and $\boldsymbol{\beta} = (0, 1, 2.5, 3)$.

R is generated from the following logistic regression:

$$\text{logit}[\hat{\pi}_{\text{true}}(\mathbf{X}, \boldsymbol{\alpha})] = \boldsymbol{\alpha}(1, I_1, I_2, I_3, I_1 I_2)^T$$

where $\boldsymbol{\alpha} = (-1, 1, 0, 0, -1)$ and I_l stands for $\mathbb{1}(X_l > 0)$.

To investigate the double robustness property, an incorrect π -model and an incorrect y -model are specified as follows:

$$s_{\text{false}}(\mathbf{X}, \boldsymbol{\beta}) = \boldsymbol{\beta}(1, X_1, X_2^2)^T$$

$$\text{logit}[\hat{\pi}_{\text{false}}(\mathbf{X}, \boldsymbol{\alpha})] = \boldsymbol{\alpha}(1, I_1, I_3)^T$$

The simulation study is based on a sample size of 500 and 1,000 simulations, with the doubly robust MI procedure based on 10 imputations. The results are shown in Table 9.1.

9.3.2 Longitudinal monotone ignorable missing data

Next, we repeat the longitudinal monotone simulation study carried out by Bang and Robins (2005), again adding our DRMI estimator as a fourth estimator to be compared with the IPWCC estimator, the OR estimator and the DR estimator.

Estimator	Bias	True variance	Estimated variance	Coverage probability
$\hat{\mu}_{IPWCC}$	-0.01	0.11	-	-
$\hat{\mu}_{OR}$	-0.00	0.04	-	-
$\hat{\mu}_{DR}$	-0.00	0.04	-	-
$\hat{\mu}_{DRMI}$	-0.00	0.04	0.04	0.95
$\hat{\mu}_{IPWCC \cdot \pi - \text{false}}$	-0.36	0.13	-	-
$\hat{\mu}_{DR \cdot \pi - \text{false}}$	-0.00	0.04	-	-
$\hat{\mu}_{DRMI \cdot \pi - \text{false}}$	-0.01	0.04	0.04	0.95
$\hat{\mu}_{OR \cdot y - \text{false}}$	-0.35	0.12	-	-
$\hat{\mu}_{DR \cdot y - \text{false}}$	-0.01	0.11	-	-
$\hat{\mu}_{DRMI \cdot y - \text{false}}$	-0.02	0.12	0.12	0.93
$\hat{\mu}_{DR \cdot \pi \oplus y - \text{false}}$	-0.35	0.13	-	-
$\hat{\mu}_{DRMI \cdot \pi \oplus y - \text{false}}$	-0.35	0.14	0.12	0.79

Table 9.1: The results of the first simulation study performed by Bang and Robins (2005) with doubly robust multiple imputation (DRMI) included in the comparison. No subscript indicates correct specification of the relevant model(s). $\pi - \text{false}$ indicates that the estimator used an incorrectly-specified π -model, $y - \text{false}$ indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y - \text{false}$ indicates that both the π - and y -models were incorrectly specified.

The OR estimator is now the solution to

$$\sum_{i=1}^n \left[H_0 \left(\mathbf{X}_i^T, \hat{\beta} \right) - \mu_{OR} \right] = 0$$

where $H_0 \left(\mathbf{X}_i^T, \hat{\beta} \right)$ is as defined in §9.2.2.

As before, $\mathbf{X} = (X_1, X_2, X_3)$ is always fully-observed with X_1, X_2, X_3 independent and identically distributed standard normal variables. Y_1 is normally distributed with mean $\tilde{s}_1^{\text{true}}(\mathbf{X}, \beta_1)$ and unit variance, where $\tilde{s}_1^{\text{true}}(\mathbf{X}, \beta_1) = \beta_1 (1, X_1, X_1 X_3)^T$ and $\beta_1 = (0, 3, 2)$. Y_2 is normally distributed with mean $s_2^{\text{true}}(\mathbf{X}, Y_1, \beta_2)$ and unit variance, where $s_2^{\text{true}}(\mathbf{X}, Y_1, \beta_2) = \beta_2 (1, X_1^2, X_2, Y_1^2, X_2 Y_1)^T$ and $\beta_2 = (0, -3, 3, 1, -2)$. The implied

$s_1^{\text{true}}(\mathbf{X}, \boldsymbol{\beta}_1)$ (as we showed in §9.2.4) is

$$s_1^{\text{true}}(\mathbf{X}, \boldsymbol{\beta}_1) = \boldsymbol{\beta}_1 (1, X_2, X_1^2, X_1 X_2, X_1^2 X_3, X_1 X_2 X_3, X_1^2 X_3^2)$$

R_1 is generated from the following logistic regression:

$$\text{logit} [\hat{\pi}_1^{\text{true}}(\mathbf{X}, \boldsymbol{\alpha}_1)] = \boldsymbol{\alpha}_1 (1, I_1^X, I_2^X, I_3^X, I_1^X I_2^X)^T$$

where $\boldsymbol{\alpha}_1 = (1, -1, -1, 1, 1)$ and I_l^X stands for $\mathbf{1}(X_l > 0)$. Conditional on $R_1 = 1$, R_2 is generated from the following logistic regression:

$$\text{logit} [\hat{\pi}_2^{\text{true}}(\mathbf{X}, Y_1, \boldsymbol{\alpha}_2)] = \boldsymbol{\alpha}_2 (1, I_1^X, I_2^X, I_3^X, I_1^X I_2^X, I_1^Y, I_3^X I_1^Y)^T$$

where $\boldsymbol{\alpha}_2 = (0, -1, -1, 0, 1, 0, 2)$ and I_1^Y stands for $\mathbf{1}(Y_1 > 0)$. If $R_1 = 0$ then $R_2 = 0$.

To investigate the double robustness property, an incorrect π -model and an incorrect y -model are specified as follows:

$$s_1^{\text{false}}(\mathbf{X}, \boldsymbol{\beta}) = \boldsymbol{\beta} (1, X_1, X_2)^T$$

$$s_2^{\text{false}}(\mathbf{X}, Y_1, \boldsymbol{\beta}) = \boldsymbol{\beta} (1, X_1, X_2^2, X_3^2, Y_1)^T$$

$$\text{logit} [\hat{\pi}_1^{\text{false}}(\mathbf{X}, \boldsymbol{\alpha})] = \boldsymbol{\alpha} (1, I_2^X, I_3^X)^T$$

$$\text{logit} [\hat{\pi}_2^{\text{false}}(\mathbf{X}, \boldsymbol{\alpha})] = \boldsymbol{\alpha} (1, I_1^Y)^T$$

The simulation study is based on a sample size of 500 and 1,000 simulations, with the DRMI procedure based on 10 imputations. The results are shown in Table 9.2.

Estimator	Bias	True variance	Estimated variance	Coverage probability
$\hat{\mu}_{IPWCC}$	-0.11	10.98	—	—
$\hat{\mu}_{OR}$	0.06	1.92	—	—
$\hat{\mu}_{DR}$	0.06	1.92	—	—
$\hat{\mu}_{DRMI}$	0.07	1.91	1.83	0.94
$\hat{\mu}_{IPWCC \cdot \pi\text{-false}}$	-3.21	5.87	—	—
$\hat{\mu}_{DR \cdot \pi\text{-false}}$	0.06	1.92	—	—
$\hat{\mu}_{DRMI \cdot \pi\text{-false}}$	0.08	1.92	1.83	0.93
$\hat{\mu}_{OR \cdot y\text{-false}}$	-4.99	3.51	—	—
$\hat{\mu}_{DR \cdot y\text{-false}}$	-0.36	10.51	—	—
$\hat{\mu}_{DRMI \cdot y\text{-false}}$	-0.37	10.63	4.28	0.74
$\hat{\mu}_{DR \cdot \pi \oplus y\text{-false}}$	-2.35	8.13	—	—
$\hat{\mu}_{DRMI \cdot \pi \oplus y\text{-false}}$	-2.37	7.38	3.67	0.57

Table 9.2: The results of the monotone longitudinal simulation study performed by Bang and Robins (2005) with doubly robust multiple imputation (DRMI) included in the comparison. No subscript indicates correct specification of the relevant model(s). π - false indicates that the estimator used an incorrectly-specified π -model, y - false indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y$ - false indicates that both the π - and y -models were incorrectly specified.

9.3.3 Longitudinal non-monotone ignorable missing data

Next, we consider a longitudinal non-monotone simulation study. In this case, neither the OR nor the DR estimator can be used and thus we compare our DRMI estimator with the IPWCC estimator and an ordinary multiple imputation (MI) estimator, i.e. an estimator identical to the DRMI estimator but without the inverse probability weights as additional covariates.

In this simulation study, X (univariate) is always observed and generated from a standard normal distribution. Y_1 is normally distributed with mean $\tilde{s}_1^{\text{true}}(X, \beta_1)$ and unit variance, where $\tilde{s}_1^{\text{true}}(X, \beta_1) = \beta_1(1, X^2)^T$ and $\beta_1 = (0, 1)$. Y_2 is normally distributed with mean $s_2^{\text{true}}(X, Y_1, \beta_2)$ and unit variance, where $s_2^{\text{true}}(X, Y_1, \beta_2) = \beta_2(1, X, Y_1)^T$

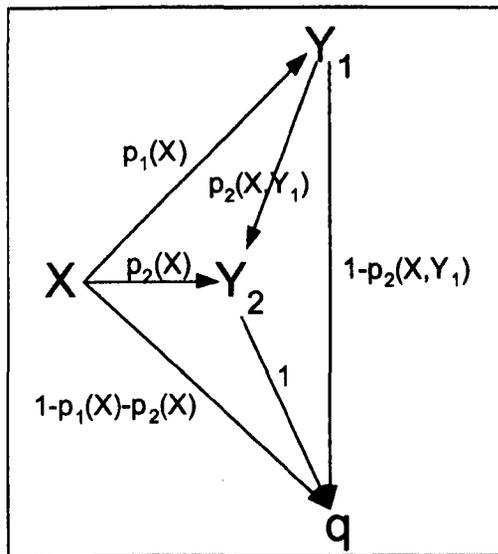


Figure 9.1: The MRMM longitudinal process used for the longitudinal non-monotone simulation study.

and $\beta_2 = (0, -1, 2)$. The implied $s_1^{\text{true}}(X, Y_2, \beta_1)$ is

$$s_1^{\text{true}}(X, Y_2, \beta_1) = \beta_1 (1, X, X^2, Y_2)$$

Note that $s_1(\cdot)$ is now a function of Y_2 . This is essential, since some subjects have Y_2 but not Y_1 observed. If Y_2 is omitted from the imputation model for Y_1 , the resulting estimator is, in general, biased since the stationary distribution to which the Gibbs sampler in the MICE procedure converges is not the correct full-data distribution, even under MAR.

The missingness model is illustrated in Fig. 9.1.

$p_1(X)$ and $p_2(X)$ are defined by the following multinomial logit model:

$$p_1^{\text{true}}(X, \alpha_{11}, \alpha_{12}) = \frac{\exp \left[\alpha_{11} \left(1, \sqrt{|X|} \right)^T \right]}{1 + \exp \left[\alpha_{11} \left(1, \sqrt{|X|} \right)^T \right] + \exp \left[\alpha_{12} \left(1, \sqrt{|X|} \right)^T \right]}$$

$$p_2^{\text{true}}(X, \alpha_{11}, \alpha_{12}) = \frac{\exp \left[\alpha_{12} \left(1, \sqrt{|X|} \right)^T \right]}{1 + \exp \left[\alpha_{11} \left(1, \sqrt{|X|} \right)^T \right] + \exp \left[\alpha_{12} \left(1, \sqrt{|X|} \right)^T \right]}$$

where $\alpha_{11} = (2, -1)$ and $\alpha_{12} = (0, 0.5)$. Conditional on Y_1 being observed at the first stage, $p_2(X, Y_1)$ is generated from the following logistic regression:

$$\text{logit} \left[p_2^{\text{true}}(X, Y_1, \alpha_2) \right] = \alpha_2 \left(1, X, Y_1^2 \right)^T$$

where $\alpha_2 = (0, -2, 0.5)$.

$\hat{\pi}_1^{\text{true}}(X, \alpha_{11}, \alpha_{12})$ and $\hat{\pi}_2^{\text{true}}(X, Y_1 \alpha_{11}, \alpha_{12}, \alpha_2)$ are then calculated as follows:

$$\hat{\pi}_1^{\text{true}}(X, \alpha_{11}, \alpha_{12}) = p_1^{\text{true}}(X, \alpha_{11}, \alpha_{12})$$

$$\hat{\pi}_2^{\text{true}}(X, Y_1 \alpha_{11}, \alpha_{12}, \alpha_2) = p_2^{\text{true}}(X, \alpha_{11}, \alpha_{12}) + p_1^{\text{true}}(X, \alpha_{11}, \alpha_{12}) p_2^{\text{true}}(X, Y_1, \alpha_2)$$

Thus, $\hat{\pi}_2^{\text{true}}(X, Y_1 \alpha_{11}, \alpha_{12}, \alpha_2)$ is missing for all subjects for whom Y_1 is missing and should be imputed (after Y_1 and before Y_2 in each chained equations cycle) deterministically based on the current imputed value of Y_1 . The code is given in the Appendix in §C.2.

To investigate the double robustness property, an incorrect π -model and an incorrect y -model are specified as follows:

$$s_1^{\text{false}}(X, Y_2, \beta_1) = \beta_1 \left(1, X, Y_2 \right)^T$$

Estimator	Bias	True variance	Estimated variance	Coverage probability
$\hat{\mu}_{IPWCC}$	0.00	0.07	—	—
$\hat{\mu}_{MI}$	-0.01	0.03	—	—
$\hat{\mu}_{DRMI}$	-0.02	0.03	0.03	0.95
$\hat{\mu}_{IPWCC \cdot \pi - \text{false}}$	-0.59	0.05	—	—
$\hat{\mu}_{DRMI \cdot \pi - \text{false}}$	-0.03	0.03	0.03	0.94
$\hat{\mu}_{MI \cdot y - \text{false}}$	3.07×10^{31}	2.16×10^{65}	—	—
$\hat{\mu}_{DRMI \cdot y - \text{false}}$	0.00	0.04	0.06	0.97
$\hat{\mu}_{DRMI \cdot \pi \oplus y - \text{false}}$	2.32	123.55	5.27×10^8	0.94

Table 9.3: The results of the non-monotone longitudinal simulation study where doubly robust multiple imputation (DRMI) is compared with IPWCC and ordinary MI. No subscript indicates correct specification of the relevant model(s). $\pi - \text{false}$ indicates that the estimator used an incorrectly-specified π -model, $y - \text{false}$ indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y - \text{false}$ indicates that both the π - and y -models were incorrectly specified.

$$s_2^{\text{false}}(X, Y_1, \beta_2) = \beta_2 (1, Y_1^2)^T$$

$$p_1^{\text{false}}(X, \alpha_{11}, \alpha_{12}) = \frac{\exp(\alpha_{11})}{1 + \exp(\alpha_{11}) + \exp(\alpha_{12})}$$

$$p_2^{\text{false}}(X, \alpha_{11}, \alpha_{12}) = \frac{\exp(\alpha_{12})}{1 + \exp(\alpha_{11}) + \exp(\alpha_{12})}$$

$$\text{logit}[p_2^{\text{true}}(X, Y_1, \alpha_2)] = \alpha_2 (1, X, Y_1)^T$$

The simulation study is based on a sample size of 500 and 1,000 simulations, with the doubly robust MI procedure based on 10 imputations and 10 cycles of the chained equations procedure (see §6.7.4). The results are shown in Table 9.3.

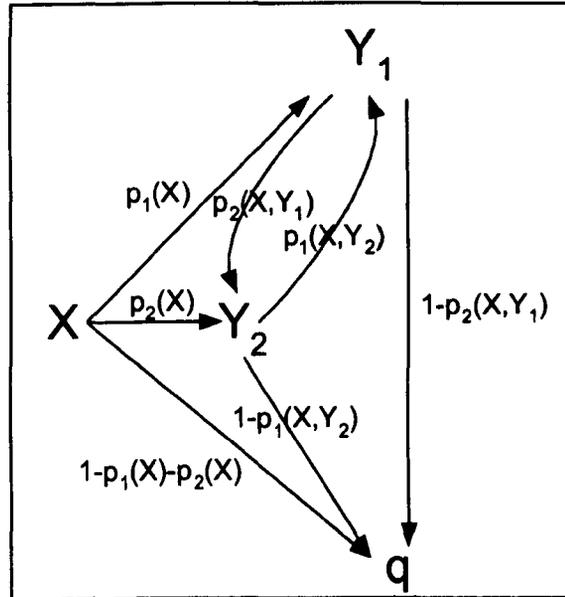


Figure 9.2: The MRMM longitudinal process used for the longitudinal non-monotone simulation study.

9.3.4 Cross-sectional non-monotone ignorable missing data

Finally, we consider a cross-sectional non-monotone simulation study. Again, neither the OR nor the DR estimator can be used and thus we compare our DRMI estimator with the IPWCC estimator and an ordinary MI estimator.

As in the previous simulation study, X (univariate) is always observed and generated from a standard normal distribution. Y_1 is normally distributed with mean $\tilde{s}_1^{\text{true}}(X, \beta_1)$ and unit variance, where $\tilde{s}_1^{\text{true}}(X, \beta_1) = \beta_1 (1, X^2)^T$ and $\beta_1 = (0, 1)$. Y_2 is normally distributed with mean $s_2^{\text{true}}(X, Y_1, \beta_2)$ and unit variance, where $s_2^{\text{true}}(X, Y_1, \beta_2) = \beta_2 (1, X, Y_1)^T$ and $\beta_2 = (0, -1, 2)$. The implied $s_1^{\text{true}}(X, Y_2, \beta_1)$ is

$$s_1^{\text{true}}(X, Y_2, \beta_1) = \beta_1 (1, X, X^2, Y_2)$$

The missingness model is illustrated in Fig. 9.2.

$p_1(X)$ and $p_2(X)$ are defined by the following multinomial logit model:

$$p_1^{\text{true}}(X, \alpha_{11}, \alpha_{12}) = \frac{\exp[\alpha_{11}(1, X, X^2)^T]}{1 + \exp[\alpha_{11}(1, X, X^2)^T] + \exp[\alpha_{12}(1, X, X^2)^T]}$$

$$p_2^{\text{true}}(X, \alpha_{11}, \alpha_{12}) = \frac{\exp[\alpha_{12}(1, X, X^2)^T]}{1 + \exp[\alpha_{11}(1, X, X^2)^T] + \exp[\alpha_{12}(1, X, X^2)^T]}$$

where $\alpha_{11} = (1, -0.5, 0.2)$ and $\alpha_{12} = (0, 0.5, -0.3)$.

Conditional on Y_1 being observed at the first stage, $p_2(X, Y_1)$ is generated from the following logistic regression:

$$\text{logit}[p_2^{\text{true}}(X, Y_1, \alpha_{22})] = \alpha_{22}(1, X, Y_1)^T$$

where $\alpha_{22} = (0, -1, 0.3)$.

Conditional on Y_2 being observed at the first stage, $p_1(X, Y_2)$ is generated from the following logistic regression:

$$\text{logit}[p_1^{\text{true}}(X, Y_2, \alpha_{21})] = \alpha_{21}(1, X, Y_2)^T$$

where $\alpha_{21} = (0, -1, 0.3)$.

$\hat{\pi}_1^{\text{true}}(X, \alpha_{11}, \alpha_{12}, \alpha_{21})$ and $\hat{\pi}_2^{\text{true}}(X, Y_1\alpha_{11}, \alpha_{12}, \alpha_{22})$ are then calculated as follows:

$$\hat{\pi}_1^{\text{true}}(X, \alpha_{11}, \alpha_{12}) = p_1^{\text{true}}(X, \alpha_{11}, \alpha_{12}, \alpha_{21}) + p_2^{\text{true}}(X, \alpha_{11}, \alpha_{12})p_1^{\text{true}}(X, Y_2, \alpha_{21})$$

$$\hat{\pi}_2^{\text{true}}(X, Y_1\alpha_{11}, \alpha_{12}, \alpha_{22}) = p_2^{\text{true}}(X, \alpha_{11}, \alpha_{12}) + p_1^{\text{true}}(X, \alpha_{11}, \alpha_{12})p_2^{\text{true}}(X, Y_1, \alpha_{22})$$

In this case, neither $\hat{\pi}_1^{\text{true}}(X, \alpha_{11}, \alpha_{12}, \alpha_{21})$ nor $\hat{\pi}_2^{\text{true}}(X, Y_1\alpha_{11}, \alpha_{12}, \alpha_{22})$ is fully-observed for all subjects and would need to be imputed deterministically based on the current imputed values of Y_1 and Y_2 . However, because of the difficulty associated

with estimating the marginal weights (discussed in §9.2.3), we cannot obtain reliable estimates of $\hat{\pi}_1^{\text{true}}(X, \alpha_{11}, \alpha_{12}, \alpha_{21})$ and $\hat{\pi}_2^{\text{true}}(X, Y_1 \alpha_{11}, \alpha_{12}, \alpha_{22})$ even for the complete cases. For the purposes of this simulation study, therefore, we will use the true (known) weights.

To investigate the double robustness property, an incorrect y -model is specified as follows:

$$s_1^{\text{false}}(X, Y_2, \beta_1) = \beta_1(1, X^2, Y_2)^T$$

$$s_2^{\text{false}}(X, Y_1, \beta_2) = \beta_2(1, Y_1)^T$$

Since the true weights are being used, no ‘ π -model’ exists. To investigate the double robustness property, we therefore define $\hat{\pi}_1^{\text{false}} = \sqrt{\hat{\pi}_1^{\text{true}}}$ and $\hat{\pi}_2^{\text{false}} = \sqrt{\hat{\pi}_2^{\text{true}}}$.

The simulation study is based on a sample size of 500 and 1,000 simulations, with the MI and doubly robust MI procedures based on 10 imputations and 10 cycles of the chained equations procedure. The results are shown in Table 9.4.

9.4 Discussion

We have seen that in both the univariate cross-sectional and longitudinal monotone cases, where the Bang and Robins (2005) method can be applied, its performance and our estimator’s performance are very similar. In addition, the variance estimates obtained using Rubin’s variance formula perform well when both models are correctly specified. As expected, the variance estimates do not share the double robustness property possessed by the estimates themselves. Our proposed variance estimator does not take into account the variability of the estimated weights but, at least in our simulations, this effect is negligible. It should in principle be possible to incorporate this variability using a sandwich estimator. Further work is needed on this.

Estimator	Bias	True variance	Estimated variance	Coverage probability
$\hat{\mu}_{IPWCC}$	0.01	0.07	—	—
$\hat{\mu}_{MI}$	0.00	0.03	—	—
$\hat{\mu}_{DRMI}$	-0.00	0.03	0.03	0.95
$\hat{\mu}_{IPWCC \cdot \pi - \text{false}}$	0.25	0.06	—	—
$\hat{\mu}_{DRMI \cdot \pi - \text{false}}$	0.00	0.03	0.03	0.95
$\hat{\mu}_{MI \cdot y - \text{false}}$	0.49	0.05	—	—
$\hat{\mu}_{DRMI \cdot y - \text{false}}$	-0.04	0.03	0.03	0.95
$\hat{\mu}_{DRMI \cdot \pi \oplus y - \text{false}}$	0.22	0.04	0.04	0.80

Table 9.4: The results of the non-monotone cross-sectional simulation study where doubly robust multiple imputation (DRMI) is compared with IPWCC and ordinary MI. The known (true or $\sqrt{\text{true}}$) probability weights were used in the IPWCC and DRMI methods. No subscript indicates correct specification of the y -model and weights (where applicable). $\pi - \text{false}$ indicates that the square root of the weights were used, $y - \text{false}$ indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y - \text{false}$ indicates that both the weights and y -model were incorrect.

When the missing data are longitudinal but *non-monotone*, the Bang and Robins (2005) method can no longer be used, but our estimator works very well: it exhibits the desired double robustness property as well as improved efficiency compared with IPWCC. The loss of efficiency relative to OR and MI is negligible in our simulation studies. Furthermore, our method is easily implemented in standard software packages such as `ice` in Stata.

We have also shown that DRMI could in principle be applied to general (non-longitudinal) non-monotone data. However, the problem of estimating the variable-specific inverse probability weights needs first to be resolved. Unfortunately, the method proposed by Robins and Gill (1997) for estimating the complete-case weights can not be used to identify the variable-specific weights. We have shown, by substituting the known true weights, that if a method were developed for estimating these probabilities, DRMI could be used and would perform very well.

Although our focus has been on examples where the aim is to estimate the marginal

mean of one of the variables, DRMI can be used much more generally (for example to estimate the parameters of a regression of one variable on another) and as easily to any appropriate analysis of the imputed data.

We have used proper imputation throughout for the simple reason that Rubin's variance formula can then be used. More efficient estimates could in principle be obtained by imputing *improperly*, but bespoke variance estimators would then be required.

Part IV

The RECORD study

10

Doubly-robust MAR analysis

10.1 Introduction

In Chapter 9, we developed a new method, doubly robust multiple imputation (DRMI), for constructing doubly robust estimates. One of our method's main advantages is the conjectured extension to the non-monotone setting. In Chapter 11, we perform MNAR sensitivity analyses on the RECORD study data (see Chapter 2 for some background

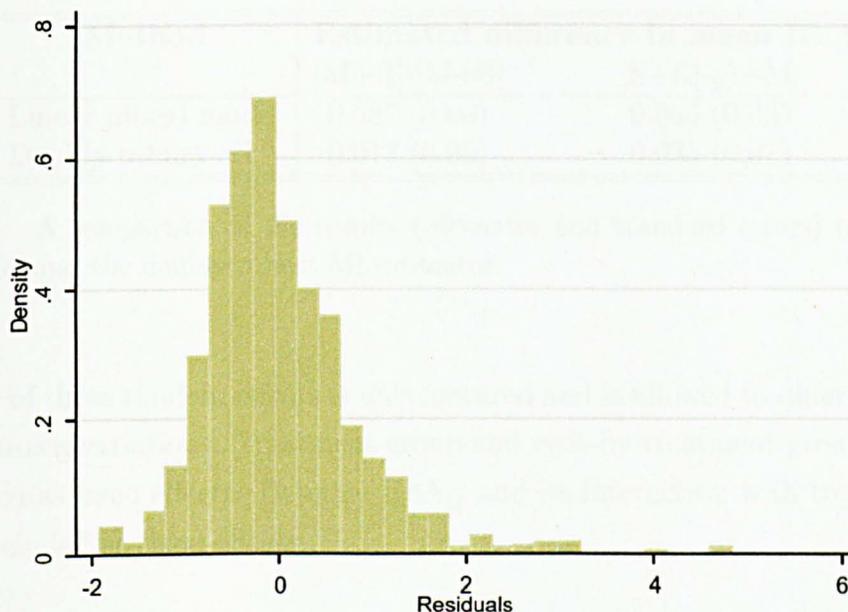


Figure 10.1: A histogram showing the distribution of the residuals from a linear regression of HbA_{1c} on treatment group and baseline HbA_{1c} for the observed data at the final timepoint.

to this study and §11.2 for a brief description of the missing data patterns). As a point of departure for our sensitivity analyses, we plan to use a direct likelihood approach under multivariate normality and MAR. Before we do this, however, it is important to explore whether or not the Gaussian direct likelihood is an appropriate choice of analysis for this purpose. The possible non-normality of the HbA_{1c} outcome variable in the RECORD data (see Fig. 10.1) suggests that a more robust analysis might be needed. In this chapter, we use the method developed in the previous chapter to perform a doubly robust analysis of the RECORD data under the MAR assumption and we compare it with the direct likelihood approach.

10.2 Methods

For the direct likelihood analysis, we use PROC MIXED in SAS. Time is included as a categorical variable with a fixed and random effect included at each visit. The

Method	Estimated difference in mean HbA _{1c}	
	[M+R]-[M+S]	[S+R]-[S+M]
Linear mixed model	0.087 (0.08)	0.066 (0.08)
Doubly robust MI	0.017 (0.09)	0.033 (0.07)

Table 10.1: A comparison of the results (estimates and standard errors) from the linear mixed model and the doubly robust MI estimator.

covariance of these random effects is unstructured and is allowed to differ by treatment group (complex variation). Treatment group and visit-by-treatment-group interactions are included as fixed effects. Baseline HbA_{1c} and its interaction with treatment group are also included as fixed effects.

Then, we apply to the same dataset the DRMI procedure described in §9.2.2. The missingness model corresponding to Fig. 9.1 is appreciably more complex with 8 variables rather than 3. It is clear that some reduction in the dimensionality of the problem must be made if the weights are to be estimated efficiently. There is a trade-off between efficiency and robustness, but this is necessary in practice with this number of timepoints. We will impose the restriction that, conditionally on the most recently observed outcome, the choice of which outcome will be the next non-missing outcome is independent of all other observed outcomes. Apart from this, the method is identical to the one described in the simulation study in §9.3.3.

10.3 Results and conclusions

The results are shown in Table 10.1. We see that the results from DRMI are similar (but not identical) to those from the direct likelihood analysis. Certainly as regards the pre-specified non-inferiority criterion of 0.4%, neither method supports the rejection of non-inferiority.

Figs. 10.2–10.5 show the profiles for each treatment group as predicted by the two

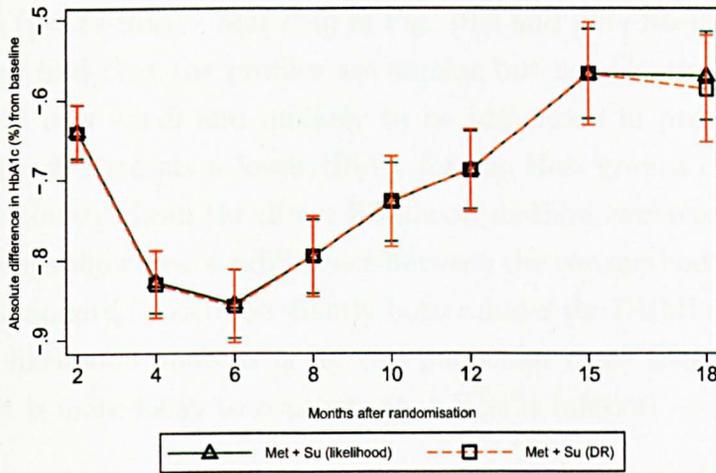


Figure 10.2: A comparison of the HbA_{1c} profiles predicted from the direct likelihood analysis and those predicted from the doubly robust multiple imputation analysis for the Met+Su arm. The green and red error bars show \pm the standard errors for the likelihood and DRMI analyses respectively.

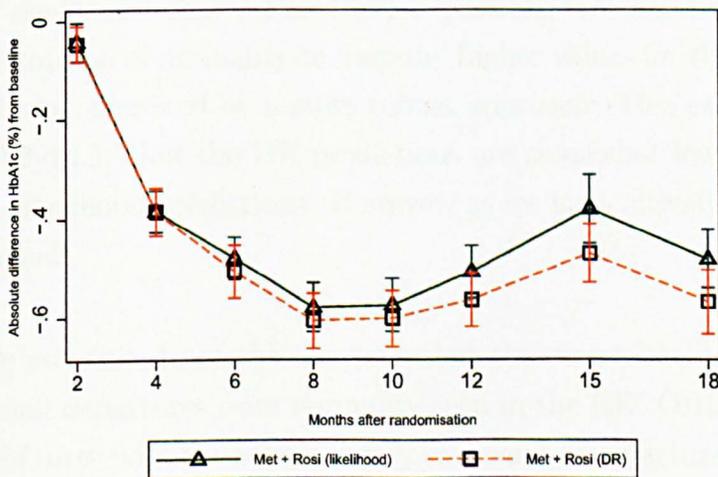


Figure 10.3: A comparison of the HbA_{1c} profiles predicted from the direct likelihood analysis and those predicted from the doubly robust multiple imputation analysis for the Met+Rosi arm. The green and red error bars show \pm the standard errors for the likelihood and DRMI analyses respectively.

models and Figs. 10.6–10.7 show the differences between these profiles for the two arms separately ($[\text{Met}+\text{Rosi}] - [\text{Met}+\text{Su}]$ in Fig. 10.6 and $[\text{Su}+\text{Rosi}] - [\text{Su}+\text{Met}]$ in Fig. 10.7). Again, we find that the profiles are similar but not identical. The differences are substantively very small and unlikely to be important in practice. If anything, the DRMI approach suggests a lower HbA_{1c} for the Rosi groups compared with the corresponding estimates from the direct likelihood method, whereas the estimates for the standard groups show less of a difference between the two methods. As a result, Rosi compared with standard looks to be slightly better under the DRMI analysis suggesting that the direct likelihood analysis is (in this particular case) slightly conservative in the sense that it is more likely to conclude that Rosi is inferior.

The reason for there being only a small difference between the two approaches is probably that the non-normality (as suggested by Fig. 10.1) is very small. We notice that what little difference there is increases over time. This could be due to the increased dependence on modelling assumptions in the direct likelihood approach as the number of missing observations increases.

Given that the skewness (seen in Fig. 10.1) is positive, we would expect an analysis based on an assumption of normality to ‘impute’ higher values for the missing observations than would be suggested by a more robust approach. This explains the pattern seen in Figs. 10.2–10.5: that the DR predictions are somewhat lower than the corresponding direct likelihood predictions. However, as we have already mentioned, these differences are small.

In summary, we conclude from this analysis that the direct likelihood is sufficiently robust to the small departures from normality seen in the RECORD study for a more robust analysis of these data to be unnecessary as a point of departure for the sensitivity analyses in the next chapter. We have also seen that the DRMI method proposed in the previous section can be applied easily in a real example with non-monotone missing data and many timepoints.

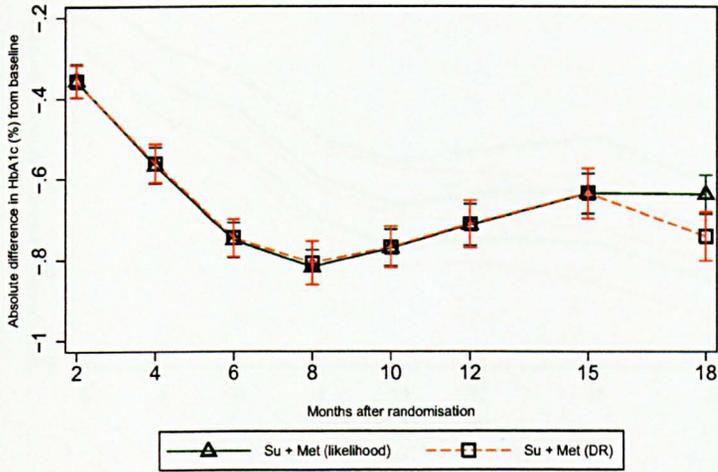


Figure 10.4: A comparison of the HbA_{1c} profiles predicted from the direct likelihood analysis and those predicted from the doubly robust multiple imputation analysis for the Su+Met arm. The green and red error bars show \pm the standard errors for the likelihood and DRMI analyses respectively.

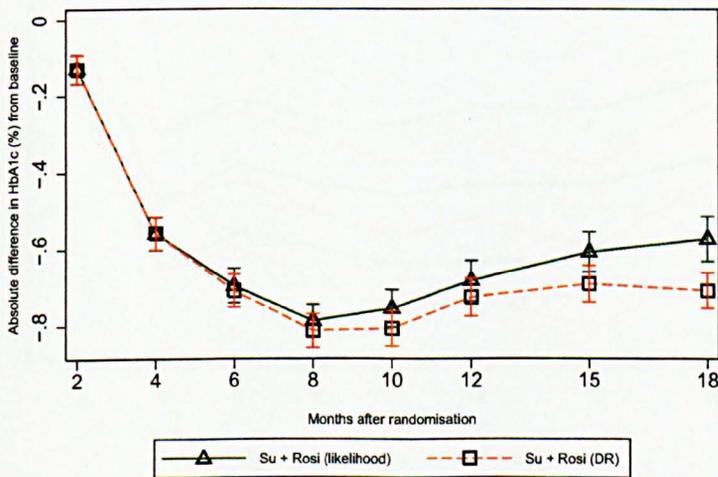


Figure 10.5: A comparison of the HbA_{1c} profiles predicted from the direct likelihood analysis and those predicted from the doubly robust multiple imputation analysis for the Su+Rosi arm. The green and red error bars show \pm the standard errors for the likelihood and DRMI analyses respectively.

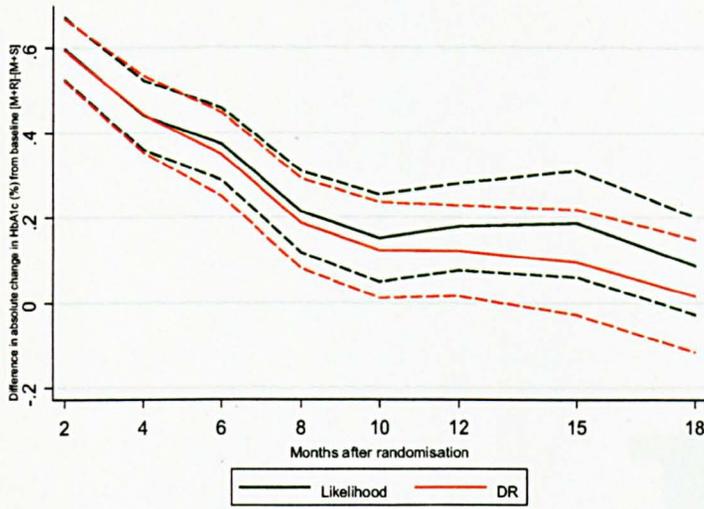


Figure 10.6: The differences between the HbA_{1c} profiles for the Met+Rosiglitazone and Met+Sulfonylurea arms. The solid green and red lines show the predicted differences from the likelihood and DRMI analyses respectively, and the dotted lines show \pm the pointwise standard errors for these differences.

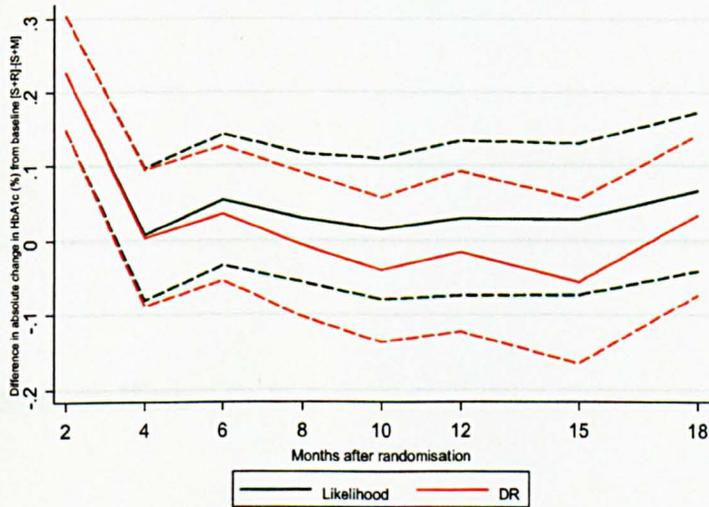


Figure 10.7: The differences between the HbA_{1c} profiles for the Sulfonylurea+Rosiglitazone and Sulfonylurea+Metformin arms. The solid green and red lines show the predicted differences from the likelihood and DRMI analyses respectively, and the dotted lines show \pm the pointwise standard errors for these differences.

11

MNAR sensitivity analyses

11.1 Aims and outline

The aim of the work presented in this chapter is to assess the potential impact of missing data and non-compliance on the conclusions drawn from the 18-month RECORD glycaemic analysis. In particular, having studied (in the previous chapter) the robustness of these conclusions to the multivariate normality assumption, we now look at robustness of a different kind: that to the MAR assumption inherent in the direct

likelihood approach.

In §11.2 we give the results of some simple analyses of the patterns of non-compliance and missing data found in the RECORD 18-month data.

Section 11.3 comprises a discussion of the possible research questions we could ask in this setting. The issue of non-compliance is key here, and the way in which the preferred question should be answered is inextricably linked to the missing data mechanism and the longitudinal structure of the data. Thinking carefully about the precise question to be answered before formulating a particular analysis is necessary for meaningful interpretation of the results, and, in the presence of non-compliance and missing data, even more care is needed.

In §11.4 we present the results of a series of sensitivity analyses carried out to assess the robustness of the results of the direct likelihood analysis as carried out by Home *et al.* (2007) to possible violations of the inherent assumptions (namely missing at random and compliance at random) made when using this approach.

11.2 Patterns of missing data and non-compliance in the 18-month RECORD data

A logistic regression of non-missingness at the final time point, with age, gender, race, and baseline HbA_{1c} as predictors provided little or no evidence of an association between age and non-missingness ($p = 0.7$), race and non-missingness ($p = 0.2$) or sex and non-missingness ($p = 0.4$). However, there was evidence of an association between treatment group and the probability of being observed ($p = 0.02$) and between baseline HbA_{1c} and the probability of being observed ($p = 0.01$). After controlling for baseline HbA_{1c}, those on Su+Met had a 51% reduction in their odds (95% CI: 20%–70%) of being observed as compared with those on Met+Su. After controlling for treatment group, an absolute increase of 1 unit (i.e. 1%) in baseline HbA_{1c} was associated

with a reduction of 28% (95% CI: 8%–43%) in the odds of being observed at the final timepoint.

A similar analysis, but this time including non-compliance as a form of missingness, showed again that there was little or no evidence of an association between age and non-missingness ($p = 0.4$), race and non-missingness ($p = 0.3$) or sex and non-missingness ($p = 0.5$). The association between treatment group and the probability of being observed (and complying) was stronger ($p = 0.002$). After controlling for baseline HbA_{1c}, those on Met+Rosi had a 47% reduction in their odds (95% CI: 29%–76%) of being observed as compared with those on Met+Su, with the corresponding reductions for Su+Met being 50% (95% CI: 31%–81%) and for Su+Rosi 46% (95% CI: 29%–73%). After controlling for treatment group, an absolute increase of 1 unit (i.e. 1%) in baseline HbA_{1c} was associated with a reduction of 64% (95% CI: 52%–78%) in the odds of being observed (and complying with protocol) at the final timepoint ($p < 0.001$). This association is also stronger than for loss-to-follow-up alone.

11.3 What are the questions and how can we answer them?

Before we can decide on an appropriate analysis plan, we must be clear on precisely what are the questions to which we wish to find answers.

11.3.1 Treatment vs. assignment to treatment

The possible questions can be divided into two broad categories:

1. questions about the actual biological effect of a treatment, and
2. questions about the effect of *being assigned* to a treatment.

Sometimes, the effect of *being assigned* to a treatment is what is of interest to us: when we wish to use the data to choose the best (in public health terms) policy for treating patients. In most other circumstances, the actual biological effect of a treatment is more likely to be of interest.

An intent-to-treat (ITT) analysis allows us to answer questions of type 2., but is often used when 1. is of interest. The reason for this is that a valid ITT analysis (when there are no missing data) is usually straightforward, since the randomisation can be relied upon to eliminate bias. Valid analyses that answer questions of type 1. are less common, apart from in the special case where everyone complies fully to the randomised treatment, with no loss-to-follow-up, in which case 1. and 2. are the same, and an ITT analysis will answer both.

In summary, if interest lies in a question of type 2., then we can analyse *as randomised* (or *by intention to treat*), which deals with the issue of non-compliance. In all other circumstances we have to choose between attempting to answer the appropriate question 1., running the risk that the effect of non-compliance has not been adequately taken into account, or answering correctly the other question (i.e. question 2.), which we didn't really want to ask.

11.3.2 Populations

Assuming that we know what it is we would like to ask, there is still the problem of "about whom do we want to ask it?". For the RECORD study, here are some possibilities:

1. All type II diabetics
2. All type II diabetics for whom single therapy is insufficient

3. All type II diabetics for whom single therapy is insufficient and for whom double therapy *is* sufficient
4. All type II diabetics for whom single therapy is insufficient at a particular point in time (e.g. the beginning of the trial)
5. All type II diabetics for whom single therapy is insufficient and for whom double therapy *is* sufficient at a particular point in time (e.g. the beginning of the trial)
6. All type II diabetics who would choose to take double therapy were it to be offered to them

An added complication is that, for example, 2. and 3. do not define a population independently of time, but rather the population they define varies over time. In other words, double therapy may be sufficient for a certain patient in April, but by May it could cease to be so.

11.3.3 Objectives of the RECORD 18-month analysis

The primary objective of the 18-month Glycaemic Control analysis of the RECORD data was given as:

“to test whether the 18-month mean change from baseline HbA_{1c} for the intention-to-treat (ITT) population (all randomised, treated and with at least one data point post-randomization) with rosiglitazone oral combination therapy was at least as good as the respective controls receiving metformin + sulphonylurea.” (Home *et al.*, 2007)

If

- i. every member of the randomised population had remained on his/her assigned treatment protocol for the entire 18-month follow up, and
- ii. there were no missing data

then an ANCOVA using the HbA_{1c} results at the 18-month observation, adjusted for baseline, would be the obvious choice of analysis. However, neither of the above conditions holds in this dataset. This means that in any analysis we choose to carry out, some assumptions will have to be made about both

- i. the mechanism which determines which subjects leave the assigned treatment protocol, and
- ii. the mechanism which determines which observations are missing.

Let us turn first to the issue of missing data, before turning to the issue of non-compliance in §11.3.3.2.

11.3.3.1 Missing data

In their monograph *Missing Data in Randomised Controlled Trials -- a Practical Guide*, Carpenter and Kenward (2008) compare two of the main MAR approaches to handling missing data in clinical trials, namely direct likelihood (or modelling) and multiple imputation (MI). They make the following remark:

“[I]t is worth noting that, as the imputation model is multivariate normal, as are the models we fit here, treatment effects can always, in principle, be estimated directly through modelling. The advantages of modelling are that it is quicker (our largest models fitted within 10 minutes), involves

fewer judgments (such as whether the an MCMC sampler has converged) and yields a unique maximum likelihood estimate. By contrast, inferences from MI are slightly different each time. Where the precise answer is critical for decision making, a substantial number of imputations may be necessary to get the Monte-Carlo variability acceptably low. We therefore advocate direct modelling, if possible.”

When the dataset is sufficiently large for the gains in efficiency offered by a more parsimonious model to be very small, Carpenter and Kenward (2008) go on to advocate the use of direct likelihood, with a separate unstructured covariance structure in each treatment group (complex variation), and an unstructured means model which includes the baseline measurement, baseline-by-visit interaction, and baseline-by-treatment interaction.

This is precisely the method adopted in the RECORD 18-month analysis (Home *et al.*, 2007). In terms of the analysis at the final timepoint, observations from earlier timepoints on incompletely observed individuals can be used—in a principled way, given the MAR assumption—to provide information on the possible values which may have been observed on these subjects had they been observed at the final timepoint. In addition, the use of an analysis which models the longitudinal profile of individuals over time means that should we be interested in anything other than the analysis at the final timepoint, this information is available to us.

Molenberghs and Kenward (2007) similarly advocate the use of direct likelihood, in the general sense described above, but both books go on to give a very important caveat. For example, Molenberghs and Kenward (2007) say:

“[R]egardless of the elegance and beauty of the direct likelihood analysis, MNAR can almost never be ruled out as a mechanism and therefore one ought to consider the possible impact of such mechanisms as well.”

In §11.4 we assess the sensitivity of the results of the direct likelihood analysis as applied to the 18-month data on HbA_{1c} to possible violations of the MAR assumption.

11.3.3.2 Non-compliance

The analysis originally undertaken of these data by Home *et al.* (2007) censored (i.e. treated as missing) any observations from subjects who had left dual therapy, from the point at which they stopped taking dual therapy. Some of these subjects went on to receive triple therapy, some started receiving insulin injections and others returned to the monotherapy they were taking before the trial commenced.

Except for two subjects, once a subject leaves dual therapy, he/she never returns. In other words, the induced missing data (induced by the adopted censoring policy) are monotone.

When the missing data are induced by non-compliance, the MAR assumption says:

If two subjects, A and B , exhibit identical behaviour up to some point t , whereafter A continues on dual therapy, while B changes to some other therapy, then the subsequent behaviour of B , had B stayed on dual therapy (something which we have not observed) is assumed to be identical *in distribution* to the subsequent behaviour of A .

We will refer to this as *Non-compliance At Random* (or NAR), since it assumes that, after conditioning on a subject's observations whilst on dual therapy, the probability of leaving dual therapy is independent of those observations which would have been observed on dual therapy, had the patient not changed to another treatment.

Ignoring for the time being the issue of missing data beyond that which is potentially induced by non-compliance, there are two broad approaches to handling the issue of

non-compliance. The first is to analyse the data from all subjects according to the treatment protocol to which they were randomised, irrespective of whether or not they adhered to this protocol. This is called *intent-to-treat* (or ITT). This has the advantages of preserving the effect of randomisation, and of allowing the practical effects of implementing such treatment policies to be assessed. For example, if a treatment has an undesirable side effect which causes many patients to withdraw from taking it, this will be reflected in the conclusions. On the other hand, if our main interest is in the biological efficacy of one treatment compared with another, then we might choose the alternative approach, which is to include in the analysis only those who adhered to the protocol to which they were assigned. This is called *per protocol* (or PP).

In his seminal book *Clinical Trials: A Practical Approach*, Pocock (1983) writes:

“[Should patients] with protocol deviations be included in the main treatment comparisons *or* should they simply be noted as being deviates and be excluded from subsequent results? In most circumstances I think the first approach is required; that is, *all eligible patients, regardless of compliance with protocol should be included in the analysis of results whenever possible*. This ‘*pragmatic approach*’ is sometimes called ‘*analysis by intention to treat*’ and is normally preferred since it provides a more valid assessment of treatment efficacy as it relates to actual clinical practice. The alternative ‘*explanatory approach*’ would confine analysis to patients who received therapy according to protocol, i.e. ‘*analysis of compliers only*’, but this can distort treatment comparisons.”

However, in a longitudinal setting, where a subject may be observed on protocol up to a certain point, after which s/he is observed as having deviated from the protocol, we are faced with a slightly different situation. The main objection to a standard PP analysis is that to analyse only those subjects who do not deviate from the protocol makes the (often implausible) assumption that those who do not deviate from the protocol form a random subset of the whole. Following the above taxonomy, this could

be thought of as *Non-compliance Completely At Random* (or NCAR). In a repeated measures setting, by including observations from subjects up to the point at which they deviate from the protocol, we assume only NAR (not NCAR), and the objections are less strong. However, in cases where the NAR assumption is violated, a PP-NAR analysis (such as the one carried out on these data by Home *et al.*, 2007) could give rise to a biased estimate of the true biological treatment effect.

This (together with the possible deviations from the MAR assumption discussed in §11.3.3.1) provide the rationale for the work described in §11.4, but first, let us look at what a true ITT analysis entails when there are missing data.

11.3.4 ITT analyses with missing data

11.3.4.1 Background

If we restrict ourselves to an ITT analysis, we automatically include in the analysis all the observations censored in the PP-NAR analysis. Specifically, this means that the data for those who changed to triple therapy, but who remained to be assessed at 2-monthly intervals, will be included.

However, there remains the problem of how to deal with the data that are truly missing. By design, any subject leaving dual therapy but not moving to triple therapy (for example, those from the Met+Su and Su+Met arms who went on to take insulin, or who returned to monotherapy) were moved to the CV outcomes stage of the trial, in which HbA_{1c} was only measured on a yearly basis. We have thorough records of the treatments taken by these subjects during this time. Furthermore, there are other subjects who genuinely dropped out and for whom we have no further data (neither for HbA_{1c} nor the treatments they were taking). The question now is, what does an ITT analysis entail in this situation?

The first thing to note is that a MAR analysis is likely to be inappropriate, unless those who drop out (or who move to the CV outcomes phase) continue to take the treatment to which they were originally assigned. We know by definition that this cannot be the case for those who move to the CV outcomes phase; also, for those who drop out completely, it is not possible for them to continue to take the additional treatment to which they were randomised once they withdraw from the trial. Thus, we are faced with a situation similar to the one described by Little and Yau (1996). The authors of this paper advocate the use of multiple imputation as a way of carrying out MNAR sensitivity analyses in this sort of situation. In the next section, we describe how MI may be used in ITT sensitivity analyses with missing data.

11.3.4.2 Multiple imputation and intent-to-treat

If, instead of fitting a model to the incomplete RECORD data by direct likelihood, we use multiple imputation to complete the data several times, and then fit direct likelihood model to the completed datasets, combining parameter estimates using Rubin's rules, then, provided we use the same model for the imputation as we use for the analysis, the results from both methods (MI and direct likelihood) will tend to the same parameter estimates (and standard errors) as the number of imputed datasets tends to infinity. Roughly speaking, this follows from the fact that multiple imputation is a Bayesian procedure, and Bayesian and likelihood analyses coincide as the sample size tends to infinity. However, varying the imputation model—specifically, allowing for an effect due to unobserved variables—allows us to fit a MNAR model. The idea is particularly natural in our setting, where we know (or can guess) what treatment is taken by those subjects for whom we have missing observations.

Our aim is first to impute the missing values using multiple imputation where the imputation model reflects the treatment *actually taken* by the subject with missing observations. Following this, the completed datasets are analysed as randomised and the results combined according to Rubin's rules in the usual way. Until we pause to think about this, we may not be entirely happy that this is a true ITT analysis, since

the first part of the analysis (the imputation step) is “as treated” and only the final part is “as randomised”. But this is exactly what we mean by an ITT analysis in the presence of missing data. Intuitively, we can think of the imputation step as imputing the sort of behaviour we would have expected to see from those who dropped out had they been observed, keeping all other factors constant. This means that if someone stops taking treatment altogether and hence experiences a sharp rise in his/her HbA_{1c}, say, this should be reflected in the imputation. When the imputation has been carried out in a way that reflects the actual or posited treatment compliance, the imputed data can be analysed “as randomised”.

11.4 Sensitivity analyses

As we saw in the last section, an ITT analysis of these data requires that any data censored due to non-compliance must be reintroduced into the analysis. An ITT analysis assuming MAR would then impute the (genuinely) missing data using a multiple linear regression imputation model with observed data and *true treatment profile* included as predictors. Our aim now is to assess the sensitivity of these results to the MAR assumption through the introduction of sensitivity parameters into the imputation model. This is described in detail in §11.4.1.

For the per protocol analysis, the direct likelihood analysis carried out by Home *et al.* (2007), represents the missing at random (MAR) and non-compliance at random (NAR) model. We can use a similar imputation approach to assess the sensitivity of this model to violations of both the MAR and NAR assumptions. This is described in detail in §11.4.2.

11.4.1 Strategy for MNAR ITT analyses on the 18-month RECORD data

A treatment profile for each individual was extracted from the various data available on treatments taken. Those who were not censored from the analysis carried out by Home *et al.* (2007) were assumed to be on dual therapy. A “triple therapy flag” variable was used to identify those on triple therapy, and the profiles for those in the CV outcomes phase were deduced from the (detailed) information available on the treatments taken at various times. For those lost-to-follow up, a profile which returned to monotherapy (equal to their original background therapy) was assumed.

For the censored individuals, and those who were lost-to-follow up, dates for their scheduled clinic visits were imputed simply by adding 61 (or 91 for the last two timepoints) days to their previous clinic visit. This was necessary in order that the treatment profile could be converted into treatments taken at different visits.

If we had observed data corresponding to each of the treatment profiles followed by the subjects with missing data, we could fit one MNAR analysis which used the observed data from the appropriate treatment profile to impute the missing observations. This is the case for those on dual or triple therapy. Unfortunately, for treatment profiles involving insulin or monotherapy, there are some timepoints for which no-one has observed data corresponding to these profiles. This means that to impute the data for these subjects, we will need to ‘borrow’ imputations from a different treatment profile, but we can allow the actual imputations to vary from the borrowed imputations by a constant parameter, and we can assess the sensitivity of our results to the value taken by this parameter by carrying out the imputations under several different values for this parameter.

The strategy can be summarised as follows:

1. Impute (5 times) intermediate missing values for those who have not deviated from the protocol (128 measurements on 32 subjects) assuming MAR conditional

only on the past. This is not fully-efficient but is still a valid approach and allows us to carry out all the imputations in Stata using the regression method with large-sample normal approximations to the Bayesian posterior distributions of the parameters, without any need for MCMC chains etc. For each timepoint in turn, the dependent variable in the linear regression imputation model is the HbA_{1c} measurement at that visit, and the independent variables are all previous measurements of HbA_{1c}, and the treatments actually taken for the previous 2/3 months.

2. Impute (5 times) monotone missing values for those who have not deviated from the protocol (163 measurements on 40 subjects) assuming MAR and the regression imputation method in Stata, as described above.
3. Impute (5 times) missing values for those who have moved to triple therapy (163 measurements on 19 subjects) using the observed data from subjects on triple therapy. For these imputations, to increase precision, use data from subjects on both background Met and background Su who have moved to triple therapy.
4. Impute (5 times) missing values for those who have moved onto insulin (48 measurements on 16 subjects) using the observed data from subjects on triple therapy but adding a constant γ_1 at each timepoint. Those who move onto insulin may have a lower HbA_{1c}, due to insulin being more effective than anti-diabetic drugs; in which case, we would expect γ_1 to be negative. On the other hand, if those who move onto insulin do so because of poor control, it is also plausible that γ_1 be positive.
5. Impute (5 times) missing values for those who have moved back to monotherapy (4 measurements on 2 subjects) using the observed data from subjects on non-experimental dual therapy Met+Su but adding a constant γ_2 at each timepoint. Those who return to monotherapy may have a higher HbA_{1c}, if one therapy is less effective than two; in which case, we would expect γ_2 to be positive.
6. Some people in the CV outcome phase receive an additional anti-diabetic drug which is not Met, Su nor Rosi. Impute (5 times) missing values for those who have moved to Met+Other or Su+Other (7 observations on 5 subjects) using

the observed data from subjects on Met+Su but adding a constant γ_3 at each timepoint.

7. Analyse and combine the 5 completed datasets using PROC MIXED and PROC MIANALYZE, separately in the two arms and analysing “as randomised”.
8. Carry out steps 1. to 7. for $(\gamma_1, \gamma_2, \gamma_3) = (0,0,0), (-0.25,0.25,0), (0.5,0.5,0), (-1,1,0), (0.25,-0.25,0), (0.5,-0.5,0), (1,-1,0)$.
9. Use the fact that HbA_{1c} is measured on some people in the CV outcomes phase at 12 months to attempt to choose a “best” combination of $(\gamma_1, \gamma_2, \gamma_3)$. Calculate multiply imputed values for this measurement as if it had not been observed, under each combination of 7 different values for $(\gamma_1, \gamma_2, \gamma_3)$, ($\gamma_j \in \{-0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6\}$), making $7^3 = 343$ combinations in total. Calculate the mean squared difference between the imputed and observed values to try to decide on a “best” set of $(\gamma_1, \gamma_2, \gamma_3)$, i.e. the set that leads to the minimum mean squared difference between the imputed and the observed outcomes at the 12-month timepoint.
10. Repeat steps 1. to 7. using this “best” combination.
11. Compare the results from each model with regards to the estimate and inference for the treatment difference at the final time point and compare the profiles implied by each model.

11.4.2 Strategy for MNAR/NNAR PP analyses on the 18-month RECORD data

In the previous two sections, we described how MNAR intent-to-treat sensitivity analyses could be carried out to assess the robustness of the MAR/NAR per protocol model (the direct likelihood approach, with individuals censored at the point of non-compliance) to violations of the MAR and NAR assumptions. There is no reason why these sensitivity analyses need be limited to the ITT setting. Supposing instead that

we are interested in true difference in biological effect between the therapies, we can again use MI to fit alternative models that allow for MNAR and NNAR mechanisms.

The strategy differs from that of §11.4.1 in a few important ways:

1. The observations that were censored in the direct likelihood analysis remain censored.
2. The 12-month observations for those in the CV outcome phase can no longer be meaningfully compared with the imputations. This is because the objective of the imputation stage has changed in moving from ITT to PP. The imputations now aim to reflect what would have been observed in these patients *had they continued to comply with the dual-therapy protocol* to the end of the 18-month follow-up. This means that we expect the imputations to differ from the observed 12-month measurements.

We draw the imputations (5 times) from a normal linear regression imputation model which exactly matches the analysis model (and thus includes all previous measurements of the outcome and the combination therapy to which the patient was randomised). Combining (using Rubin's rules) the results from a direct likelihood analysis on these completed datasets would lead to a MAR/NAR per protocol analysis, equivalent (as the number of imputations and the sample size tend to infinity) to the direct likelihood approach on the original incomplete dataset. However, to form the sensitivity analyses, we introduce four sensitivity parameters, $(\delta_1, \delta_2, \delta_3, \delta_4)$. δ_1 is added to each of the MAR/NAR imputations for those individuals lost to follow-up, δ_2 is added to each of the MAR/NAR imputations for those individuals who went on to triple therapy, δ_3 is added to each of the MAR/NAR imputations for those individuals who went on to insulin, and δ_4 is added to each of the MAR/NAR imputations for those remaining individuals who left dual therapy but did not receive triple therapy nor insulin. In other words, $(\delta_1, \delta_2, \delta_3, \delta_4)$ represent deviations from the MAR/NAR assumptions, and—if positive—represent the degree to which the condition of those who were lost

to follow-up or who left the protocol would have been worse (over and above what is predicted by the MAR/NAR assumption) than those who complied, had they stayed on dual therapy. It is important to note that the MAR/NAR model allows the HbA_{1c} of those lost to follow-up or those who did not comply with the protocol to be higher than that of the other patients, up to the level predicted by the previous measurements, and therefore $(\delta_1, \delta_2, \delta_3, \delta_4)$ represent the *additional* increase not picked up by the previous measurements. For completeness, even though less plausible, we will also consider negative values for these parameters.

We experiment with different values of the sensitivity parameters:

$$(\delta_1, \delta_2, \delta_3, \delta_4) = \left\{ \begin{array}{l} (0,0,0,0) \\ (0.25,0.25,0.25,0.25) \\ (0.5,0.5,0.5,0.5) \\ (1,1,1,1) \\ (0.25,-0.25,-0.25,0.25) \\ (0.5,-0.5,-0.5,0.5) \\ (1,-1,-1,1) \\ (-0.25,0.25,0.25,-0.25) \\ (-0.5,0.5,0.5,-0.5) \\ (-1,1,1,-1) \\ (-0.25,-0.25,-0.25,-0.25) \\ (-0.5,-0.5,-0.5,-0.5) \\ (-1,-1,-1,-1) \end{array} \right.$$

If it transpires that none of the above combinations leads to a conclusion of inferiority of Rosiglitazone (according to the pre-specified non-inferiority margin of 0.4%), more extreme combinations will be considered until a combination is found which *does* imply inferiority of Rosiglitazone in at least one of the two arms of the trial. In other words, our aim is to find a *tipping point*, described in terms of our sensitivity parameters, where if the missingness mechanism is *further* from MAR than this point, our final conclusions are affected. A clinical expert in diabetes could then give his/her opinion

on whether or not such a departure is plausible. S/he could give a (subjective) opinion on whether or not such a difference between compliers and non-compliers or between observed and unobserved subjects is plausible. By definition, we cannot use the data to answer this question.

11.4.3 Results

11.4.3.1 ITT

The top half of Table B.1 summarises the final timepoint analyses under each of the different ITT models. Step 10. in the strategy (see §11.4.1) gave rise to a combination $(\gamma_1, \gamma_2, \gamma_3) = (0.2, 0.4, 0)$ and the agreement between the observed HbA_{1c} and the imputed HbA_{1c} at the 12-month timepoint as imputed using this combination is illustrated in Fig. B.9. Fig. B.1 shows the profiles under the MAR per protocol analysis, while Figs. B.2–B.8 and B.10 show the profiles under each of the different MNAR ITT sensitivity models.

11.4.3.2 Per protocol

The bottom half of Table B.1 summarises the final timepoint analyses under each of the different PP models. None of the pre-specified combinations of sensitivity parameters led to a conclusion of inferiority of Rosi and (after increasing the sensitivity parameters in steps of 0.2), the first model to conclude Rosi to be inferior (in the background Met arm only) was the one in which $(\delta_1, \delta_2, \delta_3, \delta_4) = (3, 3, 3, 3)$. Fig. B.1 shows the profiles under the MAR per protocol analysis, while Figs. B.11–B.24 show the profiles under each of the different MNAR/NNAR per protocol sensitivity models.

11.4.4 Discussion

It should be stressed that the sensitivity analyses reported here account for only a small proportion of the huge number of sensitivity analyses that could have been considered. For example, in the ITT analyses we assume throughout that any deviation from the MAR assumption is purely due to a change in treatment on withdrawal. We may have reason to believe that those who drop out exhibit further deviations from the observed subjects, beyond what can be explained by treatment. Also, steps 4. to 7. in the strategy described in §11.4.1 may be too simplistic. We are implicitly assuming that, for example, the Met+Su and Met+Other arms deviate by a constant amount over time. Another assumption made in the ITT sensitivity analyses was that the treatment profile for any subject who dropped out from the trial completely changed down to monotherapy at the point at which drop-out occurred. This may not be valid, as some subjects presumably moved onto insulin or other additional therapies in this group also. There is no limit to the number of scenarios one could investigate in a sensitivity analysis setting, and at best we hope to select a few pertinent examples that may be representative in some way of the changes that might be seen.

The fact that some patients in the CV outcomes phase had data at the 12-month timepoint at least gave us some scope for checking how plausible our choices of $(\gamma_1, \gamma_2, \gamma_3)$ in the ITT sensitivity analyses might be. This method suggested that our more extreme choices (such as $(1, -1, 0)$) were not plausible. Table B.1 and Figs. B.1–B.8 and B.10 all demonstrate that the results are somewhat sensitive to our choice of model. Only under the model $(\gamma_1, \gamma_2, \gamma_3) = (-1, 1, 0)$ would a conclusion of inferiority of Rosiglitazone be made according to the pre-specified non-inferiority margin of 0.4%. This is illustrated on the left hand side of Fig. B.5, where the results are consistent (at a 5% level) with a difference of 0.4% in favour of Met+Su. However, when we inspect the results from our “best” model, the only slight difference is that the direction of the estimate of the difference in the background Su arm has reversed, i.e. if anything, Rosiglitazone looks *better* under this model than under the original MAR/NAR per protocol analysis.

Note that the standard errors for our estimates of the treatment differences look to be

larger for $(\gamma_1, \gamma_2, \gamma_3) = (-1, 1, 0)$ and $(\gamma_1, \gamma_2, \gamma_3) = (1, -1, 0)$. This supports the suggestion that these imputation models are incorrect, and that spurious variation is being injected into the imputations, due to the overinflated deviations between the borrowed imputations and the missing observations.

To sum up, it seems from these sensitivity analyses that a conclusion of non-inferiority is relatively stable to model changes that allow for a MNAR ITT analysis and even more stable to model changes that allow for a MNAR/NNAR per protocol analysis. In order for the non-inferiority conclusion to be challenged, it was necessary for the MAR/NAR model to underestimate the unobserved HbA_{1c} by as much as 3%. Given that all the observed differences in HbA_{1c} observed in the trial were less than 1% over the entire follow-up period, it seems implausible that the MAR/NAR model should underestimate the HbA_{1c} of the non-compliers and those who left the trial by this amount. However, by varying the assumptions of the MNAR/NNAR models, some differences were seen in the implied profiles and the resulting final timepoint analysis, even if these differences were of the same order of magnitude as (or smaller than) the associated standard errors in all models except for one. The results therefore should not be seen as casting doubt on the conclusions of the 18-month MAR/NAR per protocol analysis in this case, but they do illustrate the need to consider sensitivity analyses in this kind of setting. Certainly, had the clinical interest been in, say, the difference between therapies after 12 months, Figs. B.1 to B.24 suggest that the results would have been considerably more sensitive to violations of the MAR/NAR assumptions.

It is worth noting that if a true ITT analysis is required, then ceasing to collect data on subjects who deviate from the protocol but who remain in the trial (as was the case for the subjects in the CV outcomes phase in RECORD) is not sensible. Information on these subjects could have been very useful in drawing imputations for those who were completely lost to follow-up and this would have allowed the ITT sensitivity analyses to be more reliable.

We also note that in almost every model considered here, the only noticeable profile differences occurred in the background Met arm. This suggests that it is the presence

or absence of Su (as opposed to Rosi) which has the largest effect. This is echoed in the results of §11.2 where we noted that the largest difference in loss-to-follow-up was between the Met+Su and Su+Met arms. Is it possible that Met and Rosi are similar, with Su being better than both of them? This would explain both the difference in the (early) profiles between Met+Rosi and Met+Su, and might also explain why more people in the Su+Met arm chose to withdraw than in the Met+Su arm, where the add-on treatment was more to the patients' satisfaction. It is well-documented (Charbonnel *et al.*, 2005) that the early profile of Su is steeper than that of Met (or Rosi), and this effect may be the dominating factor in what we see over the first 18 months. Therefore, when it comes to later analyses of the RECORD data (after a longer follow-up), the patterns seen might be quite different. However, the approach described here could be applied in exactly the same way to provide sensitivity analyses for the planned PP-NAR analysis.

Future work could involve combining the work of the previous chapter with this one to carry out sensitivity analyses within the doubly robust framework. Since our methods in both chapters use multiple imputation at their centre, combining the two ideas should be possible. However, for every MNAR model considered here in the pattern-mixture framework, an equivalent selection-model representation would be required in order that the weights be modelled in an analogous way to the outcomes.

Part V

Comparing methods for incomplete longitudinal binary data

12

Motivation and simulation studies

12.1 Introduction

Principled methods for analysing incomplete continuous longitudinal data under the MAR assumption are well-understood and increasingly widely-used in medical studies. However, when the repeated outcome of interest is binary (or, more generally, discrete), the best approach to take is often harder to determine and the relative merits of different methods are less clearly understood. In the literature review, we discussed some

of the additional complexities (such as subject-specific versus population-averaged effects) which arise when analysing longitudinal binary data, and described several of the methods (likelihood methods, GEE, MI-GEE, weighted GEE and the semiparametric-efficient generalisation of GEE) that are advocated for use in this setting. In this chapter, we concentrate on the many population-averaged approaches to analysing incomplete longitudinal binary data.

Many simulation studies in the literature have compared the performance of some or all of these methods; see, for example, Fitzmaurice *et al.* (1995), Fitzmaurice *et al.* (2001), Li *et al.* (2006), Lipsitz *et al.* (2000), Beunckens *et al.* (2008) and Preisser *et al.* (2002). Inconsistencies in some aspects of the results between different simulation studies suggest a need for a more theoretical approach to comparing these methods.

In this chapter, we give some motivation for this work using our own simulation study comparing the methods described in Chapter 7. Then, in Chapter 13, we derive some theoretical results, confirming some findings suggested by our simulation study. We hope that by studying the theoretical properties of the various methods for analysing incomplete binary data, we can present a clearer picture than has been presented to date of the relationships between the available methods and their relative merits.

12.2 Simulation study

12.2.1 Methods

We consider three binary outcome variables, Y_1 , Y_2 and Y_3 , and two covariates X_1 and X_2 . The marginal distributions of Y_1 , Y_2 and Y_3 conditional on X_1 and X_2 are given

by:

$$\begin{aligned}\text{logit} [\mathbb{P} (Y_1 = 1 | X_1, X_2)] &= X_1 - \frac{1}{2} X_1 X_2 \\ \text{logit} [\mathbb{P} (Y_2 = 1 | X_1, X_2)] &= -1 + \frac{1}{4} X_1 + \frac{1}{4} X_2 - X_1 X_2 \\ \text{logit} [\mathbb{P} (Y_3 = 1 | X_1, X_2)] &= -X_1 + \frac{1}{2} X_2 - X_1 X_2\end{aligned}$$

In the first two of our three sets of simulations, X_1 and X_2 are both binary with $\mathbb{P}(X_1) = 0.5$ and $\mathbb{P}(X_2) = 0.25$. In the third scenario, X_1 is binary with $\mathbb{P}(X_1) = 0.5$ and $X_2 \sim U(0, 1)$.

As we require Y_1 , Y_2 and Y_3 to be correlated, we simulate them from a Bahadur distribution (Bahadur, 1961; Molenberghs and Verbeke, 2005). In the first and third sets of simulations, the pairwise correlation matrix of (Y_1, Y_2, Y_3) is

$$\begin{pmatrix} 1 & 0.3 & -0.15 \\ 0.3 & 1 & 0.3 \\ -0.15 & 0.3 & 1 \end{pmatrix}$$

and the higher-order correlation term, ρ_{123} , is given by

$$\rho_{123} = \mathbb{E} \left[\frac{(Y_1 - \mu_1)(Y_2 - \mu_2)(Y_3 - \mu_3)}{\sqrt{\mu_1(1 - \mu_1)\mu_2(1 - \mu_2)\mu_3(1 - \mu_3)}} \right] = -0.1$$

In the second set of simulations, the pairwise correlation matrix of (Y_1, Y_2, Y_3) is

$$\begin{pmatrix} 1 & 0.2 & 0 \\ 0.2 & 1 & 0.2 \\ 0 & 0.2 & 1 \end{pmatrix}$$

and the higher-order correlation term $\rho_{123} = -0.173$.

The missing data pattern is set to be monotone with Y_1 observed on all subjects.

Conditional on Y_1 , X_1 and X_2

$$\text{logit} [\mathbb{P} (R_2 = 1)] = \frac{1}{2}X_1 - \frac{1}{2}X_2 + 3Y_1$$

and conditional on $R_2 = 1$, Y_1 , Y_2 , X_1 and X_2

$$\text{logit} [\mathbb{P} (R_3 = 1 | R_2 = 1)] = -\frac{1}{2} - \frac{1}{2}X_1 + \frac{1}{2}X_2 + X_1Y_1 - Y_1Y_2 + 4Y_2X_1$$

This leads to approximately 20% and 60% missing data on Y_2 and Y_3 respectively. The code for the first set of simulations can be found in Appendix C.3.

A sample size of 500 is used, and 1,000 independent simulated datasets are generated under each of the three scenarios.

The methods to be compared are:

- GEE (with unstructured covariance structure) on the whole dataset;
- and, on the incomplete data:
- IEE (independence structure)
- GEE (unstructured)
- cluster-level weighted GEE (unstructured)
- observation-level weighted GEE (unstructured)
- MI-GEE (unstructured)
- semiparametric-efficient estimator as proposed by Robins and Rotnitzky (1995)
- regression-based doubly robust estimator as proposed by Bang and Robins (2005)
- doubly robust MI-GEE as we proposed in Chapter 9

To assess the robustness of these methods to misspecification of the incomplete data models, in addition to the correct models for the conditional distributions of $(Y_3 | X_1, X_2, Y_1, Y_2)$, $(Y_2 | X_1, X_2, Y_1)$, $(R_3 | X_1, X_2, Y_1, Y_2, R_2 = 1)$ and $(R_2 | X_1, X_2, Y_1)$, the following incorrect models are also defined:

$$\begin{aligned} \text{logit} [\mathbb{P} (Y_3 = 1 | X_1, X_2, Y_1, Y_2)] &= \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2 + \alpha_4 Y_1 \\ &\quad + \alpha_5 X_1 Y_1 + \alpha_6 Y_2 + \alpha_7 X_1 Y_2 \end{aligned}$$

$$\text{logit} [\mathbb{P} (Y_2 = 1 | X_1, X_2, Y_1)] = \beta_0 + \beta_1 X_1 + \beta_2 Y_1 + \beta_3 X_1 Y_1$$

$$\text{logit} [\mathbb{P} (R_3 = 1 | X_1, X_2, Y_1, Y_2, R_2 = 1)] = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_1 Y_1$$

$$\text{logit} [\mathbb{P} (R_2 = 1 | X_1, X_2, Y_1)] = \delta_0 + \delta_1 X_1 + \delta_2 X_2 Y_1$$

In each case, the parameter of interest is taken to be the coefficient of X_1 in the marginal (final timepoint) logistic regression of Y_3 on X_1 , X_2 and $X_1 X_2$. As in Chapter 9, we refer to these conditional outcome and missingness models as the y - and π -models, respectively.

12.2.2 Discussion of results

The results comparing the different methods of estimating our parameter of interest in each of the three sets of simulations are summarised in Tables 12.1–12.3 and further illustrated using kernel density plots in figures 12.1–12.12. In each of the three tables, **bias** refers to the mean bias over all simulations and **true SE** refers to the standard deviation of the parameter estimates over all simulations. **Z-score for bias** is the ratio of the first and second columns and should be used to gauge the comparative severity of the bias from methods with different efficiency. The final column, **# successful simulations**, gives the number of simulations for which the method converged and a parameter estimate was obtained. This is the number of simulations on which the estimates of the previous three columns are based.

As we would fully expect, there is a loss of information due to the missing data, reflected in the fact that the standard error of the parameter estimate from the full data is consistently smaller than the corresponding standard errors from any of the methods applied to the incomplete data. Again, as we would expect, the Z -scores for bias in both the cluster- and observation-level weighted GEE analyses increase when the π -model is misspecified. Correspondingly, the Z -score for bias in the MI-GEE analysis increases when the γ -model is misspecified. Finally, the IEE analysis is biased, suggesting that our chosen MAR mechanism represents a non-negligible departure from MCAR. While these aspects of the results agree with existing theoretical predictions, others require further explanation. We describe these aspects here and investigate them more fully (where necessary) in the next chapter.

12.2.2.1 Convergence problems with the Robins and Rotnitzky (1995) estimator

One noticeable feature of the results is that in as many as 30% of our simulations, the Robins and Rotnitzky estimator fails to converge. The iterative Fisher scoring algorithm was set to a tolerance of 10^{-5} . This means that for convergence, the absolute difference between successive parameter estimates has to be less than 10^{-5} for all parameters. To ensure a manageable computational time for the simulations, the number of iterations was limited to 100. Inspection of the parameter estimates after 100 iterations for the simulations in which the method failed to converge suggests that given further iterations the method would have converged; this is inferred from the fact that the parameter estimates after 100 iterations were—in most cases—close to the correct values, and that the cases in which the algorithm had clearly diverged were few in number.

It is possible that by updating our estimate of Λ (step 9. of the algorithm described in 7.6.5) with every iteration of the Fisher scoring algorithm as opposed to using the original fixed estimate of Λ , the convergence rate could have been improved, but the additional computational time involved in re-calculating Λ would far outweigh the time saved. In the first simulation study, the mean value of the Bang and Robins (BR)

estimate $\hat{\beta}_{\text{BR}}$ when the Robins and Rotnitzky (RR) estimator failed to converge is -0.875 , which is 0.125 higher than the true parameter value of -1 . This could explain the downward bias seen in $\hat{\beta}_{\text{RR}}$ in Table 12.1.

12.2.2.2 Superiority of doubly robust MI over the other doubly robust procedures

In contrast with the results of Chapter 9, in this simulation study doubly robust MI consistently outperforms the Bang and Robins (2005) estimator with respect to both bias and precision (see Figs. 12.22–12.24). This is a consequence of the fact that the data are binary and the different ways in which the conditional outcome distributions are computed in the two approaches in the case of non-Gaussian data. This is discussed in greater detail in the next chapter.

12.2.2.3 Lack of bias in unweighted GEE

As we have already mentioned, IEE is biased as theory predicts when the mechanism is MAR. Theory also predicts (Liang and Zeger, 1986) that unweighted GEE be biased under MAR. In our simulations, however, this bias is much smaller for GEE compared with IEE, with the bias being particularly small in the second simulation study (see Figs. 12.13–12.15). This feature is explained by the theoretical work in the next chapter.

12.2.2.4 Differences between cluster- and observation-level weighting

Observation-level weighted GEE appears to be more efficient than cluster-level weighted GEE, but the difference is very small in the first two simulation studies, where the means model is saturated, and more pronounced in the third simulation study, where

the means model is not saturated (see Figs. 12.16–12.18). This will be explored in the next chapter.

12.2.2.5 Imputation versus weighting

Comparing imputation and weighting has been the subject of many recent papers in the literature e.g. Carpenter *et al.* (2006); Wang *et al.* (2007); Beunckens *et al.* (2008). Figs. 12.19–12.21 show the comparison between MI-GEE and observation-level weighted GEE in our simulations. MI-GEE is more efficient, with the difference again more pronounced for the non-saturated means model. In the next chapter we study this comparison in more detail.

12.2.2.6 Doubly robust methods with both models misspecified

The final feature of our simulation results is that when both the y - and π -models are misspecified, the bias in the doubly robust estimators remains quite small. This is an artifact of the particular incorrect models chosen, where the bias created by the former is effectively cancelled out by the bias (in the opposite direction) due to the latter.

Estimator	Bias	True SE	Z-score for bias	# successful simulations
$\hat{\beta}_{full}$	-0.078	0.229	-0.340	1000
$\hat{\beta}_{IEE}$	0.330	0.391	0.845	998
$\hat{\beta}_{GEE}$	-0.164	0.395	-0.414	998
$\hat{\beta}_{CWGEE}$	-0.120	0.434	-0.276	998
$\hat{\beta}_{OWGEE}$	-0.100	0.428	-0.232	996
$\hat{\beta}_{MI-GEE}$	-0.051	0.402	-0.128	998
$\hat{\beta}_{RR}$	-0.136	0.435	-0.312	748
$\hat{\beta}_{BR}$	0.054	0.467	0.116	1000
$\hat{\beta}_{DRMI}$	-0.019	0.396	-0.049	998
$\hat{\beta}_{CWGEE \cdot \pi - false}$	-0.292	0.421	-0.693	996
$\hat{\beta}_{OWGEE \cdot \pi - false}$	0.348	0.396	0.879	997
$\hat{\beta}_{RR \cdot \pi - false}$	-0.199	0.482	-0.413	967
$\hat{\beta}_{BR \cdot \pi - false}$	0.047	0.470	0.100	1000
$\hat{\beta}_{DRMI \cdot \pi - false}$	-0.038	0.402	-0.094	998
$\hat{\beta}_{MI-GEE \cdot y - false}$	-0.171	0.385	-0.446	998
$\hat{\beta}_{RR \cdot y - false}$	-0.118	0.439	-0.269	682
$\hat{\beta}_{BR \cdot y - false}$	-0.141	0.437	-0.322	992
$\hat{\beta}_{DRMI \cdot y - false}$	-0.041	0.402	-0.101	998
$\hat{\beta}_{RR \cdot \pi \oplus y - false}$	-0.103	0.442	-0.232	965
$\hat{\beta}_{BR \cdot \pi \oplus y - false}$	-0.131	0.410	-0.318	1000
$\hat{\beta}_{DRMI \cdot \pi \oplus y - false}$	-0.069	0.426	-0.163	998

Table 12.1: The results of the first longitudinal binary simulation study, where the means model is saturated. In each case, β refers to the log odds ratio for X_1 at the third timepoint. The abbreviations used are: CWGEE (cluster-level weighted GEE), OWGEE (observation-level weighted GEE), RR (method proposed by Robins and Rotnitzky (1995)), BR (method proposed by Bang and Robins (2005)) and DRMI (doubly robust MI). No subscript indicates correct specification of the relevant model(s). $\pi - false$ indicates that the estimator used an incorrectly-specified π -model, $y - false$ indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y - false$ indicates that both the π - and y -models were incorrectly specified.

Estimator	Bias	True SE	Z-score for bias	# successful simulations
$\hat{\beta}_{\text{full}}$	-0.045	0.222	-0.202	1000
$\hat{\beta}_{\text{IEE}}$	0.197	0.384	0.514	998
$\hat{\beta}_{\text{GEE}}$	-0.077	0.398	-0.194	998
$\hat{\beta}_{\text{CWGEE}}$	-0.075	0.426	-0.177	998
$\hat{\beta}_{\text{OWGEE}}$	-0.063	0.421	-0.151	996
$\hat{\beta}_{\text{MI-GEE}}$	-0.005	0.390	-0.013	998
$\hat{\beta}_{\text{RR}}$	-0.082	0.429	-0.192	762
$\hat{\beta}_{\text{BR}}$	-0.036	0.423	-0.085	1000
$\hat{\beta}_{\text{DRMI}}$	0.023	0.395	0.057	998
$\hat{\beta}_{\text{CWGEE}\cdot\pi\text{-false}}$	-0.180	0.417	-0.432	998
$\hat{\beta}_{\text{OWGEE}\cdot\pi\text{-false}}$	0.194	0.386	0.502	996
$\hat{\beta}_{\text{RR}\cdot\pi\text{-false}}$	-0.199	0.501	-0.398	979
$\hat{\beta}_{\text{BR}\cdot\pi\text{-false}}$	-0.039	0.422	-0.092	1000
$\hat{\beta}_{\text{DRMI}\cdot\pi\text{-false}}$	0.000	0.402	0.001	998
$\hat{\beta}_{\text{MI-GEE}\cdot y\text{-false}}$	-0.112	0.388	-0.288	998
$\hat{\beta}_{\text{RR}\cdot y\text{-false}}$	-0.088	0.434	-0.203	777
$\hat{\beta}_{\text{BR}\cdot y\text{-false}}$	-0.161	0.404	-0.399	989
$\hat{\beta}_{\text{DRMI}\cdot y\text{-false}}$	-0.011	0.398	-0.026	998
$\hat{\beta}_{\text{RR}\cdot\pi\oplus y\text{-false}}$	-0.104	0.448	-0.233	979
$\hat{\beta}_{\text{BR}\cdot\pi\oplus y\text{-false}}$	-0.103	0.389	-0.264	1000
$\hat{\beta}_{\text{DRMI}\cdot\pi\oplus y\text{-false}}$	0.001	0.420	0.002	998

Table 12.2: The results of the second longitudinal binary simulation study, where the means model is saturated but the correlation structure is different from the one used in the first set of simulations. In each case, β refers to the log odds ratio for X_1 at the third timepoint. The abbreviations used are: CWGEE (cluster-level weighted GEE), OWGEE (observation-level weighted GEE), RR (method proposed by Robins and Rotnitzky (1995)), BR (method proposed by Bang and Robins (2005)) and DRMI (doubly robust MI). No subscript indicates correct specification of the relevant model(s). π - false indicates that the estimator used an incorrectly-specified π -model, y - false indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y$ - false indicates that both the π - and y -models were incorrectly specified.

Estimator	Bias	True SE	Z-score for bias	# successful simulations
$\hat{\beta}_{\text{full}}$	-0.049	0.386	-0.127	1000
$\hat{\beta}_{\text{IEE}}$	0.338	0.672	0.503	1000
$\hat{\beta}_{\text{GEE}}$	-0.174	0.662	-0.263	1000
$\hat{\beta}_{\text{CWGEE}}$	-0.137	0.832	-0.165	1000
$\hat{\beta}_{\text{OWGEE}}$	-0.091	0.785	-0.116	1000
$\hat{\beta}_{\text{MI-GEE}}$	-0.065	0.651	-0.100	1000
$\hat{\beta}_{\text{RR}}$	-0.079	0.813	-0.097	691
$\hat{\beta}_{\text{BR}}$	0.093	0.802	0.116	993
$\hat{\beta}_{\text{DRMI}}$	-0.011	0.673	-0.016	1000
$\hat{\beta}_{\text{CWGEE}\cdot\pi\text{-false}}$	-0.299	0.740	-0.404	987
$\hat{\beta}_{\text{OWGEE}\cdot\pi\text{-false}}$	0.332	0.710	0.468	1000
$\hat{\beta}_{\text{RR}\cdot\pi\text{-false}}$	-0.162	0.873	-0.185	910
$\hat{\beta}_{\text{BR}\cdot\pi\text{-false}}$	0.065	0.747	0.087	998
$\hat{\beta}_{\text{DRMI}\cdot\pi\text{-false}}$	-0.050	0.654	-0.076	1000
$\hat{\beta}_{\text{MI-GEE}\cdot y\text{-false}}$	-0.256	0.511	-0.501	1000
$\hat{\beta}_{\text{RR}\cdot y\text{-false}}$	-0.087	0.795	-0.110	623
$\hat{\beta}_{\text{BR}\cdot y\text{-false}}$	-0.268	0.636	-0.421	996
$\hat{\beta}_{\text{DRMI}\cdot y\text{-false}}$	-0.037	0.666	-0.056	1000
$\hat{\beta}_{\text{RR}\cdot\pi\oplus y\text{-false}}$	-0.083	0.869	-0.095	887
$\hat{\beta}_{\text{BR}\cdot\pi\oplus y\text{-false}}$	-0.293	0.488	-0.601	1000
$\hat{\beta}_{\text{DRMI}\cdot\pi\oplus y\text{-false}}$	-0.085	0.677	-0.126	1000

Table 12.3: The results of the third longitudinal binary simulation study, where the means model is *not* saturated. In each case, β refers to the log odds ratio for X_1 at the third timepoint. The abbreviations used are: CWGEE (cluster-level weighted GEE), OWGEE (observation-level weighted GEE), RR (method proposed by Robins and Rotnitzky (1995)), BR (method proposed by Bang and Robins (2005)) and DRMI (doubly robust MI). No subscript indicates correct specification of the relevant model(s). π - false indicates that the estimator used an incorrectly-specified π -model, y - false indicates that the estimator used an incorrectly-specified y -model and $\pi \oplus y$ - false indicates that both the π - and y -models were incorrectly specified.

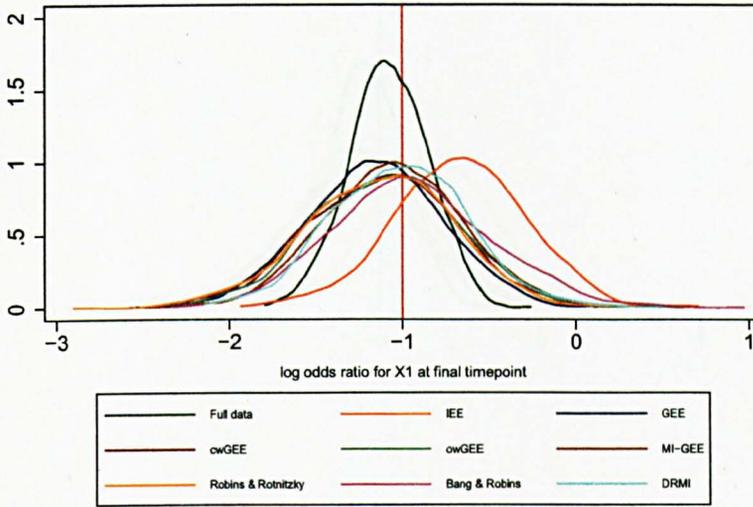


Figure 12.1: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the first set of simulations with both models correctly specified.

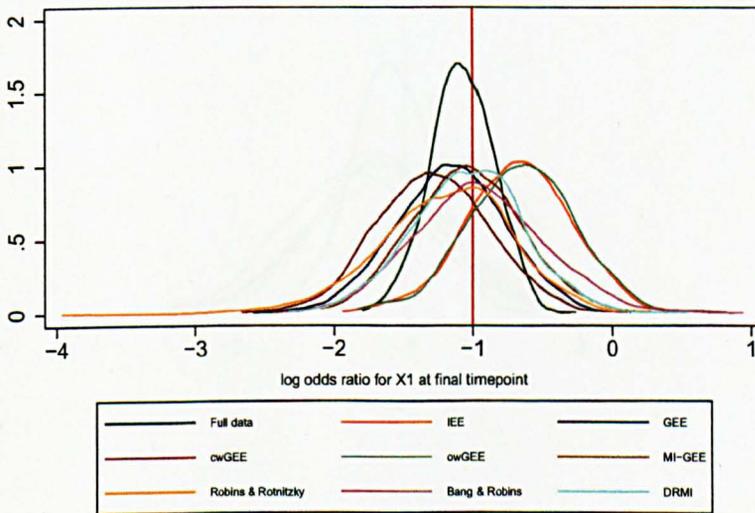


Figure 12.2: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the first set of simulations with the π -model incorrectly specified.

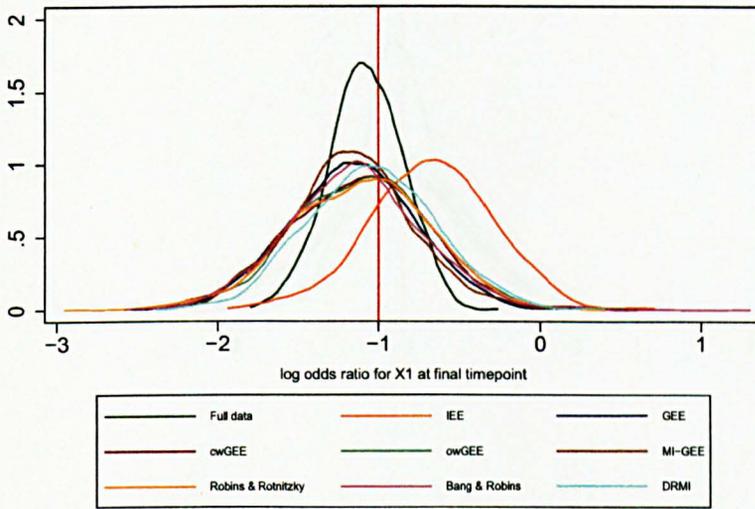


Figure 12.3: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the first set of simulations with the y -model incorrectly specified.

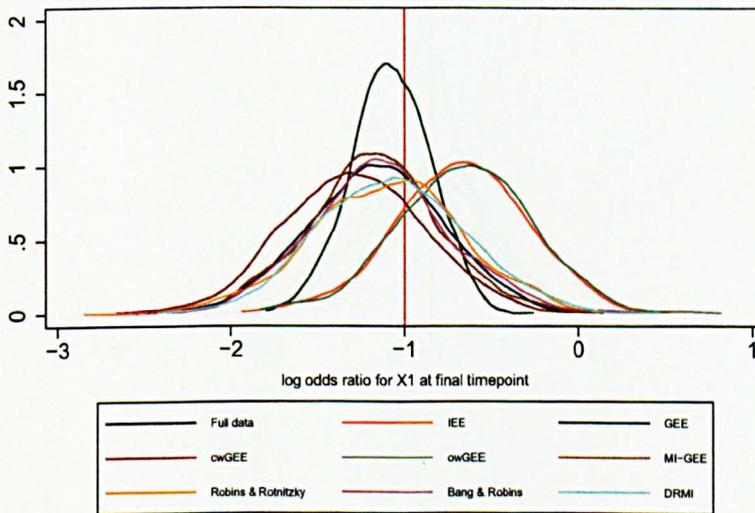


Figure 12.4: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the first set of simulations with both models incorrectly specified.

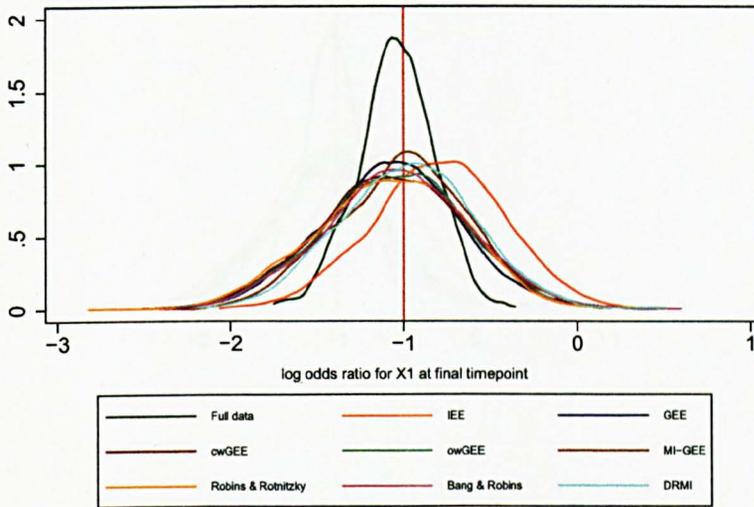


Figure 12.5: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the second set of simulations with both models correctly specified.

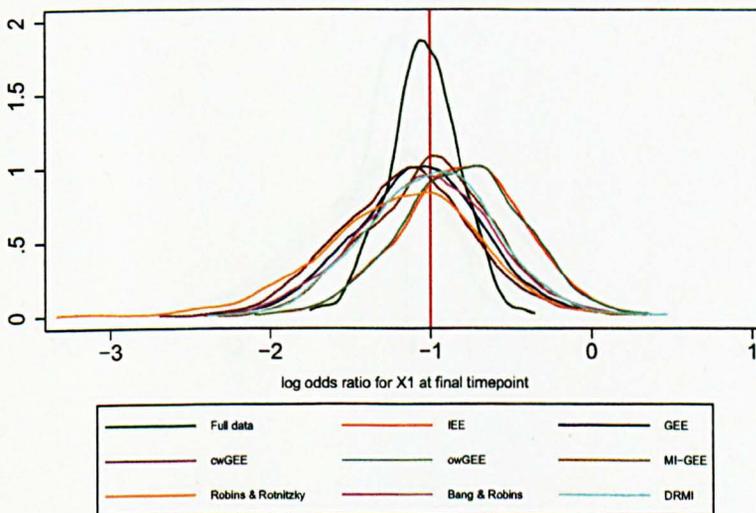


Figure 12.6: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the second set of simulations with the π -model incorrectly specified.

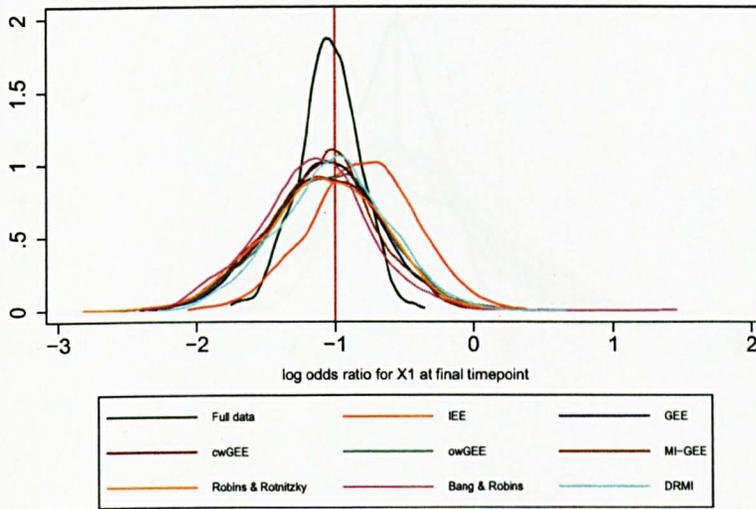


Figure 12.7: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the second set of simulations with the y -model incorrectly specified.

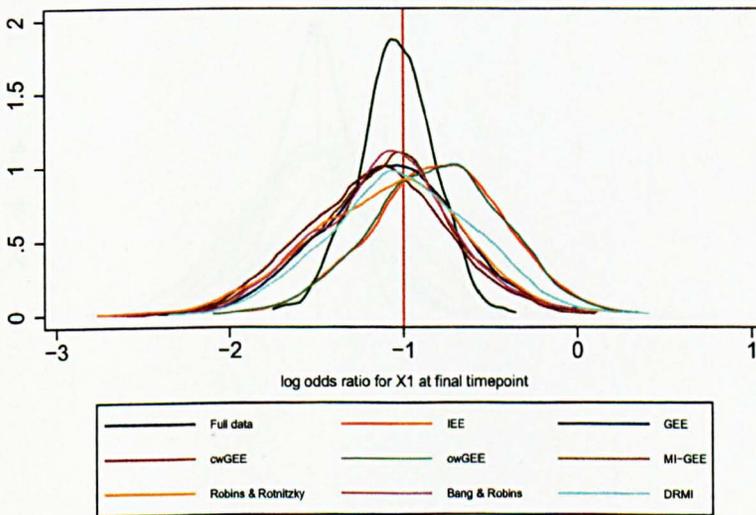


Figure 12.8: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the second set of simulations with both models incorrectly specified.

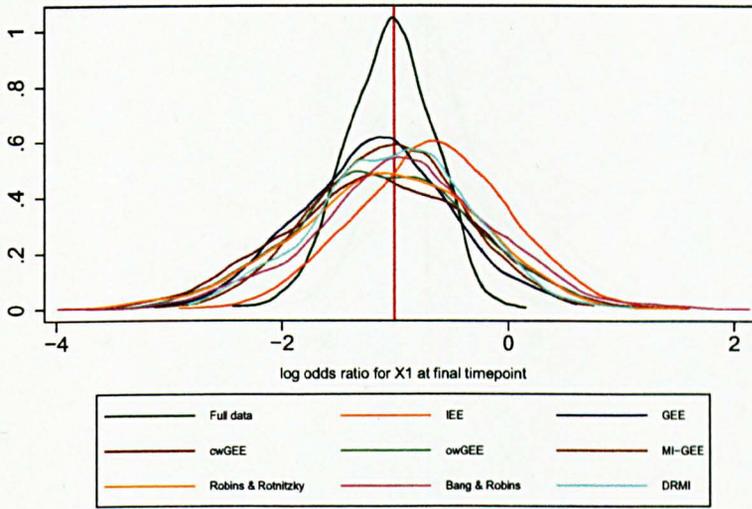


Figure 12.9: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the third set of simulations with both models correctly specified.

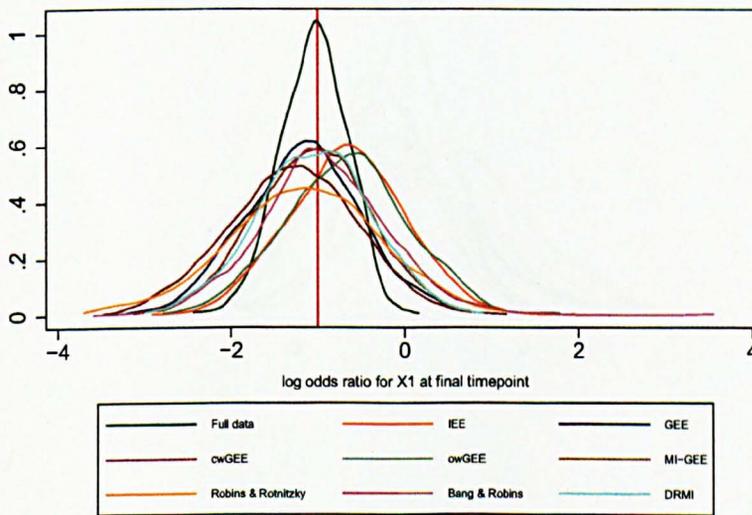


Figure 12.10: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the third set of simulations with the π -model incorrectly specified.

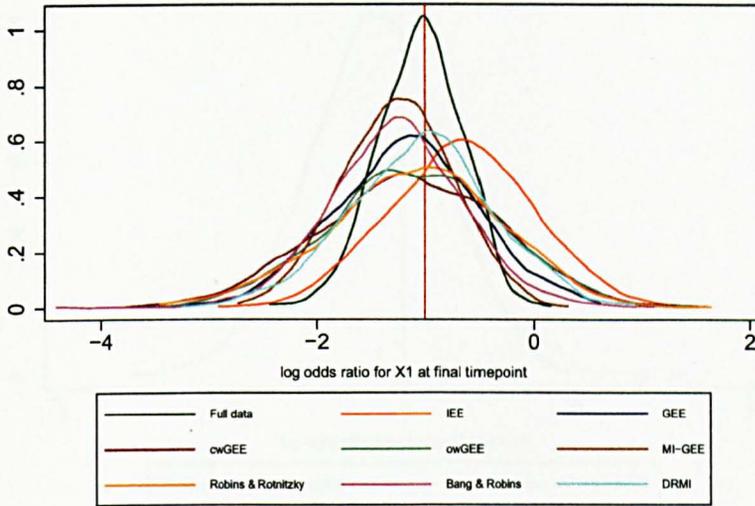


Figure 12.11: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the third set of simulations with the y -model incorrectly specified.

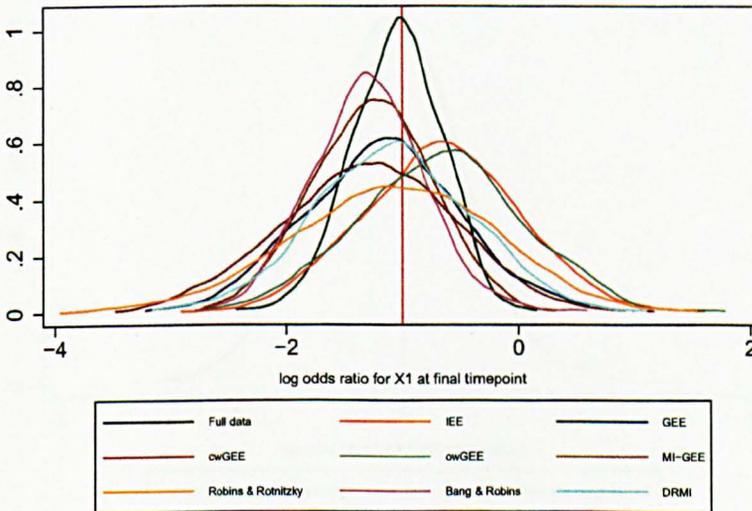


Figure 12.12: Kernel density plots for the sampling distributions of 9 different estimators of the log odds ratio for X_1 at the final timepoint. These estimates are from the third set of simulations with both models incorrectly specified.

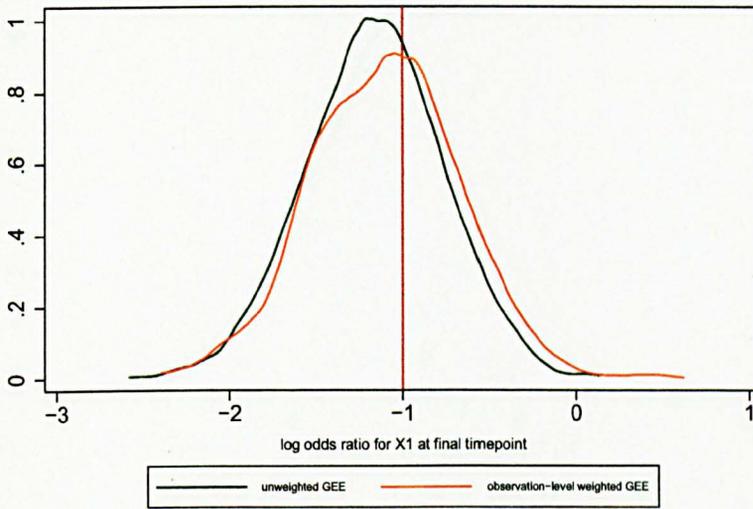


Figure 12.13: Kernel density plots comparing unweighted and observation-level weighted GEE. These estimates are from the first set of simulations with both models correctly specified.

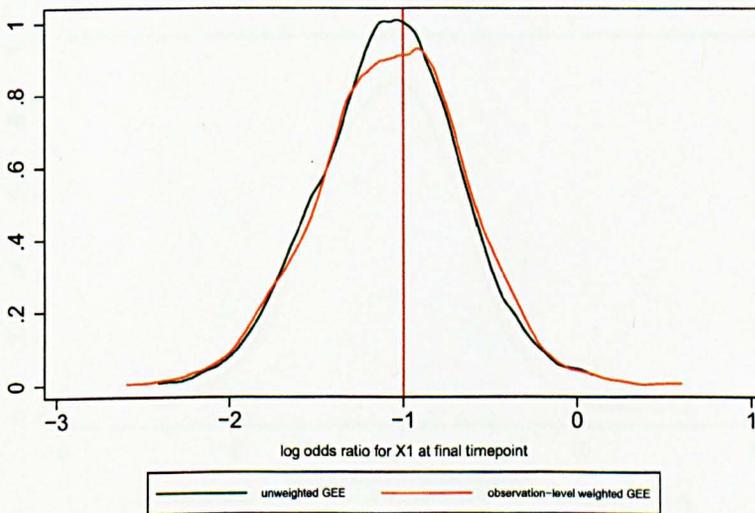


Figure 12.14: Kernel density plots comparing unweighted and observation-level weighted GEE. These estimates are from the second set of simulations with both models correctly specified.

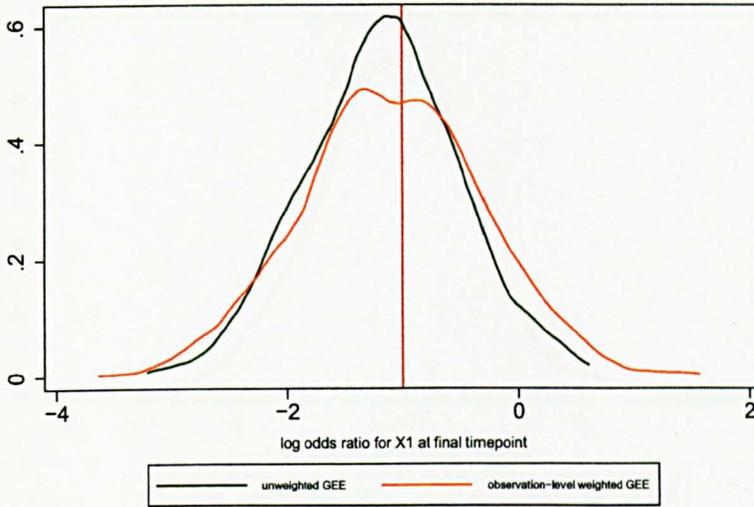


Figure 12.15: Kernel density plots comparing unweighted and observation-level weighted GEE. These estimates are from the third set of simulations with both models correctly specified.

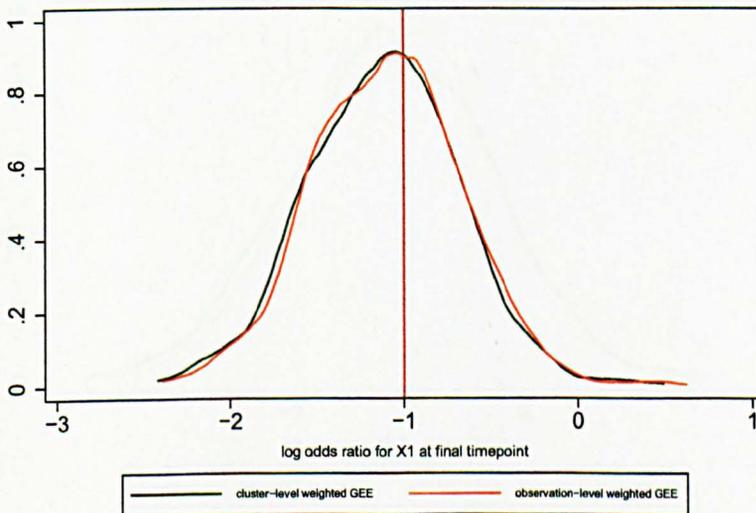


Figure 12.16: Kernel density plots comparing cluster- and observation-level weighted GEE. These estimates are from the first set of simulations with both models correctly specified.

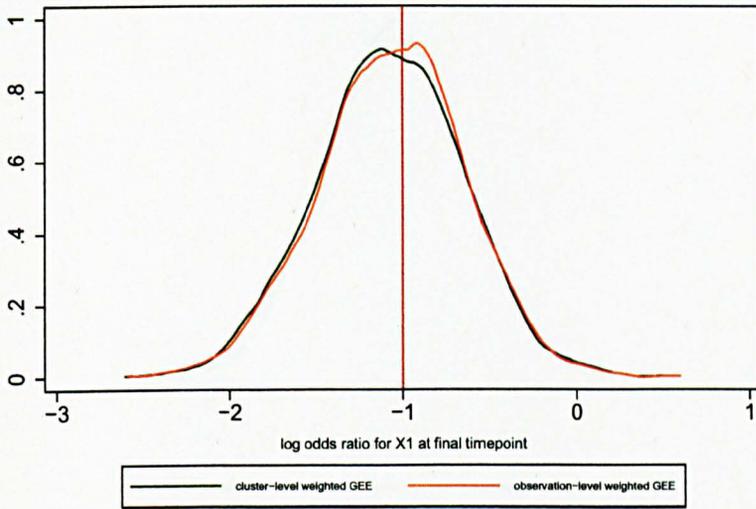


Figure 12.17: Kernel density plots comparing cluster- and observation-level weighted GEE. These estimates are from the second set of simulations with both models correctly specified.

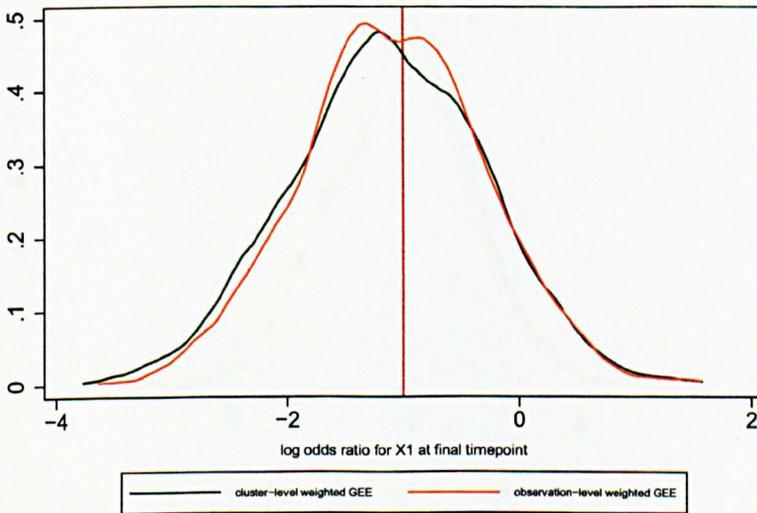


Figure 12.18: Kernel density plots comparing cluster- and observation-level weighted GEE. These estimates are from the third set of simulations with both models correctly specified.

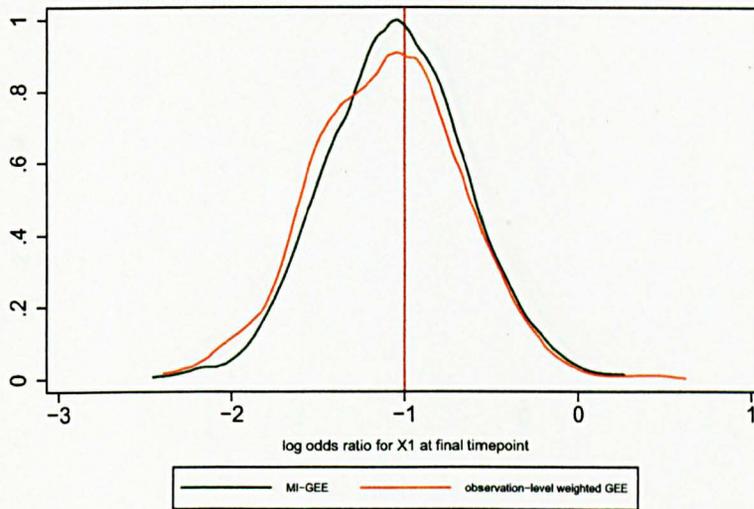


Figure 12.19: Kernel density plots comparing MI-GEE and observation-level weighted GEE. These estimates are from the first set of simulations with both models correctly specified.

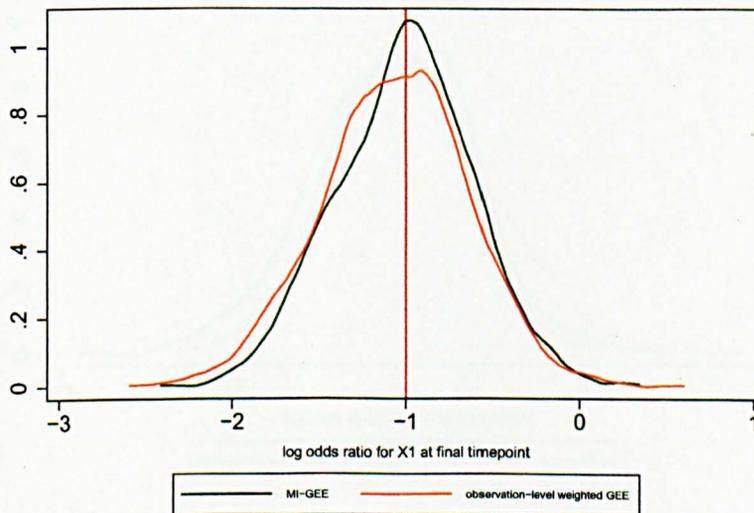


Figure 12.20: Kernel density plots comparing MI-GEE and observation-level weighted GEE. These estimates are from the second set of simulations with both models correctly specified.

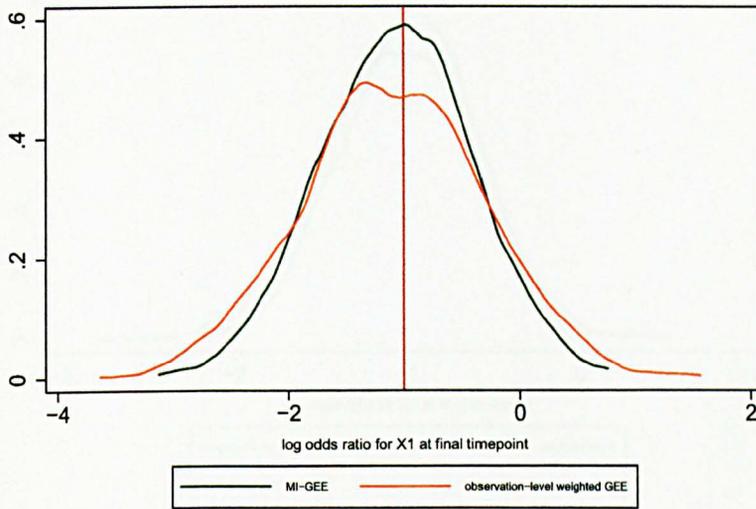


Figure 12.21: Kernel density plots comparing MI-GEE and observation-level weighted GEE. These estimates are from the third set of simulations with both models correctly specified.

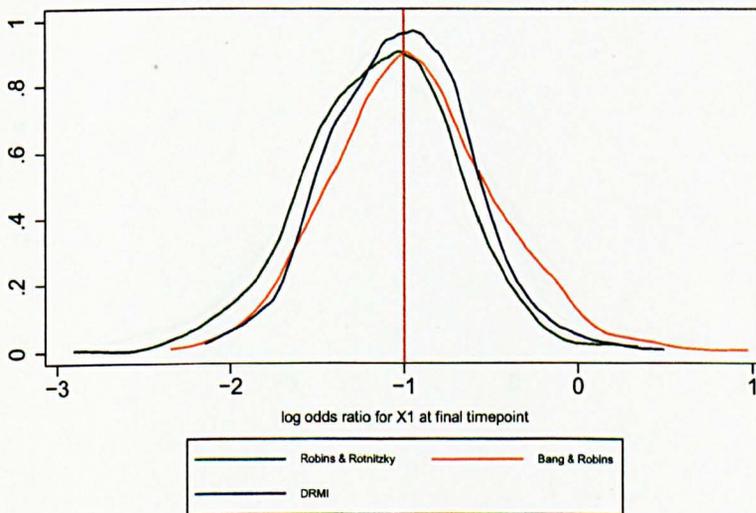


Figure 12.22: Kernel density plots comparing the three doubly robust procedures. These estimates are from the first set of simulations with both models correctly specified.

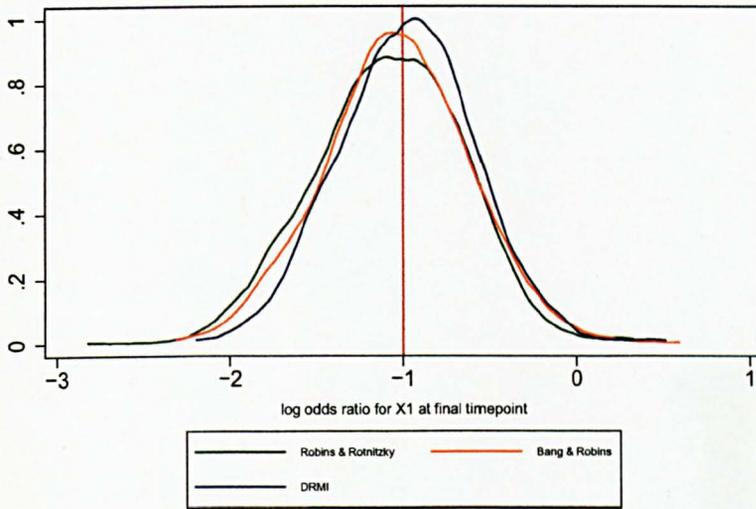


Figure 12.23: Kernel density plots comparing the three doubly robust procedures. These estimates are from the second set of simulations with both models correctly specified.

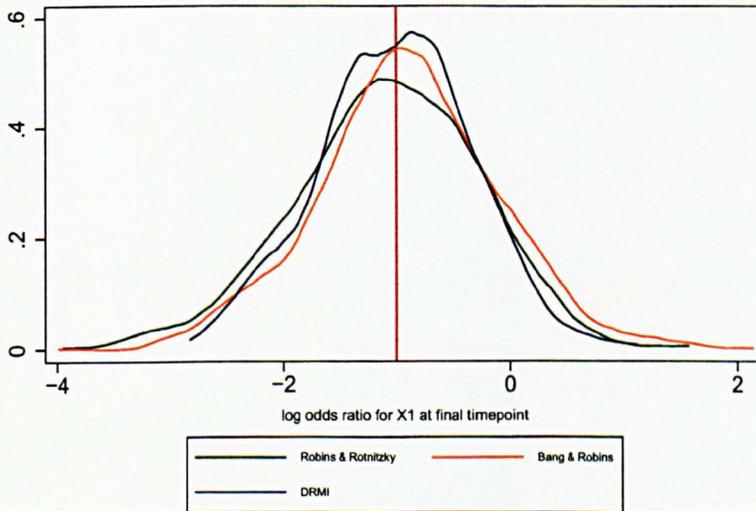


Figure 12.24: Kernel density plots comparing the three doubly robust procedures. These estimates are from the third set of simulations with both models correctly specified.

13

Theoretical comparison of GEE and related methods

13.1 Aims and outline

In this chapter we bring together different strands of theoretical work appearing in the literature on marginal models for repeated binary data, and derive some further theoretical results of our own, with the aim of presenting a clearer picture of the rela-

tionships between the available methods and their relative merits. We are motivated throughout by the results of the simulation studies described in the previous chapter.

We start, in §13.2, with a result which underpins much of the following results, namely the numerical equivalence of observation-level weighted GEE, augmented observation-level weighted GEE and nonparametric mean quasi score imputation under certain special conditions. This extends the work done by Wang *et al.* (2007). In §13.3 we derive conditions under which unweighted GEE is consistent and semiparametric-efficient under MAR, extending results given in Robins and Rotnitzky (1995) and Lipsitz *et al.* (2000). In §13.4 we show that observation-level weighting is always at least as good as cluster-level weighting, and in §13.5 we argue that MI-GEE is approximately equivalent to observation-level weighted GEE when the means model is saturated. Finally, the difference between doubly robust (DR) MI and the other DR procedures is explained in §13.6, justifying our preference for the former over the latter for non-Gaussian data.

13.2 Conditions under which observation-level weighted GEE, augmented observation-level weighted GEE and nonparametric mean quasi score imputation are numerically equivalent

The situation considered by Wang *et al.* (2007) is one in which Y is the (univariate) outcome variable for the regression analysis of interest, \mathbf{X} are partially observed covariates, \mathbf{Z} are always-observable covariates, and \mathbf{W} are observable surrogates for \mathbf{X} , with Y and \mathbf{W} conditionally independent given (\mathbf{X}, \mathbf{Z}) . They show that in this situation, nonparametric mean score imputation, IPWCC and AIPW are numerically equivalent. By following their argument closely, we show a similar result for the longitudinal binary case. More explicitly, we prove the following theorem.

Theorem 13.1 (Numerical equivalence of OWGEE, AIPW and nonparametric mean quasi-score imputation). *Using the notation of §7.4 we assume that all covariates*

are discrete, that the means model is saturated and that the missing data mechanism is monotone and MAR. Under these conditions, OWGEE, AIPW and nonparametric mean quasi-score imputation, where the quasi-score function $\mathbf{Q}(\mathbf{X}_i, \mathbf{Y}_i, \boldsymbol{\mu}) = \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)$ replaces $\mathbf{S}_{\boldsymbol{\theta}}^F(\mathbf{Z}_i, \boldsymbol{\theta})$ in (7.2.3), give numerically identical results, provided that the weights in OWGEE and AIPW are also estimated nonparametrically.

Proof. We consider each covariate combination separately and the OWGEE equation (7.4.2) simplifies to

$$\sum_{i=1}^n \Phi_i \left(\tilde{\mathbf{Y}}_i - \tilde{\boldsymbol{\mu}}_i \right) = \mathbf{0} \tag{13.2.1}$$

applied to distinct subsets of the data corresponding to each covariate combination. Without loss of generality, we continue as if there were only one covariate combination without explicitly stating at each stage that we are dealing with the distinct subsets of the data separately.

(13.2.1) can be rewritten as

$$\sum_{i=1}^n \frac{R_{t,i}}{\hat{\pi}_{t,i}} (Y_{t,i} - \mu_t) = 0 \quad \forall t \tag{13.2.2}$$

Also, since $\pi_{t,i}$ is estimated nonparametrically, we can write

$$\hat{\pi}_{t,i} = \frac{\sum_{j=1}^n R_{t,j} \mathbb{1}(\bar{\mathbf{Y}}_{t-1,j} = \bar{\mathbf{Y}}_{t-1,i})}{\sum_{k=1}^n \mathbb{1}(\bar{\mathbf{Y}}_{t-1,k} = \bar{\mathbf{Y}}_{t-1,i})} \hat{\pi}_{t-1,i}$$

which in turn means that the LHS of (13.2.2) can be rewritten as

$$\begin{aligned} \sum_{i=1}^n \frac{R_{t,i} \sum_{k=1}^n \mathbb{1}(\bar{\mathbf{Y}}_{t-1,k} = \bar{\mathbf{Y}}_{t-1,i})}{\hat{\pi}_{t-1,i} \sum_{j=1}^n R_{t,j} \mathbb{1}(\bar{\mathbf{Y}}_{t-1,j} = \bar{\mathbf{Y}}_{t-1,i})} (Y_{t,i} - \mu_t) \\ = \sum_{i=1}^n \sum_{k=1}^n \frac{R_{t,i} \mathbb{1}(\bar{\mathbf{Y}}_{t-1,k} = \bar{\mathbf{Y}}_{t-1,i})}{\hat{\pi}_{t-1,i} \sum_{j=1}^n R_{t,j} \mathbb{1}(\bar{\mathbf{Y}}_{t-1,j} = \bar{\mathbf{Y}}_{t-1,i})} (Y_{t,i} - \mu_t) \end{aligned}$$

$$\begin{aligned}
 &= \sum_{k=1}^n \frac{\sum_{i=1}^n R_{t,i} \mathbf{1}(\bar{Y}_{t-1,k} = \bar{Y}_{t-1,i}) (Y_{t,i} - \mu_t)}{\hat{\pi}_{t-1,k} \sum_{j=1}^n R_{t,j} \mathbf{1}(\bar{Y}_{t-1,j} = \bar{Y}_{t-1,k})} \\
 &= \sum_{k=1}^n \frac{R_{t-1,k} \hat{\mathbb{E}}(Q_{t,k} | R_{t-1,k} = 1, \bar{Y}_{t-1,k})}{\hat{\pi}_{t-1,k}}
 \end{aligned}$$

where $Q_{t,i} = Q_t(\mathbf{X}_i, \mathbf{Y}_i, \boldsymbol{\mu})$ is the t^{th} element of $\mathbf{Q}(\mathbf{X}_i, \mathbf{Y}_i, \boldsymbol{\mu})$ and $\hat{\mathbb{E}}(\cdot)$ is the nonparametric estimator of $\mathbb{E}(\cdot)$.

Thus we have shown that

$$\sum_{k=1}^n \frac{R_{t,k} Q_{t,k}}{\hat{\pi}_{t,k}} = \sum_{k=1}^n \frac{R_{t-1,k} \hat{\mathbb{E}}(Q_{t,k} | R_{t-1,k} = 1, \bar{Y}_{t-1,k})}{\hat{\pi}_{t-1,k}} \tag{13.2.3}$$

We can iterate this to show that

$$\begin{aligned}
 \sum_{k=1}^n \frac{R_{t,k} Q_{t,k}}{\hat{\pi}_{t,k}} &= \sum_{k=1}^n \frac{R_{t-1,k} \hat{\mathbb{E}}(Q_{t,k} | R_{t-1,k} = 1, \bar{Y}_{t-1,k})}{\hat{\pi}_{t-1,k}} = \dots \\
 &\dots = \sum_{k=1}^n \frac{R_{1,k} \hat{\mathbb{E}}(Q_{t,k} | R_{1,k} = 1, Y_{1,k})}{\hat{\pi}_{1,k}} = \sum_{k=1}^n \hat{\mathbb{E}}(Q_{t,k}) \tag{13.2.4}
 \end{aligned}$$

where at each stage $\hat{\mathbb{E}}(\cdot)$ is defined sequentially and nonparametrically in the sense that $\hat{\mathbb{E}}(Q_{t,k} | R_{l,k} = 1, \bar{Y}_{l,k})$ is the nonparametric estimator of

$$\mathbb{E} \left[\hat{\mathbb{E}}(Q_{t,k} | R_{l+1,k} = 1, \bar{Y}_{l+1,k}) | R_{l,k} = 1, \bar{Y}_{l,k} \right]$$

Thus observation-level weighted GEE is equivalent to solving

$$\sum_{k=1}^n \hat{\mathbb{E}}(Q_{t,k}) = 0 \quad \forall t \tag{13.2.5}$$

Using again the fact that the means model is saturated and using the same sequential

nonparametric estimator of $\hat{\mathbb{E}}(Q_{t,k} | R_{l,1} = 1, \bar{Y}_{l,k})$, the sequential nonparametric mean quasi-score imputation estimating equation is equivalent to

$$\sum_{i=1}^n \left[R_{t,i} Q_{t,i} + (1 - R_{t,i}) \hat{\mathbb{E}}(Q_{t,i} | \mathbf{R}_i, \mathbf{Y}_i^{\text{obs}}) \right] = 0 \quad \forall t \quad (13.2.6)$$

Since $Q_{t,i} = \hat{\mathbb{E}}(Q_{t,i} | \mathbf{R}_i, \mathbf{Y}_i^{\text{obs}})$ when each component of $\bar{Y}_{t,i}$ is observed, $R_{t,i} Q_{t,i} = R_{t,i} \hat{\mathbb{E}}(Q_{t,i} | \mathbf{R}_i, \mathbf{Y}_i^{\text{obs}})$ and therefore (13.2.6) can be simplified to

$$\sum_{i=1}^n \hat{\mathbb{E}}(Q_{t,i} | \mathbf{R}_i, \mathbf{Y}_i^{\text{obs}}) = 0 \quad \forall t \quad (13.2.7)$$

We can write $\hat{\mathbb{E}}(Q_{t,i} | \mathbf{R}_i, \mathbf{Y}_i^{\text{obs}})$ as

$$\begin{aligned} \hat{\mathbb{E}}(Q_{t,i} | \mathbf{R}_i, \mathbf{Y}_i^{\text{obs}}) &= R_{t,i} Q_{t,i} + R_{t-1,i} (1 - R_{t,i}) \hat{\mathbb{E}}(Q_{t,i} | R_{t-1,i} = 1, \bar{Y}_{t-1,i}) + \cdots \\ &\quad \cdots + R_{1,i} (1 - R_{2,i}) \hat{\mathbb{E}}(Q_{t,i} | R_{1,i} = 1, Y_{1,i}) + (1 - R_{1,i}) \hat{\mathbb{E}}(Q_{t,i}) \\ &= R_{t,i} \left[\hat{\mathbb{E}}(Q_{t,i} | R_{t,i} = 1, \bar{Y}_{t,i}) - \hat{\mathbb{E}}(Q_{t,i} | R_{t-1,i} = 1, \bar{Y}_{t-1,i}) \right] \\ &\quad + R_{t-1,i} \left[\hat{\mathbb{E}}(Q_{t,i} | R_{t-1,i} = 1, \bar{Y}_{t-1,i}) - \hat{\mathbb{E}}(Q_{t,i} | R_{t-2,i} = 1, \bar{Y}_{t-2,i}) \right] \\ &\quad + \cdots + R_{1,i} \left[\hat{\mathbb{E}}(Q_{t,i} | R_{1,i} = 1, Y_{1,i}) - \hat{\mathbb{E}}(Q_{t,i}) \right] + \hat{\mathbb{E}}(Q_{t,i}) \end{aligned}$$

But due to our sequential definition of the nonparametric mean quasi-score, we can show that $\sum_{i=1}^n R_{l,i} \left[\hat{\mathbb{E}}(Q_{t,i} | R_{l,i} = 1, \bar{Y}_{l,i}) - \hat{\mathbb{E}}(Q_{t,i} | R_{l-1,i} = 1, \bar{Y}_{l-1,i}) \right] = 0$ for all l between 1 and t , as follows:

$$\begin{aligned} \sum_{i=1}^n R_{l,i} \hat{\mathbb{E}}(Q_{t,i} | R_{l-1,i} = 1, \bar{Y}_{l-1,i}) \\ = \sum_{i=1}^n R_{l,i} \frac{\sum_{j=1}^n R_{l,j} \mathbf{1}(\bar{Y}_{l-1,j} = \bar{Y}_{l-1,i}) \hat{\mathbb{E}}(Q_{t,i} | \bar{Y}_{l,i})}{\sum_{k=1}^n R_{l,k} \mathbf{1}(\bar{Y}_{l-1,k} = \bar{Y}_{l-1,i})} \end{aligned}$$

$$\begin{aligned}
 &= \sum_{i=1}^n \sum_{j=1}^n \frac{R_{l,i} R_{l,j} \mathbb{1}(\bar{Y}_{l-1,j} = \bar{Y}_{l-1,i}) \hat{\mathbb{E}}(Q_{t,j} | R_{l,j} = 1, \bar{Y}_{l,j})}{\sum_{k=1}^n R_{l,k} \mathbb{1}(\bar{Y}_{l-1,k} = \bar{Y}_{l-1,i})} \\
 &= \sum_{j=1}^n R_{l,j} \hat{\mathbb{E}}(Q_{t,j} | R_{l,j} = 1, \bar{Y}_{l,j}) \frac{\sum_{i=1}^n R_{l,i} \mathbb{1}(\bar{Y}_{l-1,j} = \bar{Y}_{l-1,i})}{\sum_{k=1}^n R_{l,k} \mathbb{1}(\bar{Y}_{l-1,k} = \bar{Y}_{l-1,i})} \\
 &= \sum_{j=1}^n R_{l,j} \hat{\mathbb{E}}(Q_{t,j} | R_{l,j} = 1, \bar{Y}_{l,j}) \\
 &\Rightarrow \sum_{i=1}^n R_{l,i} \left[\hat{\mathbb{E}}(Q_{t,i} | R_{l,i} = 1, \bar{Y}_{l,i}) - \hat{\mathbb{E}}(Q_{t,i} | R_{l-1,i} = 1, \bar{Y}_{l-1,i}) \right] = 0
 \end{aligned}$$

And thus (13.2.7) can be rewritten as

$$\sum_{i=1}^n \hat{\mathbb{E}}(Q_{t,i}) = 0 \quad \forall t$$

which is identical to (13.2.5). That is, when the means model is saturated, observation-level GEE and sequential nonparametric mean quasi-score lead to numerically identical estimates.

Now we consider the augmented version of (13.2.2). From (9.2.6), this can be written as

$$\sum_{i=1}^n \left[\frac{R_{t,i}}{\hat{\pi}_{t,i}} Q_{t,i} + \sum_{k=1}^t \left(\frac{R_{k-1,i}}{\hat{\pi}_{k-1,i}} - \frac{R_{k,i}}{\hat{\pi}_{k,i}} \right) \hat{\mathbb{E}}(Q_{t,i} | R_{k-1,i} = 1, \bar{Y}_{k-1,i}) \right] = 0 \quad \forall t \quad (13.2.8)$$

But by (13.2.4),

$$\sum_{i=1}^n \left[\sum_{k=1}^t \left(\frac{R_{k-1,i}}{\hat{\pi}_{k-1,i}} - \frac{R_{k,i}}{\hat{\pi}_{k,i}} \right) \hat{\mathbb{E}}(Q_{t,i} | R_{k-1,i} = 1, \bar{Y}_{k-1,i}) \right] \equiv 0 \quad \forall t$$

and thus (13.2.8) reduces to (13.2.2), which completes the proof of this theorem, that when the means model is saturated, observation-level GEE, augmented observation-level GEE and sequential nonparametric mean quasi-score all lead to numerically identical estimates. □

This numerical equivalence is not precisely reflected in our simulations for a number of reasons. First, mean quasi-score in its exact form was not included in the simulation study. However, MI-GEE is related to mean quasi-score and the equivalence between mean quasi-score and OWGEE gives rise to a near equivalence between MI-GEE and OWGEE, which is discussed further in §13.5.

As for the equivalence between OWGEE and its augmented counterpart, we would expect to see this manifested in the comparison between OWGEE and the estimators of Robins and Rotnitzky (1995) and Bang and Robins (2005). We do indeed see a greater difference between these methods and OWGEE when the means model is not saturated: for example, BR is more efficient than OWGEE only in the third set of simulations. The fact that they are not numerically the same when the means model is saturated is due in part to the fact that the model for the probability weights is not saturated. The weights are generated from a model which contains no interactions between the covariates and the previous outcomes and this is reflected in the model used to estimate the weights; in other words the coefficients for the interaction terms are fixed at their true values of zero. This is enough to cause the final estimates to differ slightly in their exact numerical values, although they are asymptotically equivalent. In the third set of simulations—when the means model is non-saturated—BR is more efficient than OWGEE, since it uses the information on the incomplete cases to learn about the parameters of the semiparametric model for the observed data distribution, but when this model is nonparametric, there is nothing to learn.

When the RR model converges, its estimates are very similar to those from OWGEE and the main difference between the results from these two methods comes as a result of the poor convergence of the RR method. If the probability weights model were saturated then we would expect the augmented estimating equation and the non-augmented OWGEE to share the same root, but we would not necessarily expect the two methods to converge to this root at the same rate and with the same success, which is a point made by Wang *et al.* (2007).

We wouldn't expect DRMI to be numerically equivalent to OWGEE, since DRMI is

merely an approximation to BR, as was discussed at length in Chapter 9.

13.3 Unweighted GEE: conditions for consistency and semiparametric efficiency

When there are no missing data and assuming throughout that the true correlation matrix is known, (7.3.1) can be written as

$$\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} (\tilde{\mathbf{Y}}_i - \tilde{\boldsymbol{\mu}}) = \mathbf{0}$$

where the $\tilde{}$ above each matrix is used to emphasise its full-data dimension.

When some components of $\tilde{\mathbf{Y}}_i$ are missing, we can write $\tilde{\mathbf{Y}}_i$ as $(\mathbf{Y}_i^{\text{obs}T}, \mathbf{Y}_i^{\text{mis}T})^T$. Consistent estimates of $\tilde{\boldsymbol{\mu}}$ could then in theory be obtained by solving

$$\sum_{i=1}^n \mathbb{E} \left[\tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} (\tilde{\mathbf{Y}}_i - \tilde{\boldsymbol{\mu}}) \mid \mathbf{Y}_i^{\text{obs}}, \mathbf{X}_i \right] = \mathbf{0}$$

which is equivalent to

$$\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \left[\mathbb{E} \left(\tilde{\mathbf{Y}}_i \mid \mathbf{Y}_i^{\text{obs}}, \mathbf{X}_i \right) - \tilde{\boldsymbol{\mu}} \right] = \mathbf{0} \quad (13.3.1)$$

As pointed out by Lipsitz *et al.* (2000), the GEE method of Liang and Zeger (1986) can be viewed as an approximate solution to (13.3.1) where the true conditional expectation $\mathbb{E} \left(\tilde{\mathbf{Y}}_i \mid \mathbf{Y}_i^{\text{obs}}, \mathbf{X}_i \right)$ is replaced by its approximation under the assumption of multivariate normality.

More specifically, as shown in the appendix of Lipsitz *et al.* (2000), if we partition $\tilde{\mathbf{W}}_i$

as

$$\tilde{\mathbf{W}}_i = \text{Var} \left(\begin{array}{c} \mathbf{Y}_i^{\text{mis}} \\ \mathbf{Y}_i^{\text{obs}} \end{array} \middle| \mathbf{X}_i \right) = \begin{pmatrix} \mathbf{W}_i^{\text{m}} & \mathbf{W}_i^{\text{m,o}} \\ \mathbf{W}_i^{\text{m,o}T} & \mathbf{W}_i^{\text{o}} \end{pmatrix}$$

then, under an assumption of multivariate normality,

$$\mathbb{E} \left(\begin{array}{c} \mathbf{Y}_i^{\text{mis}} \\ \mathbf{Y}_i^{\text{obs}} \end{array} \middle| \mathbf{Y}_i^{\text{obs}}, \mathbf{X}_i \right) = \begin{bmatrix} \boldsymbol{\mu}_i^{\text{mis}} + \mathbf{W}_i^{\text{m,o}} \mathbf{W}_i^{\text{o}-1} (\mathbf{Y}_i^{\text{obs}} - \boldsymbol{\mu}_i^{\text{obs}}) \\ \mathbf{Y}_i^{\text{obs}} \end{bmatrix}$$

Substituting in (13.3.1),

$$\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \begin{pmatrix} \mathbf{W}_i^{\text{m,o}} \mathbf{W}_i^{\text{o}-1} \\ \mathbf{I} \end{pmatrix} (\mathbf{Y}_i^{\text{obs}} - \boldsymbol{\mu}_i^{\text{obs}}) = \mathbf{0} \quad (13.3.2)$$

Using the form of the inverse of a partitioned matrix (Seber, 2008, p. 293), the left-hand side of (13.3.2) can be re-written as

$$\begin{aligned} & \sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \left[\begin{pmatrix} \mathbf{W}_i^{\text{m}-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \right. \\ & \quad \left. + \begin{pmatrix} \mathbf{W}_i^{\text{m}-1} \mathbf{W}_i^{\text{m,o}} \\ \mathbf{I} \end{pmatrix} \left(\mathbf{W}_i^{\text{o}} - \mathbf{W}_i^{\text{m,o}T} \mathbf{W}_i^{\text{m}-1} \mathbf{W}_i^{\text{m,o}} \right)^{-1} \left(-\mathbf{W}_i^{\text{m,o}T} \mathbf{W}_i^{\text{m}-1}, \mathbf{I} \right) \right] \\ & \quad \cdot \begin{pmatrix} \mathbf{W}_i^{\text{m,o}} \mathbf{W}_i^{\text{o}-1} \\ \mathbf{I} \end{pmatrix} (\mathbf{Y}_i^{\text{obs}} - \boldsymbol{\mu}_i^{\text{obs}}) \end{aligned}$$

$$\begin{aligned} & = \sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \left[\begin{pmatrix} \mathbf{W}_i^{\text{m}-1} \mathbf{W}_i^{\text{m,o}} \mathbf{W}_i^{\text{o}-1} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} -\mathbf{W}_i^{\text{m}-1} \mathbf{W}_i^{\text{m,o}} \\ \mathbf{I} \end{pmatrix} \right. \\ & \quad \left. \left(\mathbf{W}_i^{\text{o}} - \mathbf{W}_i^{\text{m,o}T} \mathbf{W}_i^{\text{m}-1} \mathbf{W}_i^{\text{m,o}} \right)^{-1} \left(-\mathbf{W}_i^{\text{m,o}T} \mathbf{W}_i^{\text{m}-1} \mathbf{W}_i^{\text{m,o}} + \mathbf{W}_i^{\text{o}} \right) \mathbf{W}_i^{\text{o}-1} \right] \\ & \quad \cdot (\mathbf{Y}_i^{\text{obs}} - \boldsymbol{\mu}_i^{\text{obs}}) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \begin{pmatrix} \mathbf{0} \\ \mathbf{W}_i^{\text{o-1}} \end{pmatrix} (\mathbf{Y}_i^{\text{obs}} - \boldsymbol{\mu}_i^{\text{obs}}) \\
&= \sum_{i=1}^n \mathbf{X}_i^{\text{o}T} \mathbf{D}_i^{\text{o}} \mathbf{W}_i^{\text{o-1}} (\mathbf{Y}_i^{\text{obs}} - \boldsymbol{\mu}_i^{\text{obs}})
\end{aligned}$$

This explicitly confirms the observation made by Lipsitz *et al.* (2000) that unweighted GEE under MAR corresponds to an assumption of multivariate normality when the true correlation matrix is assumed known. This is also consistent with the frequently quoted property of GEEs—that they give consistent estimates under MAR for Gaussian data when the correlation structure is correct and its parameters consistently estimated.

A corollary for binary data is that if, conditional on \mathbf{X}_i and each variable in $\mathbf{Y}_i^{\text{obs}}$, $\mathbf{E}(\mathbf{Y}_i^{\text{mis}})$ is independent of all pairwise and higher-order interactions between variables in $\mathbf{Y}_i^{\text{obs}}$, then the ‘multivariate normality assumption’ holds and GEE gives consistent estimates under MAR, again under the assumption that the working correlation structure is the true one. We refer to this condition henceforth as the linearity condition.

In our simulation studies in the previous chapter, the true correlation structure was changed between the first and second sets of simulations in order to investigate this. In the logistic regression of Y_3 on X_1 , X_2 , X_1X_2 , Y_1 , Y_2 and Y_1Y_2 , the coefficient of Y_1Y_2 is zero in the second set of simulations, compared with 1.75 in the first set. Correspondingly, we see that the bias in GEE is smaller in the second set of simulations. The small residual variance could be a consequence of the inconsistency in estimating the parameters of the correlation matrix which could be reduced using quadratic estimation as described in §7.3.1.

In Section 4 of Robins and Rotnitzky (1995), the paper which exhibits the semiparametric-efficient estimator, which we refer to as RR, the authors derive the condition needed under MCAR for GEE to be semiparametric-efficient among all estimators belonging to the restricted moment class \mathcal{R} (see Definition 3.20), and this condition is precisely the linearity condition described above. Their argument appeals

to the same multivariate normality assumption described above and the MCAR condition is only necessary to ensure consistency. Now that we have established that the linearity condition leads to consistency under MAR, we have the following result:

Theorem 13.2 (Conditions for the consistency and semiparametric efficiency of GEE).

The solution to

$$\sum_{i=1}^n \mathbf{X}_i^{\circ T} \mathbf{D}_i^{\circ} \mathbf{W}_i^{\circ -1} (\mathbf{Y}_i^{\text{obs}} - \hat{\boldsymbol{\mu}}_i^{\text{obs}}) = \mathbf{0}$$

is a consistent estimator of $\boldsymbol{\mu}$ under MAR if

- *The correlation structure is correctly specified and its parameters consistently estimated.*
- *Conditional on \mathbf{X}_i and each variable in $\mathbf{Y}_i^{\text{obs}}$, $\mathbb{E}(\mathbf{Y}_i^{\text{mis}})$ is independent of all pairwise and higher-order interactions between variables in $\mathbf{Y}_i^{\text{obs}}$.*

Furthermore, under these conditions, the asymptotic variance of $\hat{\boldsymbol{\mu}}$ attains the semiparametric efficiency bound for all estimators in the restricted moment class \mathcal{R} .

13.4 Cluster- versus observation-level weighting

If we consider the simplest case of cluster- versus observation-level independence estimating equations (CWIEE and OWIEE, respectively), we see that CWIEE is intuitively less satisfactory than OWIEE. Under the independence structure, the parameter estimates are equivalent to estimates from separate logistic regression analyses, one for each timepoint, and all available observations contribute to both CWIEE and OWIEE. These available observations are weighted differently, however, in the two analyses, except for the final timepoint, since—under monotonicity—the probability of dropping out *after* the final timepoint is equal to the probability of being observed *at* the final timepoint, and the cluster- and observation-level weights coincide. However, the

analyses at the other two timepoints are, in general, different. At the first timepoint, when we assume that all subjects are observed, the OWIEE analysis weights each subject equally, whereas the CWIEE analysis weights the subjects differently according to their subsequent dropout pattern. This doesn't make CWIEE inconsistent, since (for example, when there are three timepoints):

$$\mathbb{E} \left[\frac{\sum_{i=1}^n \left(\frac{1-R_2}{1-\pi_2} Y_1 + \frac{R_2-R_3}{\pi_2-\pi_3} Y_1 + \frac{R_3}{\pi_3} Y_1 \right)}{\sum_{i=1}^n \left(\frac{1-R_2}{1-\pi_2} + \frac{R_2-R_3}{\pi_2-\pi_3} + \frac{R_3}{\pi_3} \right)} \right] \rightarrow \mathbb{E}(Y_1)$$

and

$$\mathbb{E} \left[\frac{\sum_{i=1}^n \left(\frac{R_2-R_3}{\pi_2-\pi_3} Y_2 + \frac{R_3}{\pi_3} Y_2 \right)}{\sum_{i=1}^n \left(\frac{R_2-R_3}{\pi_2-\pi_3} + \frac{R_3}{\pi_3} \right)} \right] \rightarrow \mathbb{E}(Y_2)$$

as $n \rightarrow \infty$.

However, the comparison between CWIEE and OWIEE can be thought of as a comparison between two sets of weighted logistic regressions, where, at every timepoint except for the final one, the former uses weights which are a 'noisier' version of the weights used by the latter. For example, for the second timepoint (again, in an example where there are three timepoints), when the weights used in OWIEE are $w_2 = \frac{1}{\pi_2}$, the weights used in CWIEE are $w'_2 = \frac{w_2}{\mathbb{P}(R_3=1|R_2=1)}$ if Y_3 is observed, and $w''_2 = \frac{w_2}{\mathbb{P}(R_3=0|R_2=1)}$ otherwise. If we consider the weighted average Y_2 for two subjects who both share identical values of the covariates and Y_1 , then they will also both share the same observation-level weight, w_2 . However, if one of these two has an observed value of Y_3 , whereas the other subject's Y_3 is missing, they will have two different values (w'_2 and w''_2) of the cluster-level weight, where both w'_2 and w''_2 are greater than w_2 . If the variance of Y_2 is σ^2 then the variance of the weighted average of these two observations will be

$$\frac{w_2^2 + w_2^2}{(w_2 + w_2)^2} \sigma^2 = \frac{1}{2} \sigma^2$$

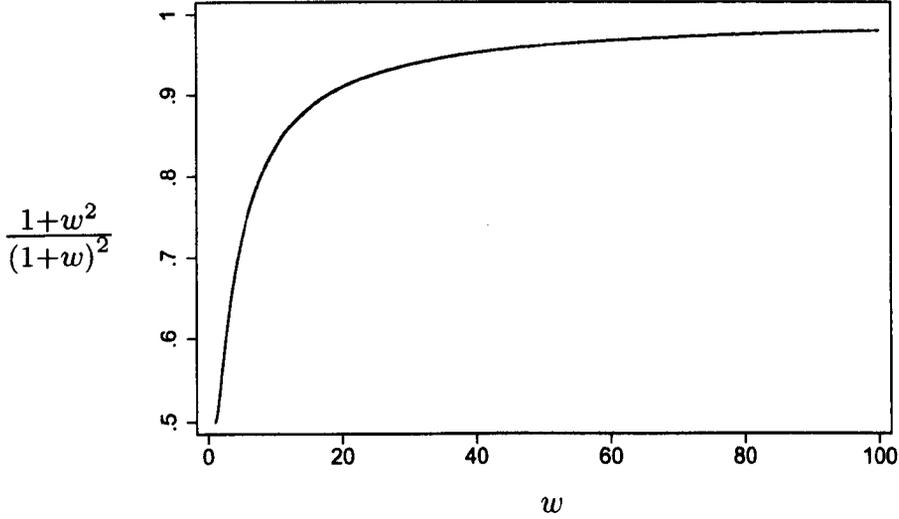


Figure 13.1: The variance of the weighted average of two second timepoint observations from a three timepoint CWIEE. The corresponding OWIEE variance occurs when $w = 1$.

in the OWIEE analysis, and

$$\frac{w_2'^2 + w_2''^2}{(w_2' + w_2'')^2} \sigma^2 = \frac{\frac{1}{\pi_3} + \frac{1}{(\pi_2 - \pi_3)^2}}{\left(\frac{1}{\pi_3} + \frac{1}{\pi_2 - \pi_3}\right)^2} \sigma^2 = \frac{1 + \left(\frac{\pi_3}{\pi_2 - \pi_3}\right)^2}{\left(1 + \frac{\pi_3}{\pi_2 - \pi_3}\right)^2} \sigma^2 = \frac{1 + w^2}{(1 + w)^2} \sigma^2$$

where $w = \frac{\pi_3}{\pi_2 - \pi_3}$ and, without loss of generality, we can assume that $w > 1$. There is no loss of generality here, since, if $w < 1$, we could redefine $w = \frac{\pi_2 - \pi_3}{\pi_3}$ and

$$\frac{w_2'^2 + w_2''^2}{(w_2' + w_2'')^2} \sigma^2 = \frac{w^2 + 1}{(w + 1)^2} \sigma^2$$

A graph of the function $f(w) = \frac{1+w^2}{(1+w)^2}$ is shown in Fig. 13.1 and this shows that the relative efficiency of CWIEE gets worse as w gets larger and the optimal efficiency is at $w = 1$, which corresponds to the OWIEE analysis. This argument generalises to n subjects and to any of the T timepoints, leading to the conclusion that OWIEE is

always at least as efficient as CWIEE.

We have already made the observation that at the final timepoint, CWIEE and OWIEE coincide. There is another special case in which CWIEE and OWIEE coincide at *all* timepoints, and this is when the covariates are categorical, the means model saturated and the weights estimated nonparametrically, as we will now show.

Theorem 13.3 (Numerical equivalence of CWIEE and OWIEE). *We assume that all covariates are discrete, that the means model is saturated and that the missing data mechanism is monotone and MAR. Under these conditions, CWIEE and OWIEE, give numerically identical results, provided that the weights in both methods are estimated nonparametrically.*

Proof. Since we are assuming a saturated means model, it is sufficient to show that cluster- and observation-level weighted averages of Y_t are equal, i.e. we need to show that

$$\frac{\sum_{i=1}^n \frac{R_{t,i} Y_{t,i}}{\hat{\pi}_{t,i}}}{\sum_{i=1}^n \frac{R_{t,i}}{\hat{\pi}_{t,i}}} = \frac{\sum_{i=1}^n \left(\sum_{k=t}^{T-1} \frac{R_{k,i} - R_{k+1,i}}{\hat{\pi}_{k,i} - \hat{\pi}_{k+1,i}} Y_{t,i} + \frac{R_{T,i}}{\hat{\pi}_{T,i}} Y_{t,i} \right)}{\sum_{i=1}^n \left(\sum_{k=t}^{T-1} \frac{R_{k,i} - R_{k+1,i}}{\hat{\pi}_{k,i} - \hat{\pi}_{k+1,i}} + \frac{R_{T,i}}{\hat{\pi}_{T,i}} \right)} \tag{13.4.1}$$

By (13.2.3), we have that

$$\sum_{i=1}^n \frac{R_{t,i} Y_{t,i}}{\hat{\pi}_{t,i}} = \sum_{i=1}^n \frac{R_{t+1,i} Y_{t,i}}{\hat{\pi}_{t+1,i}} = \dots = \sum_{i=1}^n \frac{R_{T,i} Y_{t,i}}{\hat{\pi}_{T,i}}$$

which also implies that, by setting $Y_{t,i} = 1$,

$$\sum_{i=1}^n \frac{R_{t,i}}{\hat{\pi}_{t,i}} = \sum_{i=1}^n \frac{R_{t+1,i}}{\hat{\pi}_{t+1,i}} = \dots = \sum_{i=1}^n \frac{R_{T,i}}{\hat{\pi}_{T,i}}$$

We can re-write the other terms using

$$\frac{(R_{k,i} - R_{k+1,i}) Y_{k,i}}{\hat{\pi}_{k,i} - \hat{\pi}_{k+1,i}} = \frac{D_{k+1,i} Y_{t,i}}{\hat{\pi}'_{k+1,i}} \tag{13.4.2}$$

and

$$\frac{R_{k,i} - R_{k+1,i}}{\hat{\pi}_{k,i} - \hat{\pi}_{k+1,i}} = \frac{D_{k+1,i}}{\hat{\pi}'_{k+1,i}}$$

where $D_{k+1,i}$ is the dropout indicator for dropping out at time $k + 1$ and $\hat{\pi}'_{k+1,i}$ is the probability of dropping out at time $k + 1$, conditional on $\bar{Y}_{k,i}$.

But the sum over i of (13.4.2) can be rewritten as

$$\begin{aligned} \sum_{i=1}^n \frac{D_{k+1,i} Y_{t,i}}{\hat{\pi}'_{k+1,i}} &= \sum_{i=1}^n \frac{\sum_{j=1}^n \mathbf{1}(\bar{Y}_{k,j} = \bar{Y}_{k,i})}{(1 - \hat{\pi}'_{k,i} - \hat{\pi}'_{k-1,i} - \dots - \hat{\pi}'_{1,i}) \sum_{l=1}^n D_{k+1,l} \mathbf{1}(\bar{Y}_{k,l} = \bar{Y}_{k,i})} D_{k+1,i} Y_{t,i} \\ &= \sum_{i=1}^n \sum_{j=1}^n \frac{\mathbf{1}(\bar{Y}_{k,j} = \bar{Y}_{k,i})}{\hat{\pi}_{k,i} \sum_{l=1}^n D_{k+1,l} \mathbf{1}(\bar{Y}_{k,l} = \bar{Y}_{k,i})} D_{k+1,i} Y_{t,i} \\ &= \sum_{j=1}^n \frac{\sum_{i=1}^n \mathbf{1}(\bar{Y}_{k,j} = \bar{Y}_{k,i}) D_{k+1,i} Y_{t,i}}{\hat{\pi}_{k,j} \sum_{l=1}^n D_{k+1,l} \mathbf{1}(\bar{Y}_{k,l} = \bar{Y}_{k,j})} \\ &= \sum_{j=1}^n \frac{R_{k,j} Y_{t,j} \sum_{i=1}^n D_{k+1,i} \mathbf{1}(\bar{Y}_{k,j} = \bar{Y}_{k,i})}{\hat{\pi}_{k,j} \sum_{l=1}^n D_{k+1,l} \mathbf{1}(\bar{Y}_{k,l} = \bar{Y}_{k,j})} \\ &= \sum_{j=1}^n \frac{R_{k,j} Y_{t,j}}{\hat{\pi}_{k,j}} \end{aligned} \tag{13.4.3}$$

which also implies that, by setting $Y_{t,i} = 1$,

$$\sum_{i=1}^n \frac{D_{k+1,i}}{\hat{\pi}'_{k+1,i}} = \sum_{j=1}^n \frac{R_{k,j}}{\hat{\pi}_{k,j}}$$

and thus both the numerator and the denominator on the RHS of (13.4.1) are $(T - t + 1)$ times the corresponding numerator and denominator on the LHS, which proves the result. □

Corollary 13.4 (Numerical equivalence of CWIEE and OWGEE). *Assuming again that all covariates are discrete, that the means model is saturated and that the missing data mechanism is monotone and MAR, then CWIEE and OWGEE give numerically identical results, for any choice of covariance structure, provided that the weights in both methods are estimated nonparametrically.*

Proof. This follows trivially from Theorem 13.3 and the fact that when the means model is saturated, full data GEE is equivalent to full data IEE for all choices of the covariance structure for GEE (O'Brien *et al.*, 2006). As was discussed in §7.4, the way in which OWGEE is formulated creates what is effectively a full data structure from the incomplete data. This proves the corollary. \square

It remains now to compare OWIEE and CWGEE when the means model is saturated and finally to compare OWGEE and CWGEE when the means model is not saturated.

Lemma 13.5. *Let C_d be the upper left ($d \times d$) block of the correlation matrix for $\{Y_{1,i}, Y_{2,i}, \dots, Y_{T,i}\}$ and \mathbf{b}_d be the first $d - 1$ elements of the d^{th} column of C_d . Then*

$$\begin{aligned} C_t^{-1} &= \begin{pmatrix} C_1^{-1} & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix} + \frac{\det C_1}{\det C_2} \begin{pmatrix} C_1^{-1} \mathbf{b}_2 \mathbf{b}_2^T C_1^{-1} & -C_1^{-1} \mathbf{b}_2 & 0 \\ -\mathbf{b}_2^T C_1^{-1} & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \dots \\ &+ \frac{\det C_{t-2}}{\det C_{t-1}} \begin{pmatrix} C_{t-2}^{-1} \mathbf{b}_{t-1} \mathbf{b}_{t-1}^T C_{t-2}^{-1} & -C_{t-2}^{-1} \mathbf{b}_{t-1} & 0 \\ -\mathbf{b}_{t-1}^T C_{t-2}^{-1} & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ &+ \frac{\det C_{t-1}}{\det C_t} \begin{pmatrix} C_{t-1}^{-1} \mathbf{b}_t \mathbf{b}_t^T C_{t-1}^{-1} & -C_{t-1}^{-1} \mathbf{b}_t \\ -\mathbf{b}_t^T C_{t-1}^{-1} & 1 \end{pmatrix} \end{aligned}$$

Proof. This follows immediately from repeated applications of

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix} + \begin{pmatrix} -\mathbf{A}_{11}^{-1} \mathbf{A}_{12} \\ \mathbf{I} \end{pmatrix} \mathbf{A}_{22 \cdot 1}^{-1} (-\mathbf{A}_{21} \mathbf{A}_{11}^{-1}, \mathbf{I})$$

(Seber, 2008, p. 293), where $\mathbf{A}_{22 \cdot 1} = \mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12}$, and

$$\det \begin{pmatrix} \mathbf{T} & \mathbf{U} \\ \mathbf{V} & \mathbf{W} \end{pmatrix} = \det \mathbf{T} \det (\mathbf{W} - \mathbf{V} \mathbf{T}^{-1} \mathbf{U})$$

(Harville, 1997, p. 189). □

Lemma 13.6. *Under the assumption that all covariates are discrete, that the means model is saturated and that the missing data mechanism is monotone and MAR, then the expression*

$$\sum_{i=1}^n \frac{R_{t,i} - R_{t+1,i}}{\hat{\pi}_{t,i} - \hat{\pi}_{t+1,i}} (Y_{s,i} - \mu_s)$$

has numerically exactly the same value for all t and s satisfying $t \geq s$.

Proof. This follows automatically from (13.4.3) and (13.2.4). □

Theorem 13.7 (Numerical equivalence of CWGEE and OWGEE). *Assuming again that all covariates are discrete, that the means model is saturated and that the missing data mechanism is monotone and MAR, then CWGEE and OWGEE give numerically identical results, for any choice of covariance structure (correct or incorrect), provided that the weights in both methods are estimated nonparametrically.*

Proof. Recall the cluster-weighted estimating equation (7.4.1):

$$\sum_{i=1}^n \frac{1}{\mathbf{P}(D_i = d_i | \mathbf{X}_i, \mathbf{Y}_i)} \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

When the means model is saturated, without loss of generality, by considering each covariate combination as a separate dataset, this can be rewritten as

$$\sum_{i=1}^n \frac{1}{\mathbf{P}(D_i = d_i | \mathbf{Y}_i)} \mathbf{D}_i \mathbf{W}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

Note that, unlike observation-level weighting, \mathbf{W}_i^{-1} is not constant for cluster-level weighted GEE. Since \mathbf{D}_i is a diagonal matrix, its effect is to multiply each row, t , on the LHS by a constant factor $\mu_t(1 - \mu_t)$ and thus the combined effect of $\mathbf{D}_i \mathbf{W}_i^{-1}$ is to multiply each row by $\sqrt{\mu_t(1 - \mu_t)}$. Thus we can rewrite CWGEE as T separate scalar

equations:

$$\sum_{i=1}^n \sum_{d=s}^T \frac{\mathbb{1}(D_i = d + 1)}{\hat{\pi}'_{d+1,i}} (\mathbf{C}_d^{-1})_{(s)} \bar{\boldsymbol{\varepsilon}}_{d,i} = 0 \quad \forall s \in \{1, \dots, T\} \tag{13.4.4}$$

where $\hat{\pi}'_{d+1,i} = \hat{\pi}_{d,i} - \hat{\pi}_{d+1,i}$ is the nonparametric estimator of $\mathbb{P}(D_i = d + 1 | \bar{\mathbf{Y}}_{d,i})$, \mathbf{C}_d is the upper left $(d \times d)$ block of the correlation matrix for $(Y_{1,i}, Y_{2,i}, \dots, Y_{T,i})$, $(\mathbf{A})_{(s)}$ is the s^{th} row of \mathbf{A} and

$$\bar{\boldsymbol{\varepsilon}}_{d,i} = \begin{pmatrix} \frac{Y_{1,i} - \mu_1}{\sqrt{\mu_1(1-\mu_1)}} \\ \frac{Y_{2,i} - \mu_2}{\sqrt{\mu_2(1-\mu_2)}} \\ \vdots \\ \frac{Y_{d,i} - \mu_d}{\sqrt{\mu_d(1-\mu_d)}} \end{pmatrix}$$

By Lemma 13.5, the LHS of (13.4.4) can be rewritten as

$$\begin{aligned} & \sum_{i=1}^n \left\{ \left(\sum_{k=s}^T \frac{\mathbb{1}(D_i = k + 1)}{\hat{\pi}'_{k+1,i}} \right) \frac{\det \mathbf{C}_{s-1}}{\det \mathbf{C}_s} (-\mathbf{b}_s^T \mathbf{C}_{s-1}^{-1} \bar{\boldsymbol{\varepsilon}}_{s-1} + \varepsilon_s) \right. \\ & + \left(\sum_{k=s+1}^T \frac{\mathbb{1}(D_i = k + 1)}{\hat{\pi}'_{k+1,i}} \right) \frac{\det \mathbf{C}_s}{\det \mathbf{C}_{s+1}} \left[(\mathbf{C}_s^{-1} \mathbf{b}_{s+1} \mathbf{b}_{s+1}^T \mathbf{C}_s^{-1})_{(s)} \bar{\boldsymbol{\varepsilon}}_s - (\mathbf{C}_s^{-1} \mathbf{b}_{s+1})_{(s)} \varepsilon_{s+1} \right] + \dots \\ & \left. + \frac{\mathbb{1}(D_i = T + 1) \det \mathbf{C}_{T-1}}{\hat{\pi}'_{T+1,i} \det \mathbf{C}_T} \left[(\mathbf{C}_{T-1}^{-1} \mathbf{b}_T \mathbf{b}_T^T \mathbf{C}_{T-1}^{-1})_{(s)} \bar{\boldsymbol{\varepsilon}}_{T-1} - (\mathbf{C}_{T-1}^{-1} \mathbf{b}_T)_{(s)} \varepsilon_T \right] \right\} \\ & = \sum_{i=1}^n \sum_{t=1}^T \sum_{d=\max\{t,s\}}^T a_{d+1,t} \frac{\mathbb{1}(D_i = d + 1)}{\hat{\pi}'_{d+1,i}} (Y_{t,i} - \mu_t) \end{aligned}$$

where $a_{d+1,t}$ are constants, not dependent on i .

By Lemma 13.6, this can be rewritten as

$$\sum_{i=1}^n \sum_{t=1}^T \frac{\mathbb{1}(D_i = t + 1)}{\hat{\pi}'_{t+1,i}} (Y_{t,i} - \mu_t) \sum_{d=\max\{t,s\}}^T a_{d+1,t}$$

which, by (13.4.3), can be rewritten as

$$\sum_{i=1}^n \sum_{t=1}^T \frac{R_{t,i}}{\hat{\pi}_{t,i}} (Y_{t,i} - \mu_t) \sum_{d=\max\{t,s\}}^T a_{d+1,t}$$

But this proves that the OWIEE estimator is also a root of the cluster-weighted estimating equation (7.4.1) in the saturated means case, which—if we ignore the possibility of multiple roots (Heyde and Morton, 1998)—together with Corollary 13.4 proves the result. □

It remains now to compare OWGEE and CWGEE when the means model is *not* saturated.

Recall that the cluster-weighted estimating equation (7.4.1) is:

$$\sum_{i=1}^n \omega_i \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \beta_i^{\text{CWGEE}}) = 0$$

where

$$\omega_i = \frac{1}{\mathbb{P}(D_i = d_i | \mathbf{X}_i, \mathbf{Y}_i)}$$

Treating the weights as fixed (rather than estimated) and using the variance formula derived by Robins *et al.* (1995), the variance-covariance matrix of $\hat{\beta}^{\text{CWGEE}}$ can be estimated using the following sandwich estimator:

$$\begin{aligned} & \left[\left(\sum_{i=1}^n \omega_i \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right)^{-1} \right]^T \\ & \cdot \left[\sum_{i=1}^n \omega_i^2 \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta}_i^{\text{CWGEE}}) (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta}_i^{\text{CWGEE}})^T \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right] \\ & \cdot \left(\sum_{i=1}^n \omega_i \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right)^{-1} \end{aligned}$$

Similarly, the variance-covariance matrix of $\hat{\beta}^{\text{OWGEE}}$, the estimator which solves

$$\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \Phi_i (\tilde{\mathbf{Y}}_i - \tilde{\mathbf{X}}_i \tilde{\beta}_i^{\text{OWGEE}}) = \mathbf{0} \tag{13.4.5}$$

can be consistently estimated using the following sandwich estimator:

$$\begin{aligned} & \left[\left(\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \Phi_i \tilde{\mathbf{D}}_i \tilde{\mathbf{X}}_i \right)^{-1} \right]^T \\ & \cdot \left(\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \Phi_i (\tilde{\mathbf{Y}}_i - \tilde{\mathbf{X}}_i \hat{\beta}_i^{\text{OWGEE}}) (\tilde{\mathbf{Y}}_i - \tilde{\mathbf{X}}_i \hat{\beta}_i^{\text{OWGEE}})^T \Phi_i \tilde{\mathbf{W}}_i^{-1} \tilde{\mathbf{D}}_i \tilde{\mathbf{X}}_i \right) \\ & \cdot \left(\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \Phi_i \tilde{\mathbf{D}}_i \tilde{\mathbf{X}}_i \right)^{-1} \end{aligned}$$

To show that OWGEE is always at least as efficient as CWGEE, we must show the following:

Proposition 13.8.

$$\begin{aligned} & \mathbf{E} \left(\left\{ \left[\left(\sum_{i=1}^n \omega_i \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right)^{-1} \right]^T \right. \right. \\ & \cdot \left[\sum_{i=1}^n \omega_i^2 \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta}_i^{\text{CWGEE}}) (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta}_i^{\text{CWGEE}})^T \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right] \\ & \cdot \left. \left(\sum_{i=1}^n \omega_i \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right)^{-1} \right\} - \left\{ \left[\left(\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \Phi_i \tilde{\mathbf{D}}_i \tilde{\mathbf{X}}_i \right)^{-1} \right]^T \right. \\ & \cdot \left[\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \Phi_i (\tilde{\mathbf{Y}}_i - \tilde{\mathbf{X}}_i \hat{\beta}_i^{\text{OWGEE}}) (\tilde{\mathbf{Y}}_i - \tilde{\mathbf{X}}_i \hat{\beta}_i^{\text{OWGEE}})^T \Phi_i \tilde{\mathbf{W}}_i^{-1} \tilde{\mathbf{D}}_i \tilde{\mathbf{X}}_i \right] \\ & \cdot \left. \left. \left(\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \Phi_i \tilde{\mathbf{D}}_i \tilde{\mathbf{X}}_i \right)^{-1} \right\} \right) \tag{13.4.6} \end{aligned}$$

is non-negative definite.

Demonstration using computer simulations. Due to the complex nature of (13.4.6), we have not been able to prove Proposition 13.8 mathematically. However, for 2, 3 and 4 timepoints, we can demonstrate the results in reasonable generality using computer simulations, as we now describe.

First we generate two independent covariates, X_1 and X_2 , each from a $N(0, 1)$ distribution. Then we generate each $Y_{j,i}$ from a Bahadur distribution with

$$\mathbb{P}(Y_{j,i} = 1 | X_{1,i}, X_{2,i}) = \frac{\exp(c_{j,0} + c_{j,1}X_{1,i} + c_{j,2}X_{2,i} + c_{j,3}X_{1,i}X_{2,i})}{1 + \exp(c_{j,0} + c_{j,1}X_{1,i} + c_{j,2}X_{2,i} + c_{j,3}X_{1,i}X_{2,i})}$$

where each $c_{j,k}$ is generated independently and at random from a $U(-1, 1)$ distribution, and correlations $\rho_{j_1,j_2}, \rho_{j_1,j_2,j_3}, \dots$ also generated from $U(-1, 1)$ distributions. The algorithm searches for combinations of these coefficients and correlations which give rise to a well-defined joint distribution, i.e. one in which

$$0 \leq \mathbb{P}(Y_{j,i} = 1 | \bar{Y}_{j-1,i}, X_{1,i}, X_{2,i}) \leq 1$$

for each i and j .

The data are then subjected to monotone missingness according to a MAR mechanism, for example:

$$\mathbb{P}(R_{3,i} = 1 | \bar{Y}_{2,i}, X_{1,i}, X_{2,i}, R_{2,i} = 1) = \frac{\exp \left(\begin{aligned} &c_{3,4} + c_{3,5}X_{1,i} + c_{3,6}X_{2,i} + c_{3,7}X_{1,i}X_{2,i} + c_{3,8}Y_{1,i} + c_{3,9}X_{1,i}Y_{1,i} + c_{3,10}X_{2,i}Y_{1,i} \\ &+ c_{3,11}X_{1,i}X_{2,i}Y_{1,i} + c_{3,12}Y_{2,i} + c_{3,13}X_{1,i}Y_{2,i} + c_{3,14}X_{2,i}Y_{2,i} + c_{3,15}X_{1,i}X_{2,i}Y_{2,i} \\ &+ c_{3,16}Y_{1,i}Y_{2,i} + c_{3,17}X_{1,i}Y_{1,i}Y_{2,i} + c_{3,18}X_{2,i}Y_{1,i}Y_{2,i} + c_{3,19}X_{1,i}X_{2,i}Y_{1,i}Y_{2,i} \end{aligned} \right)}{1 + \exp \left(\begin{aligned} &c_{3,4} + c_{3,5}X_{1,i} + c_{3,6}X_{2,i} + c_{3,7}X_{1,i}X_{2,i} + c_{3,8}Y_{1,i} + c_{3,9}X_{1,i}Y_{1,i} + c_{3,10}X_{2,i}Y_{1,i} \\ &+ c_{3,11}X_{1,i}X_{2,i}Y_{1,i} + c_{3,12}Y_{2,i} + c_{3,13}X_{1,i}Y_{2,i} + c_{3,14}X_{2,i}Y_{2,i} + c_{3,15}X_{1,i}X_{2,i}Y_{2,i} \\ &+ c_{3,16}Y_{1,i}Y_{2,i} + c_{3,17}X_{1,i}Y_{1,i}Y_{2,i} + c_{3,18}X_{2,i}Y_{1,i}Y_{2,i} + c_{3,19}X_{1,i}X_{2,i}Y_{1,i}Y_{2,i} \end{aligned} \right)}$$

where each $c_{3,4} \dots c_{3,19}$ is generated independently and at random from a $U(-1, 1)$ distribution.

$$\begin{aligned}
 & \left\{ \left[\left(\sum_{i=1}^n \omega_i \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right)^{-1} \right]^T \right. \\
 & \quad \cdot \left[\sum_{i=1}^n \omega_i^2 \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}_i) (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}_i)^T \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right] \\
 & \quad \cdot \left. \left(\sum_{i=1}^n \omega_i \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right)^{-1} \right\} - \left\{ \left[\left(\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \boldsymbol{\Phi}_i \tilde{\mathbf{D}}_i \tilde{\mathbf{X}}_i \right)^{-1} \right]^T \right. \\
 & \quad \cdot \left[\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \boldsymbol{\Phi}_i (\tilde{\mathbf{Y}}_i - \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\beta}}_i) (\tilde{\mathbf{Y}}_i - \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\beta}}_i)^T \boldsymbol{\Phi}_i \tilde{\mathbf{W}}_i^{-1} \tilde{\mathbf{D}}_i \tilde{\mathbf{X}}_i \right] \\
 & \quad \cdot \left. \left(\sum_{i=1}^n \tilde{\mathbf{X}}_i^T \tilde{\mathbf{D}}_i \tilde{\mathbf{W}}_i^{-1} \boldsymbol{\Phi}_i \tilde{\mathbf{D}}_i \tilde{\mathbf{X}}_i \right)^{-1} \right\} \quad (13.4.7)
 \end{aligned}$$

is then evaluated using the true (known) values of ω_i , $\boldsymbol{\Phi}_i$ and $\boldsymbol{\beta}$, using the true (known) pairwise correlations to evaluate \mathbf{W}_i and $\tilde{\mathbf{W}}_i$.

A sample size of 10,000 is used, to keep the Monte Carlo error low. The eigenvalues of (13.4.7) are then evaluated using Mathematica[®]. This is repeated for 1,000 datasets with 2 timepoints, 1,000 datasets with 3 timepoints and 1,000 datasets with 4 timepoints. The eigenvalues are plotted in Figs. 13.2–13.4. Although some eigenvalues are negative, the magnitude of these is sufficiently small to be explained by Monte Carlo error. These plots constitute strong evidence that the eigenvalues of (13.4.6) are all non-negative, which implies (see Harville, 1997, p. 543) that (13.4.6) is non-negative definite.

□

The decreased efficiency of cluster-level weighting compared with observation-level weighting can also be seen in Figs. 13.5 and 13.6. Since cluster-level weights are inverse probabilities of dropout, as opposed to inverse probabilities of being observed, the problem gets worse as the number of timepoints increases and the probability of

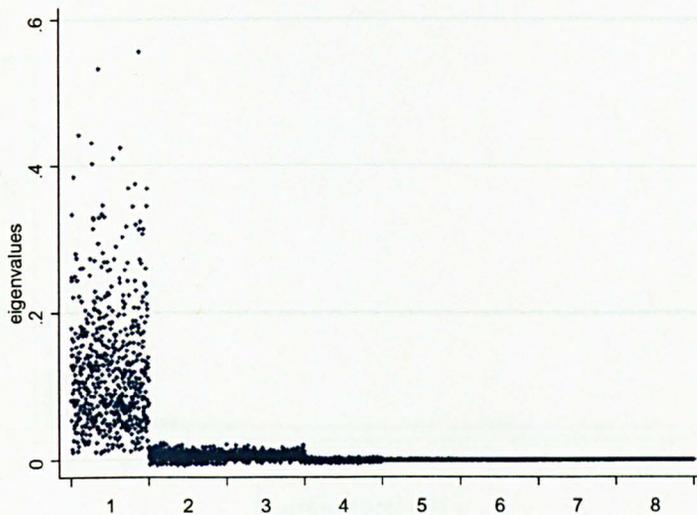


Figure 13.2: The 8 eigenvalues of (13.8) with 2 timepoints evaluated for 1,000 different datasets.

dropping out at the exact time of dropout decreases. This can be seen in Fig. 13.4 where the eigenvalues increase as the number of timepoints increases.

In summary, we feel that there is never a reason to prefer cluster- over observation-level weighting, and (except for the equivalence situations described above) observation- is more efficient than cluster-level weighting.

13.5 MI-GEE and its relationship with observation-level weighted GEE

In the saturated means model case, if the imputations in MI are drawn nonparametrically, then, in the limit as the number of imputations tends to infinity, the estimates from MI-GEE are equivalent to those from nonparametric sequential mean quasi-score.

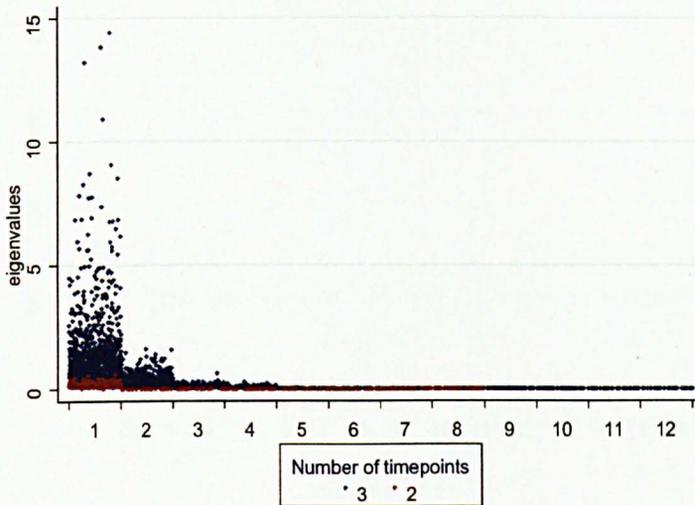


Figure 13.3: The 12 eigenvalues of (13.8) with 3 timepoints evaluated for 1,000 different datasets, with the 8 eigenvalues for the 2 timepoints case superimposed.

This follows from the fact that—in the saturated case—imputation of the quasi-scores is equivalent to the imputation of the missing outcomes, since the only part of the score which need be considered is $\mathbf{Y} - \boldsymbol{\mu}$. As the number of imputations increases, the proportion of imputed ones will tend to the corresponding nonparametric estimate of the expectation of that outcome. Thus, given Theorem 13.1, in the case where the imputations are drawn nonparametrically, we would expect MI-GEE to be approximately equivalent to OWGEE, with a slight reduction in efficiency due to a finite number of imputations. In our simulation results, however, MI-GEE appears to be more efficient than OWGEE, even in the case when the means model is saturated. This is due to the fact that our imputations were not drawn nonparametrically, but rather from the correct parametric model (when the y -model is correct). This increases the efficiency, but of course there is a corresponding decrease in robustness, as can be seen in the simulations when the y -model is incorrect.

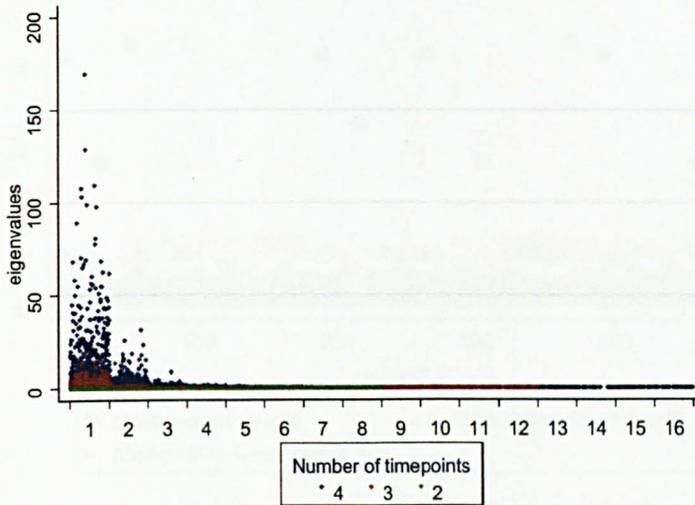


Figure 13.4: The 16 eigenvalues of (13.8) with 4 timepoints evaluated for 1,000 different datasets, with the 12 eigenvalues for the 3 timepoints case and the 8 eigenvalues for the 2 timepoints case superimposed.

13.6 A comparison of doubly robust MI and other doubly robust procedures

As we explained in Chapter 9, DRMI is an approximation to BR and we would not have expected it to perform as well in the examples where BR can be applied. However, in our simulations, DRMI was less biased and more efficient. We believe that this is partly due to a limitation of BR (and RR) which is overcome in DRMI, but is also due to the way in which we chose to implement BR and RR.

Let us consider an example (similar to our simulations) in which there are three timepoints. The BR implementation involves fitting a 'suitable regression model' to the predictions $\hat{E}(Y_3 | Y_1, Y_2)$ on Y_1 and the RR implementation involves fitting a 'suitable regression model' to the weighted residuals $K_{j,t,i} = \hat{\pi}_{t,i} \hat{\pi}_{j,i}^{-1} (Y_{j,i} - \mu_{j,i})$. In both cases, these quantities are not binary and yet a linear regression is certainly not sensible.

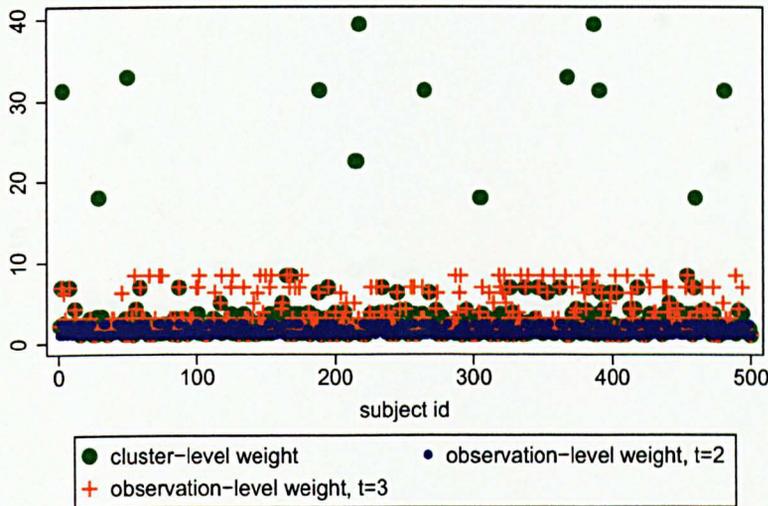


Figure 13.5: A comparison of the cluster-level and observation-level weights for the first simulation in the first set.

Robins and Rotnitzky (1995) point this out by saying “let $\hat{\tau}_j^{(t)}$ be the (possibly non-linear) least squares estimator of $\tau_j^{(t)}$ ”. However, Bang and Robins (2005) make no reference to this problem and claim that their method can be implemented using only “standard off-the-shelf regression software”. Since the predictions in the BR method come from a logistic regression, and lie strictly between 0 and 1, it would be possible in theory to use the logistic regression Fisher scoring algorithm to obtain estimates under the correct nonlinear model, but (at least in Stata) this requires some tweaking of the logistic regression command, and thus the claim that only “off-the-shelf” software need be used is not strictly true. For this reason, we decided in our simulation study to use linear regression at this stage in the procedure (and in the corresponding stages in the RR procedure). We believe that this is the reason for the small bias seen in these methods compared with DRMI. In DRMI, $\mathbb{E}(Y_3 | Y_1)$ is estimated by first drawing binary imputations \tilde{Y}_2 from $\mathbb{E}(Y_2 | Y_1)$ and then fitting a logistic regression to $\mathbb{E}(Y_3 | Y_1, \tilde{Y}_2)$ and this reversal produces in the order of the imputation leads to less biased and more stable estimates of $\mathbb{E}(Y_3 | Y_1)$. For this reason, we believe that DRMI offers advantages over BR and RR in the non-Gaussian case. This is in addition to the

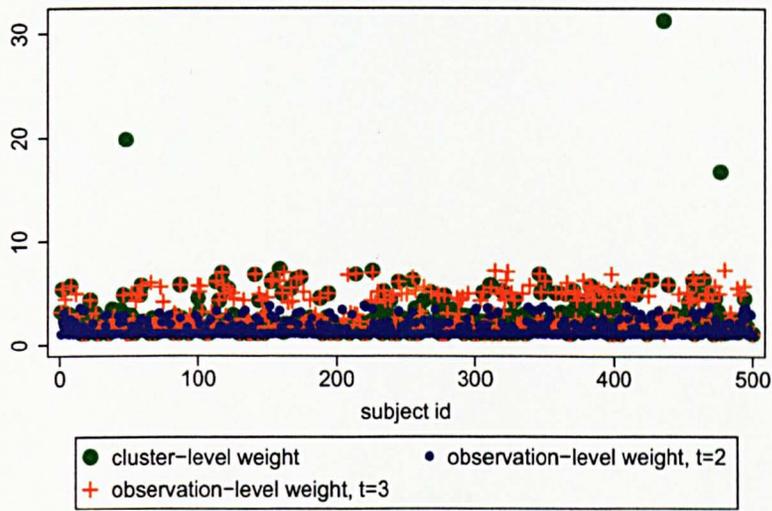


Figure 13.6: A comparison of the cluster-level and observation-level weights for the first simulation in the third set.

superiority of the convergence of the algorithm when compared with RR.

Part VI

Discussion

14

Discussion

14.1 Main conclusions

The main achievement of this thesis has been to show that multiple imputation (MI) can be used as a tool to obtain doubly robust (DR) estimators. This has built on the work done by Bang and Robins (2005), and, in our opinion, offers some advantages, such as an easily computable variance formula courtesy of Rubin's rules for MI. A larger advantage, as we showed in Chapter 9, is the conjectured extension to non-monotone

MAR data. In practice, however, the approach is limited to non-monotone longitudinal data, where a method for identifying the inverse probability weights exists. We have demonstrated, by repeating the simulation studies carried out by Bang and Robins (2005) that in settings where their method can also be applied, DRMI is only slightly inferior, as the theory would predict. A general method for obtaining DR estimators for MAR non-monotone data has not previously been proposed. In her recently published overview of inverse probability weighted methods in Fitzmaurice *et al.* (2008), Andrea Rotnitzky (under the heading *Discussion: A look into the future*) writes:

“in some models, such as CAR models with non-monotone data, doubly robust estimators could in principle be constructed, but their implementation is not clear.”

It is important to note, however, that a method for obtaining DR estimators in non-ignorable non-monotone longitudinal data has been proposed by Vansteelandt *et al.* (2007), and for reasons discussed on page 55 of this thesis, the mechanism considered by Vansteelandt *et al.* (2007) is usually more realistic in the non-monotone longitudinal setting than the RMM mechanism considered here.

In Chapters 12 and 13 we have shown that for binary data, DRMI outperforms the two existing methods for constructing DR methods in this setting. This is a consequence both of the computational power of MI, and the natural way in which it approximates conditional distributions when the data are not Gaussian. To estimate $\mathbb{E}(Y_3 | Y_1)$ using either of the methods proposed by Bang and Robins (2005) and Robins and Rotnitzky (1995), one must first estimate $\mathbb{E}(Y_3 | Y_1, Y_2)$ and then fit a linear regression to $\mathbb{E}(Y_3 | Y_1, Y_2)$ conditional on Y_1 . But since the predictions are not Gaussian, this leads to bias and instability in the estimates. In DRMI, however, $\mathbb{E}(Y_3 | Y_1)$ is estimated by first drawing binary imputations \tilde{Y}_2 from $\mathbb{E}(Y_2 | Y_1)$ and then fitting a logistic regression to $\mathbb{E}(Y_3 | Y_1, \tilde{Y}_2)$. Essentially, this reversal produces less biased and more stable estimates of $\mathbb{E}(Y_3 | Y_1)$.

14.2 Other conclusions

During the course of this thesis, we have also drawn several other conclusions. In Chapter 11 we demonstrated another unconventional use of MI, which is in sensitivity analyses. The basic principle here was introduced by Little and Yau (1996) and it exploits the fact that the imputation and analysis models need not be the same. By varying the imputation model, we can vary the assumptions made about the missing data mechanism. While Little and Yau (1996) concentrated on analyses within the intent-to-treat framework, we applied the same idea to per protocol analyses, where the additional potential violations of the ‘noncompliance at random’ (NAR) assumption were investigated. The conclusions as regards the findings of the glycaemia analysis from the RECORD study were that the original analysis, assuming multivariate normality, MAR and NAR, was reliable and not unduly sensitive to possible departures from the assumptions made.

In Chapter 13, we combined the results of Lipsitz *et al.* (2000) and Robins and Rotnitzky (1995) in order to derive conditions under which an unweighted GEE analysis gives consistent estimates under MAR. We also extended the result of Wang *et al.* (2007) to the monotone longitudinal case (for binary data, but the result automatically applies to any discrete-data GLM). In particular, we showed that in the longitudinal setting, the augmented estimating equation introduced by Robins and his colleagues gives numerically identical results to its non-augmented counterpart when the means model is saturated. This should not be surprising: double robustness and augmentation are intrinsically linked to the different smoothing implied by the different models. When these modelling assumptions are potentially incorrect, we gain robustness by protecting ourselves using two different sets of smoothing assumptions. When no such assumption is being made (as in the nonparametric case), there is no protection needed, and correspondingly no efficiency can be gained. In summary, we concluded that cluster-level weighting need never be used and that augmentation should only be contemplated when the means model is not saturated. For small samples, we concluded that even in saturated means models MI-GEE is more efficient than OWGEE, but that

as the sample size increases, the difference reverses and OWGEE becomes superior. In our simulations, a sample size of 5,000 was needed to see this reversal. When the means model is saturated, we see that MI-GEE is more efficient than OWGEE, but it is of course also less robust to model misspecification. In our simulation studies DRMI was only slightly less efficient than MI-GEE but it exhibited double robustness. For reasons outlined above, we believe that DRMI is the best and most practicably useable of the three DR estimators considered, and would recommend its use above other methods in this setting.

In Chapter 9, we also drew attention to a practical problem with the Bang and Robins (2005) method for constructing DR estimators for longitudinal data, namely that a model for $\mathbb{E}(Y_3 | Y_1)$ is required but might not be easily postulated.

14.3 Future work

Although DRMI has been shown to be a promising new approach, several limitations remain. First, in practice it is not possible to apply the method to non-monotone data except in the special case of longitudinal data, and even then, the claimed double-robustness has not been rigorously proved. If a method could be developed for calculating marginal inverse probability weights in general randomised monotone missingness mechanisms, DRMI estimators could be explored in this more general setting.

DRMI for general non-monotone patterns relies on MICE, a method which—although shown to be very effective in simulations—does not have a firm theoretical justification. Any developments in this area would be highly relevant to strengthen the theoretical justification for DRMI.

In this thesis, we have used MI in two different ways: to construct DR estimators and to carry out sensitivity analyses for parametric models. A possible extension would be to combine these two approaches and to use MI to construct sensitivity analyses

within a DR framework.

In the chapters on binary data, the conditions for the consistency and asymptotic efficiency of GEE were derived for any means model (saturated or otherwise), as was the comparison between cluster- and observation-level weighted GEE, but the other results derived related only to the case when the means model is saturated. This is a good starting point for understanding how these methods relate to each other, but more needs to be done on the comparisons in the non-saturated case.

Also, Chapters 12 and 13 considered only monotone missing data patterns. More work is needed on the comparison of methods for non-monotone incomplete binary data, but this work is likely to be mathematically more challenging.

A possible extension of the work on binary data is to consider the case when parameters are shared across timepoints. The theory in this case would be more complex, but potentially a two-stage process could be envisaged, where first models are fitted with distinct parameters at each timepoint and then, using a least squares or similar procedure, inference could be made about a suitable weighted average of these considered to approximate the shared parameter. Asymptotic properties of the different methods could then be derived using the two-stage approximation.

In our motivation for the work on binary data, we considered examples in which there were only three timepoints. As the number of timepoints increases, there is likely to be perfect prediction, the phenomenon in which estimated conditional probabilities (such as $\hat{P}(Y_{t,i} = 1 | \bar{Y}_{t-1,i})$) are either 0 or 1. This can cause problems, particularly in methods that use multiple imputation since the normal approximation to the Bayesian posterior distribution of the parameters becomes very poor in the extremes of the distribution. The latest version of `ice` in Stata incorporates a solution to this problem, but some further work to investigate how this influences the comparisons between the various methods would be useful.

Finally, this research has been largely confined to problems in missing data. Many

aspects of the problems described are more generally encountered in the field of causal inference. We believe that some of our proposed methodology, for example DRMI, could be adapted for use in this wider setting.

Bibliography

- Anderson, T. W. (1957) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, **52**(278), 200–203.
- Bahadur, R. R. (1961) A representation of the joint distribution of responses to n dichotomous items. Taken from: *Studies in Item Analysis and Prediction*, H. Solomon, ed., pp. 158–168. Stanford Mathematical Studies in the Social Sciences VI. Stanford, California: Stanford University Press.
- Bang, H. and Robins, J. M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**, 962–973.
- Beunckens, C., Sotto, C. and Molenberghs, G. (2008) A simulation study comparing weighted estimating equations with multiple imputation based estimating equations. *Computational Statistics & Data Analysis*, **52**(3), 1533–1548.
- Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.
- Brown, H. and Prescott, R. (2006) *Applied Mixed Models in Medicine*. Wiley, Chichester.
- van Buuren, S., Boshuizen, H. C. and Knook, D. L. (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, **18**, 681–694.

- van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. and Rubin, D. B. (2006) Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, **76**, 1049–1064.
- Carpenter, J., Kenward, M., Evans, S. and White, I. (2004) Last observation carry-forward and last observation analysis. *Statistics in Medicine*, **23**(20), 3241–3242.
- Carpenter, J. C. and Kenward, M. G. (2008) *Missing Data in Randomised Controlled Trials — a Practical Guide*. Birmingham: National Institute for Health Research, Publication RM03/JH17/MK. Available at http://www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml.
- Carpenter, J. C., Kenward, M. G. and Vansteelandt, S. (2006) A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **169**, 571–584.
- Charbonnel, B., Scherthaner, G., Brunetti, P., Matthews, D. R., Urquhart, R., Tan, M. H. and Hanefeld, M. (2005) Long-term efficacy and tolerability of add-on pioglitazone therapy to failing monotherapy compared with addition of gliclazide or metformin in patients with type 2 diabetes. *Diabetologia*, **48**(6), 1093–1104.
- Clayton, D. G. (1996) *Markov Chain Monte Carlo in Practice*, chapter Generalized linear mixed models, pp. 275–301. Chapman and Hall, London.
- Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. Chapman and Hall, London.
- Crowder, M. (1985) Gaussian estimation for correlated binomial data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **47**(2), 229–237.
- Crowder, M. (1992) Contribution to the discussion of “Multivariate regression analyses for categorical data” by Liang, Zeger and Qaqish. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **54**, 3–40.
- Dale, J. R. (1986) Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**(4), 909–917.

- Davidian, M., Tsiatis, A. A. and Leon, S. (2005) Semiparametric estimation of treatment effect in a pretest–posttest study with missing data (with discussion). *Statistical Science*, **20**, 261–301.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.
- Diggle, P. J., Heagerty, P., Liang, K. and Zeger, S. L. (2002) *Analysis of Longitudinal Data*. Oxford University Press, Oxford.
- Engel, B. and Keen, A. (1994) A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, **48**, 1–22.
- Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (eds) (2008) *Longitudinal Data Analysis*. Chapman and Hall/CRC.
- Fitzmaurice, G. M. and Laird, N. M. (1993) A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**(1), 141–151.
- Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S. R. (1995) Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(4), 691–704.
- Fitzmaurice, G. M., Lipsitz, S. R., Molenberghs, G. and Ibrahim, J. G. (2001) Bias in estimating association parameters for longitudinal binary responses with drop-outs. *Biometrics*, **57**(1), 15–21.
- Gelman, A. and Raghunathan, T. E. (2001) Using conditional distributions for missing-data imputation, in discussion of ‘Conditionally specified distributions’ by Arnold et al. *Statistical Science*, **3**, 268–269.
- Gill, R. and Robins, J. M. (1996) Missing at random from an algorithmic viewpoint. *Proceedings of the First Seattle Symposium on Survival Analysis*.
- Gill, R., van der Laan, M. and Robins, J. (1996) Coarsening at random: characterizations, conjectures and counterexamples. *Proceedings of the First Seattle Symposium on Survival Analysis*, pp. 255–294.

- Harville, D. A. (1997) *Matrix Algebra from a Statistician's Perspective*. Springer, New York.
- Heckman, J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, **5**, 475–492.
- Heitjan, D. F. and Rubin, D. B. (1991) Ignorability and coarse data. *The Annals of Statistics*, **19**(4), 2244–2253.
- Heyde, C. C. and Morton, R. (1998) Multiple roots in general estimating equations. *Biometrika*, **85**, 954–959.
- Home, P. D., Pocock, S. J., Beck-Nielsen, H., Gomis, R., Hanefeld, M., Dargic, H., Komajda, M., Gubb, J., Biwas, N. and Jones, N. P. (2005) Rosiglitazone Evaluated for Cardiac Outcomes and Regulation of glycaemia in Diabetes (RECORD): study design and protocol. *Diabetologia*, **48**, 1726–1735.
- Home, P. D., Jones, N. P., Pocock, S. J., Beck-Nielsen, H., Gomis, R., Hanefeld, M., Komajda, M. and Curtis, P. (2007) Rosiglitazone RECORD study: glucose control outcomes at 18 months. *Diabetic Medicine*, **24**, 626–634.
- Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**(260), 663–685.
- Kang, J. D. Y. and Schafer, J. L. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, **22**(4), 523–580.
- Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**(4), 963–974.
- Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with Discussion),. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 619–678.

- Li, X., Mehrotra, D. V. and Barnard, J. (2006) Analysis of incomplete longitudinal binary data using multiple imputation. *Statistics in Medicine*, **25**(12), 2107–2124.
- Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.
- Lipsitz, S. R., Molenberghs, G., Fitzmaurice, G. M. and Ibrahim, J. (2000) GEE with Gaussian estimation of the correlations when data are incomplete. *Biometrics*, **56**(2), 528–536.
- Little, R. and Yau, L. (1996) Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, **52**(4), 1324–1333.
- Little, R. J. and Rubin, D. B. (2000) Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, **21**, 121–45.
- Little, R. J. A. (1993) Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**(421), 125–134.
- Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*. Wiley, New York.
- McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. Chapman and Hall, London.
- Meng, X.-L. (1994) Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, **9**(4), 538–558.
- Molenberghs, G. and Kenward, M. G. (2007) *Missing Data in Clinical Studies*. Statistics in Practice. Wiley.
- Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data*. Springer, New York.
- Molenberghs, G., Michiels, B., Kenward, M. G. and Diggle, P. J. (1998) Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, **52**, 153–161.

- Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C. and Carroll, R. J. (2004) Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, **5**, 445–464.
- Molenberghs, G., Beunckens, C., Sotito, C. and Kenward, M. G. (2008) Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(2), 371–388.
- Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, **135**, 370–384.
- Newey, W. K. (1990) Semiparametric efficiency bounds. *Journal of Applied Econometrics*, **5**(2), 99–135.
- Nissen, S. E. and Wolski, K. (2007) Effect of Rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *The New England Journal of Medicine*, **356**(24), 2457–2471.
- O'Brien, L. M., Fitzmaurice, G. M. and Horton, N. J. (2006) Maximum likelihood estimation of marginal pair-wise associations with multiple source predictors. *Biometrical Journal*, **48**(5), 860–875.
- Orchard, T. and Woodbury, M. A. (1972) A missing information principle: theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics*, **1**, 697–715.
- Paik, M. C. (1997) The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, **92**(440), 1320–1329.
- Patterson, H. D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**(3), 545–554.
- Pocock, S. J. (1983) *Clinical Trials: A Practical Approach*. Wiley & sons.

- Preisser, J. S., Lohman, K. K. and Rathouz, P. J. (2002) Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, **21**(20), 3035–3054.
- Rasch, D. and Guiard, V. (2004) The robustness of parametric statistical methods. *Psychology Science*, **46**, 175–208.
- Reilly, M. and Pepe, M. S. (1995) A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, **82**(2), 299–314.
- Robins, J. M. (1999) Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran and D. Berry (eds), *Statistical Models in Epidemiology: The Environment and Clinical Trials*, IMA **116**, 95–134. New York: Springer-Verlag.
- Robins, J. M. and Gill, R. D. (1997) Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, **16**(1), 39–56.
- Robins, J. M. and Rotnitzky, A. (1992) Recovery of information and adjustment for dependent censoring using surrogate markers. In N. Jewell, K. Dietz and V. Farewell (eds), *AIDS Epidemiology—Methodological Issues*, 297–331. Boston, MA: Birkhäuser.
- Robins, J. M. and Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90**(429), 122–129.
- Robins, J. M. and Wang, N. (2000) Inference for imputation estimators. *Biometrika*, **87**(1), 113–124.
- Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**(427), 846–866.
- Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**(429), 106–121.

- Rotnitzky, A. and Robins, J. M. (1997) Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine*, **16**, 81–102.
- Rotnitzky, A. and Wypij, D. (1994) A note on the bias of estimators with missing data. *Biometrics*, **50**(4), 1163–1170.
- Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**(3), 581–592.
- Rubin, D. B. (1978) Multiple imputations in sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 20–34.
- Seber, G. A. F. (2008) *A Matrix Handbook for Statisticians*. Wiley.
- Shao, J. and Zhong, B. (2004) Last observation carry-forward and last observation analysis. *Statistics in Medicine*, **22**, 2429–2441.
- Smith, D. M. and Kenward, M. G. (2000) Extensions of multiple linear regression. *Communications in Statistics: Theory and Methods*, **29**(9), 2033–2053.
- Tsiatis, A. A. (2006) *Semiparametric Theory and Missing Data*. Springer, New York.
- Vansteelandt, S., Rotnitzky, A. and Robins, J. (2007) Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, **94**, 841–860.
- Verbeke, G. and Molenberghs, G. (1997) *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York.
- Wang, C. Y., Lee, S.-M. and Chao, E. C. (2007) Numerical equivalence of imputing scores and weighted estimators in regression analysis with missing covariates. *Biostatistics*, **8**(2), 468–473.
- Wang, N. and Robins, J. M. (1998) Large-sample theory for parametric multiple imputation procedures. *Biometrika*, **85**(4), 935–948.
- Williams, D. (1991) *Probability with Martingales*. Cambridge University Press, Cambridge.

- Wu, M. C. and Carroll, R. J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**(1), 175–188.
- Young, M. L., Preisser, J. S., Qaqish, B. F. and Wolfson, M. (2007) Comparison of subject-specific and population averaged models for count data from cluster-unit intervention trials. *Statistical Methods in Medical Research*, **16**, 167–184.
- Yun, S.-C., Lee, Y. and Kenward, M. G. (2007) Using hierarchical likelihood for missing data problems. *Biometrika*, **94**, 905–919.
- Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.



Proofs omitted from the main text

A.1 Proof of Lemma 8.1

The estimating equation for the j th imputed dataset in ordinary MI is

$$\sum_{i=1}^n \mathbf{S}_{\boldsymbol{\theta}}^F \left(\mathbf{z}_{ij}, \hat{\boldsymbol{\theta}}_{nj}^{\text{ord}} \right) = \mathbf{0}$$

but for robust MI is

$$\sum_{i=1}^n \left(\frac{R_i}{\pi_i} \mathbf{S}_{\theta}^F (\mathbf{z}_i^*, \hat{\theta}_j^{*\text{rob}}) + \left(1 - \frac{R_i}{\pi_i} \right) \mathbf{S}_{\theta}^F \left\{ \mathbf{z}_{ij}^* [\hat{\theta}_I^{(j)}], \hat{\theta}_j^{*\text{rob}} \right\} \right) = 0$$

which we can re-write as

$$\sum_{i=1}^n \left[\mathbf{S}_{\theta}^F \left\{ \mathbf{z}_{ij}^* [\hat{\theta}_I^{(j)}], \hat{\theta}_j^{*\text{rob}} \right\} + \left(\frac{1 - \pi_i}{\pi_i} \right) \left(\mathbf{S}_{\theta}^F \left\{ \mathbf{z}_{ij}^* [\hat{\theta}_I^{(j)}], \hat{\theta}_j^{*\text{rob}} \right\} - \mathbf{S}_{\theta}^F \left\{ \tilde{\mathbf{z}}_{ij}^* [\hat{\theta}_I^{(j)}], \hat{\theta}_j^{*\text{rob}} \right\} \right) \right] = 0$$

where $\mathbf{z}_{ij}^* [\hat{\theta}_I^{(j)}]$ and $\tilde{\mathbf{z}}_{ij}^* [\hat{\theta}_I^{(j)}]$ are as defined above.

Expanding in a Taylor series about θ_0 :

$$\begin{aligned} & \sum_{i=1}^n \left[\mathbf{S}_{\theta}^F \left\{ \mathbf{z}_{ij}^* [\hat{\theta}_I^{(j)}], \theta_0 \right\} + \left(\frac{1 - \pi_i}{\pi_i} \right) \left(\mathbf{S}_{\theta}^F \left\{ \mathbf{z}_{ij}^* [\hat{\theta}_I^{(j)}], \theta_0 \right\} - \mathbf{S}_{\theta}^F \left\{ \tilde{\mathbf{z}}_{ij}^* [\hat{\theta}_I^{(j)}], \theta_0 \right\} \right) \right. \\ & \quad \left. + \frac{\partial \mathbf{S}_{\theta}^F [\mathbf{z}_{ij}^* (\theta_0), \theta_0]}{\partial \theta^T} (\hat{\theta}_j^{*\text{rob}} - \theta_0) \right. \\ & \left. + \left(\frac{1 - \pi_i}{\pi_i} \right) \underbrace{\left\{ \frac{\partial \mathbf{S}_{\theta}^F [\mathbf{z}_{ij}^* (\theta_0), \theta_0]}{\partial \theta^T} - \frac{\partial \mathbf{S}_{\theta}^F [\tilde{\mathbf{z}}_{ij}^* (\theta_0), \theta_0]}{\partial \theta^T} \right\}}_{=0} (\hat{\theta}_j^{*\text{rob}} - \theta_0) \right] + o_p \left(n^{-\frac{1}{2}} \right) = 0 \end{aligned}$$

$$\begin{aligned} \Rightarrow & \sum_{i=1}^n \left[\mathbf{S}_{\theta}^F \left\{ \mathbf{z}_{ij}^* [\hat{\theta}_I^{(j)}], \theta_0 \right\} + \left(\frac{1 - \pi_i}{\pi_i} \right) \left(\mathbf{S}_{\theta}^F \left\{ \mathbf{z}_{ij}^* [\hat{\theta}_I^{(j)}], \theta_0 \right\} - \mathbf{S}_{\theta}^F \left\{ \tilde{\mathbf{z}}_{ij}^* [\hat{\theta}_I^{(j)}], \theta_0 \right\} \right) \right] \\ & = \sum_{i=1}^n - \frac{\partial \mathbf{S}_{\theta}^F [\mathbf{z}_{ij}^* (\theta_0), \theta_0]}{\partial \theta^T} (\hat{\theta}_j^{*\text{rob}} - \theta_0) + o_p \left(n^{-\frac{1}{2}} \right) \\ & = n \mathbb{E} \left\{ - \frac{\partial \mathbf{S}_{\theta}^F [\mathbf{z}_{ij}^* (\theta_0), \theta_0]}{\partial \theta^T} \right\} (\hat{\theta}_j^{*\text{rob}} - \theta_0) + o_p \left(n^{-\frac{1}{2}} \right) \end{aligned}$$

$$= nI_{\theta\theta}^F(\theta_0) \left(\hat{\theta}_j^{*rob} - \theta_0 \right) + o_p \left(n^{-\frac{1}{2}} \right)$$

$$\begin{aligned} \therefore n^{\frac{1}{2}} \left(\hat{\theta}_j^{*rob} - \theta_0 \right) &= n^{-\frac{1}{2}} \sum_{i=1}^n [I_{\theta\theta}^F(\theta_0)]^{-1} \left[\mathbf{S}_{\theta}^F \left\{ \mathbf{Z}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} \right. \\ &\quad \left. + \left(\frac{1 - \pi_i}{\pi_i} \right) \left(\mathbf{S}_{\theta}^F \left\{ \mathbf{Z}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} - \mathbf{S}_{\theta}^F \left\{ \tilde{\mathbf{Z}}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} \right) \right] + o_p(1) \end{aligned}$$

$$\begin{aligned} \therefore n^{\frac{1}{2}} \left(\hat{\theta}^{*rob} - \theta_0 \right) &= n^{-\frac{1}{2}} \sum_{i=1}^n [I_{\theta\theta}^F(\theta_0)]^{-1} \left[m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left\{ \mathbf{Z}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} \right. \\ &\quad \left. + \left(\frac{1 - \pi_i}{\pi_i} \right) \left(m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left\{ \mathbf{Z}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} - m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left\{ \tilde{\mathbf{Z}}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} \right) \right] + o_p(1) \\ &= n^{-\frac{1}{2}} \sum_{i=1}^n [I_{\theta\theta}^F(\theta_0)]^{-1} \left[m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\mathbf{Z}_{ij}^*(\theta_0), \theta_0 \right] \right. \\ &\quad \left. + \left(\frac{1 - \pi_i}{\pi_i} \right) \left\{ m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\mathbf{Z}_{ij}^*(\theta_0), \theta_0 \right] - m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\tilde{\mathbf{Z}}_{ij}^*(\theta_0), \theta_0 \right] \right\} \right. \\ &\quad \left. + m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left\{ \mathbf{Z}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} \right. \\ &\quad \left. + \left(\frac{1 - \pi_i}{\pi_i} \right) \left(m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left\{ \mathbf{Z}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} - m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left\{ \tilde{\mathbf{Z}}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} \right) \right. \\ &\quad \left. - m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\mathbf{Z}_{ij}^*(\theta_0), \theta_0 \right] \right. \\ &\quad \left. - \left(\frac{1 - \pi_i}{\pi_i} \right) \left\{ m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\mathbf{Z}_{ij}^*(\theta_0), \theta_0 \right] - m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\tilde{\mathbf{Z}}_{ij}^*(\theta_0), \theta_0 \right] \right\} \right] + o_p(1) \end{aligned}$$

Now, if $R_i = 1$,

$$\mathbf{Z}_{ij}^* \left[\hat{\boldsymbol{\theta}}_I^{(j)} \right] = \mathbf{Z}_{ij}^* (\boldsymbol{\theta}_0)$$

and if $R_i = 0$,

$$\mathbf{Z}_{ij}^* \left[\hat{\boldsymbol{\theta}}_I^{(j)} \right] = \tilde{\mathbf{Z}}_{ij}^* \left[\hat{\boldsymbol{\theta}}_I^{(j)} \right]$$

and

$$\mathbf{Z}_{ij}^* (\boldsymbol{\theta}_0) = \tilde{\mathbf{Z}}_{ij}^* (\boldsymbol{\theta}_0)$$

$$\begin{aligned} \therefore n^{\frac{1}{2}} \left(\hat{\boldsymbol{\theta}}^{*\text{rob}} - \boldsymbol{\theta}_0 \right) &= n^{-\frac{1}{2}} \sum_{i=1}^n \left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F (\boldsymbol{\theta}_0) \right]^{-1} \left[m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\mathbf{Z}_{ij}^* (\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right] \right. \\ &+ \left. \left(\frac{1 - \pi_i}{\pi_i} \right) \left\{ m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\mathbf{Z}_{ij}^* (\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right] - m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\tilde{\mathbf{Z}}_{ij}^* (\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right] \right\} \right. \\ &+ m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left\{ \mathbf{Z}_{ij}^* \left[\hat{\boldsymbol{\theta}}_I^{(j)} \right], \boldsymbol{\theta}_0 \right\} \\ &- m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\mathbf{Z}_{ij}^* (\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right] \\ &\left. - \left(\frac{1 - \pi_i}{\pi_i} \right) \left(m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left\{ \tilde{\mathbf{Z}}_{ij}^* \left[\hat{\boldsymbol{\theta}}_I^{(j)} \right], \boldsymbol{\theta}_0 \right\} - m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left\{ \tilde{\mathbf{Z}}_{ij}^* \left[\boldsymbol{\theta}_0 \right], \boldsymbol{\theta}_0 \right\} \right) \right] + o_p(1) \end{aligned}$$

□

A.2 Proof of Lemma 8.2

We know (see Tsiatis, 2006, Theorem 14.3 on p.350) that, for ordinary improper MI

$$\begin{aligned} & n^{-\frac{1}{2}} \sum_{i=1}^n \left(m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left\{ \mathbf{Z}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} - m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\mathbf{Z}_{ij}^* \left(\theta_0 \right), \theta_0 \right] \right) \\ &= n^{-\frac{1}{2}} \sum_{i=1}^n \left[I_{\theta\theta}^F \left(\theta_0 \right) - I_{\theta\theta} \left(\theta_0 \right) \right] q \left[R_i, G_{R_i} \left(\mathbf{Z}_i \right) \right] + o_p \left(1 \right) \end{aligned} \tag{A.2.1}$$

and, provided the response probabilities, π_i , are bounded away from zero, for robust improper MI

$$\begin{aligned} & n^{-\frac{1}{2}} \sum_{i=1}^n \left(\frac{1 - \pi_i}{\pi_i} \right) \left(m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left\{ \tilde{\mathbf{Z}}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} - m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\tilde{\mathbf{Z}}_{ij}^* \left(\theta_0 \right), \theta_0 \right] \right) \\ &= n^{-\frac{1}{2}} \overline{\left(\frac{1 - \pi_i}{\pi_i} \right)} \sum_{i=1}^n \left[I_{\theta\theta}^F \left(\theta_0 \right) - I_{\theta\theta}^U \left(\theta_0 \right) \right] q \left[R_i, G_{R_i} \left(\mathbf{Z}_i \right) \right] + o_p \left(1 \right) \end{aligned} \tag{A.2.2}$$

where $\overline{\left(\frac{1 - \pi_i}{\pi_i} \right)}$ is the mean of $\left(\frac{1 - \pi_i}{\pi_i} \right)$ over all i , and

$$I_{\theta\theta}^U \left(\theta \right) = \mathbb{E} \left\{ - \frac{\partial \left[\mathbf{S}_{\theta}^U \left(\mathbf{Z}, \theta \right)^T \right]}{\partial \theta} \right\}$$

where

$$\mathbf{S}_{\theta}^U \left(\mathbf{Z}, \theta \right) = \mathbb{E} \left(\mathbf{S}_{\theta}^F \mid Y \right)$$

Thus,

$$\begin{aligned}
 n^{\frac{1}{2}} \left(\hat{\boldsymbol{\theta}}^{*rob} - \boldsymbol{\theta}_0 \right) &= n^{-\frac{1}{2}} \sum_{i=1}^n \left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) \right]^{-1} \left(m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\mathbf{Z}_{ij}^*(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right] \right. \\
 &+ \left. \left(\frac{1 - \pi_i}{\pi_i} \right) \left\{ m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\mathbf{Z}_{ij}^*(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right] - m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\tilde{\mathbf{Z}}_{ij}^*(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right] \right\} \right. \\
 &+ \left. \left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) - I_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0) \right] q \left[R_i, G_{R_i}(\mathbf{Z}_i) \right] \right. \\
 &\left. - \left(\frac{1 - \pi_i}{\pi_i} \right) \left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) - I_{\boldsymbol{\theta}\boldsymbol{\theta}}^U(\boldsymbol{\theta}_0) \right] q \left[R_i, G_{R_i}(\mathbf{Z}_i) \right] \right) + o_p(1)
 \end{aligned}$$

This means that the i th influence function for the robust improper multiple imputation estimator is:

$$\begin{aligned}
 &\left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) \right]^{-1} \left(m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\mathbf{Z}_{ij}^*(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right] \right. \\
 &+ \left. \left(\frac{1 - \pi_i}{\pi_i} \right) \left\{ m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\mathbf{Z}_{ij}^*(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right] - m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\tilde{\mathbf{Z}}_{ij}^*(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right] \right\} \right. \\
 &+ \left. \left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) - I_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0) \right] q \left[R_i, G_{R_i}(\mathbf{Z}_i) \right] - \left(\frac{1 - \pi_i}{\pi_i} \right) \left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) - I_{\boldsymbol{\theta}\boldsymbol{\theta}}^U(\boldsymbol{\theta}_0) \right] q \left[R_i, G_{R_i}(\mathbf{Z}_i) \right] \right) \\
 &= \left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) \right]^{-1} \left\{ \overbrace{\frac{1}{\pi_i} m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\mathbf{Z}_{ij}^*(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right]}^{\text{Term1}} - \overbrace{\left(\frac{1 - \pi_i}{\pi_i} \right) m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\tilde{\mathbf{Z}}_{ij}^*(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right]}^{\text{Term2}} \right. \\
 &\left. + \underbrace{\left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) - I_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0) \right] q \left[R_i, G_{R_i}(\mathbf{Z}_i) \right]}_{\text{Term3}} - \underbrace{\left(\frac{1 - \pi_i}{\pi_i} \right) \left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F(\boldsymbol{\theta}_0) - I_{\boldsymbol{\theta}\boldsymbol{\theta}}^U(\boldsymbol{\theta}_0) \right] q \left[R_i, G_{R_i}(\mathbf{Z}_i) \right]}_{\text{Term4}} \right\}
 \end{aligned}$$

(A.2.3)

□

A.3 Proof of Lemma 8.3

The variance of the i th influence function (A.2.3) is given by

$$\begin{aligned}
 & [I_{\theta\theta}^F(\theta_0)]^{-1} [\text{Var}(\text{Term1}) + \text{Var}(\text{Term2}) + \text{Var}(\text{Term3}) + \text{Var}(\text{Term4}) \\
 & - \mathbb{E}(\text{Term1} \cdot \text{Term2}^T) + \mathbb{E}(\text{Term1} \cdot \text{Term3}^T) - \mathbb{E}(\text{Term1} \cdot \text{Term4}^T) \\
 & - \mathbb{E}(\text{Term2} \cdot \text{Term1}^T) - \mathbb{E}(\text{Term2} \cdot \text{Term3}^T) + \mathbb{E}(\text{Term2} \cdot \text{Term4}^T) \\
 & + \mathbb{E}(\text{Term3} \cdot \text{Term1}^T) - \mathbb{E}(\text{Term3} \cdot \text{Term2}^T) - \mathbb{E}(\text{Term3} \cdot \text{Term4}^T) \\
 & - \mathbb{E}(\text{Term4} \cdot \text{Term1}^T) + \mathbb{E}(\text{Term4} \cdot \text{Term2}^T) - \mathbb{E}(\text{Term4} \cdot \text{Term3}^T)] [I_{\theta\theta}^F(\theta_0)]^{-1}
 \end{aligned}$$

Some of these are evaluated by Tsiatis (2006) (pp. 355-357), and the others follow by similar arguments, giving the required expression for the variance of the i th influence function.

□

A.4 Proof of Lemma 8.5

Tsiatis (2006) shows that if the initial estimator is proper, (A.2.1) becomes

$$\begin{aligned}
 & n^{-\frac{1}{2}} \sum_{i=1}^n \left(m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left\{ \mathbf{Z}_{ij}^* \left[\hat{\theta}_I^{(j)} \right], \theta_0 \right\} - m^{-1} \sum_{j=1}^m \mathbf{S}_{\theta}^F \left[\mathbf{Z}_{ij}^* (\theta_0), \theta_0 \right] \right) \\
 & = n^{-\frac{1}{2}} \sum_{i=1}^n [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)] q[R_i, G_{R_i}(\mathbf{Z}_i)] \\
 & + m^{-1} \sum_{j=1}^m [I_{\theta\theta}^F(\theta_0) - I_{\theta\theta}(\theta_0)] n^{\frac{1}{2}} \left[\hat{\theta}_I^{(j)} - \hat{\theta}_I^{\text{improper}} \right] + o_p(1) \tag{A.4.1}
 \end{aligned}$$

Analogously, (A.2.2) becomes

$$\begin{aligned}
 & n^{-\frac{1}{2}} \sum_{i=1}^n \left(\frac{1 - \pi_i}{\pi_i} \right) \left(m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left\{ \tilde{\mathbf{Z}}_{ij}^* \left[\hat{\boldsymbol{\theta}}_I^{(j)} \right], \boldsymbol{\theta}_0 \right\} - m^{-1} \sum_{j=1}^m \mathbf{S}_{\boldsymbol{\theta}}^F \left[\tilde{\mathbf{Z}}_{ij}^* (\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \right] \right) \\
 &= n^{-\frac{1}{2}} \overline{\left(\frac{1 - \pi_i}{\pi_i} \right)} \sum_{i=1}^n \left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F (\boldsymbol{\theta}_0) - I_{\boldsymbol{\theta}\boldsymbol{\theta}}^U (\boldsymbol{\theta}_0) \right] q [R_i, G_{R_i} (\mathbf{Z}_i)] \\
 &+ m^{-1} \overline{\left(\frac{1 - \pi_i}{\pi_i} \right)} \sum_{j=1}^m \left[I_{\boldsymbol{\theta}\boldsymbol{\theta}}^F (\boldsymbol{\theta}_0) - I_{\boldsymbol{\theta}\boldsymbol{\theta}}^U (\boldsymbol{\theta}_0) \right] n^{\frac{1}{2}} \left[\hat{\boldsymbol{\theta}}_I^{(j)} - \hat{\boldsymbol{\theta}}_I^{\text{improper}} \right] + o_p(1) \quad (\text{A.4.2})
 \end{aligned}$$

Since the variation of the j th initial estimate of $\boldsymbol{\theta}$ about its mean is independent of the first term (in both (A.4.1) and (A.4.2)), the variance of $n^{\frac{1}{2}} \left(\hat{\boldsymbol{\theta}}^{*\text{rob}} - \boldsymbol{\theta}_0 \right)$ for robust proper MI is as required.

□

B

Further tables and figures

The tables and figures excluded from the main text are given here.

Model	Background Treatment			
	Met		Su	
	[M+R]-[M+S]	SE	[S+R]-[S+M]	SE
MAR PP	0.087	0.08	0.066	0.08
MNAR ITT, $(\gamma_1, \gamma_2, \gamma_3) =$				
(0,0,0)	0.068	0.08	-0.067	0.07
(-0.25,0.25,0)	0.101	0.08	-0.121	0.08
(-0.5,0.5,0)	0.133	0.09	-0.173	0.10
(-1,1,0)	0.197	0.13	-0.280	0.16
(0.25,-0.25,0)	0.036	0.08	-0.013	0.08
(0.5,-0.5,0)	0.004	0.09	0.041	0.10
(1,-1,0)	-0.060	0.12	0.149	0.16
(0.4,0.2,0)	0.090	0.08	-0.110	0.08
MNAR/NNAR PP, $(\delta_1, \delta_2, \delta_3, \delta_4) =$				
(0,0,0,0)	0.078	0.09	0.029	0.08
(0.25,0.25,0.25,0.25)	0.100	0.08	-0.003	0.08
(0.5,0.5,0.5,0.5)	0.112	0.08	0.003	0.08
(1,1,1,1)	0.127	0.09	-0.031	0.08
(0.25,-0.25,-0.25,0.25)	0.107	0.08	0.020	0.07
(0.5,-0.5,-0.5,0.5)	0.112	0.09	-0.005	0.08
(1,-1,-1,1)	0.126	0.09	-0.034	0.09
(-0.25,0.25,0.25,-0.25)	0.097	0.08	0.054	0.09
(-0.5,0.5,0.5,-0.5)	0.088	0.08	0.043	0.08
(-1,1,1,-1)	0.074	0.08	0.076	0.08
(-0.25,-0.25,-0.25,-0.25)	0.097	0.08	0.049	0.08
(-0.5,-0.5,-0.5,-0.5)	0.092	0.08	0.033	0.07
(-1,-1,-1,-1)	0.096	0.08	0.099	0.08
(3,3,3,3)	0.179	0.12	-0.139	0.12

Table B.1: Estimates and SEs of the treatment difference (change in HbA_{1c} from baseline to 18 months) between Met+Rosi and Met+Su, and Su+Rosi and Su+Met, respectively, for each of the models considered.

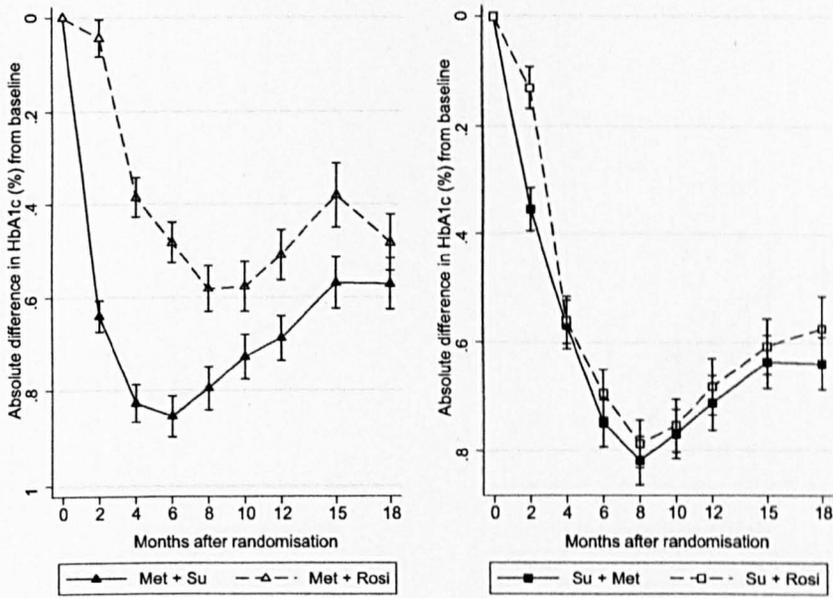


Figure B.1: The profiles (mean \pm SE) implied by the MAR per protocol analysis

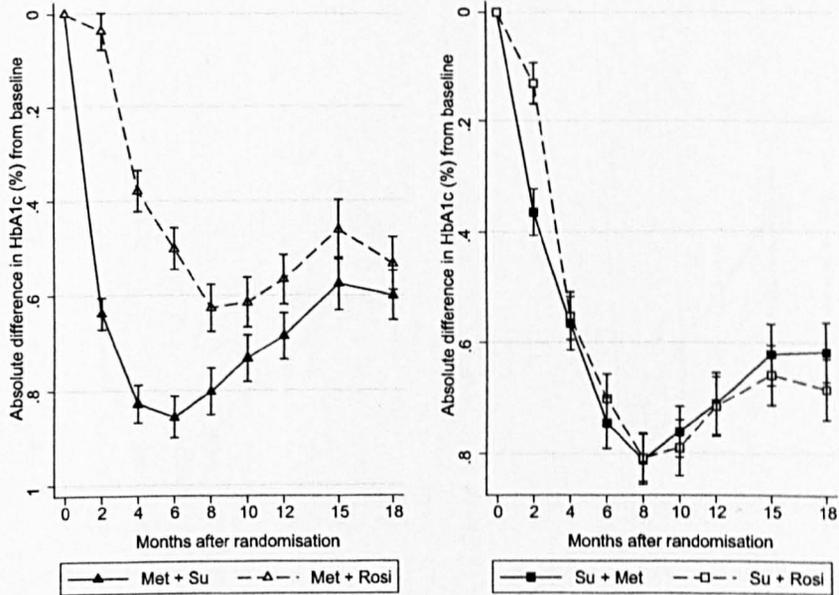


Figure B.2: The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (0, 0, 0)$

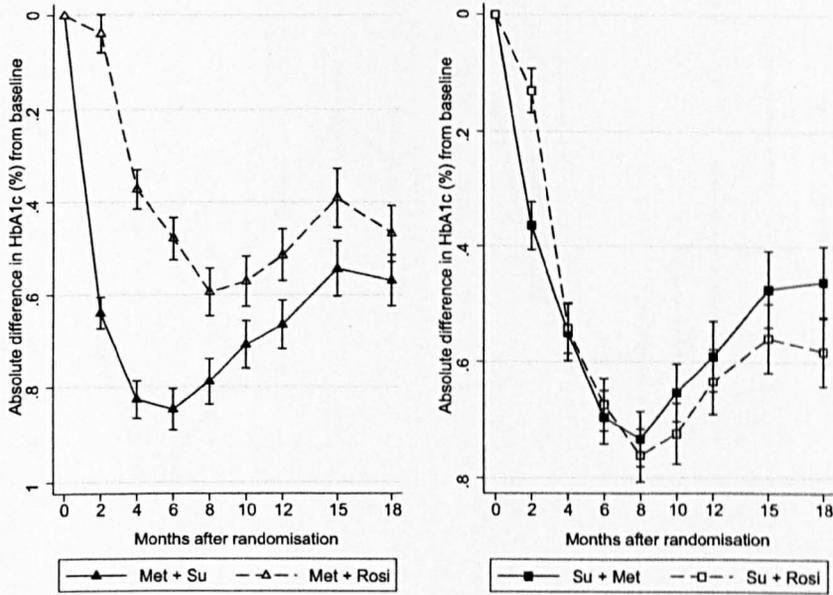


Figure B.3: The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (-0.25, 0.25, 0)$

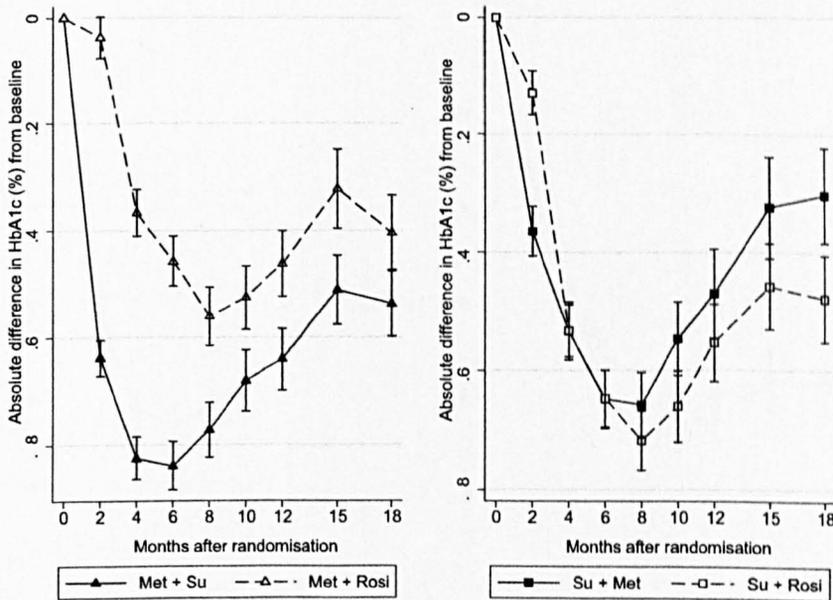


Figure B.4: The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (-0.5, 0.5, 0)$

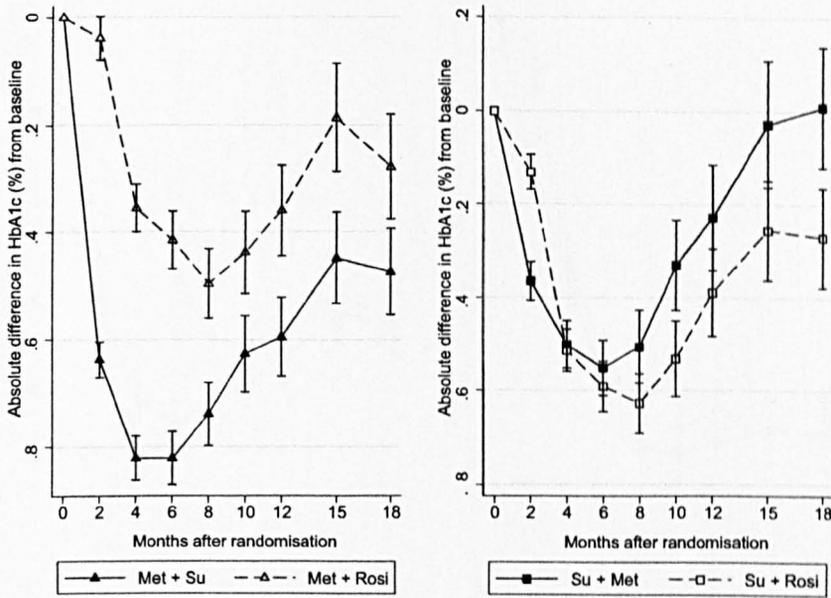


Figure B.5: The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (-1, 1, 0)$

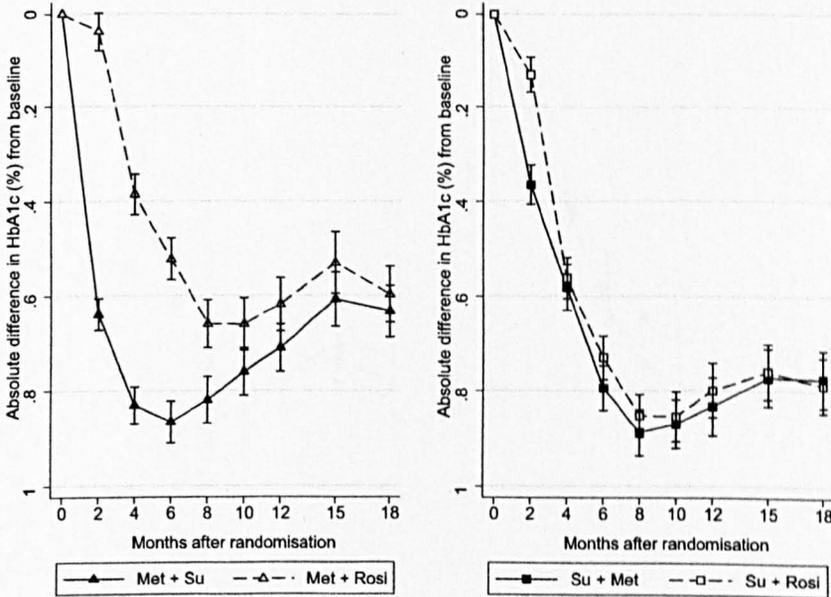


Figure B.6: The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (0.25, -0.25, 0)$

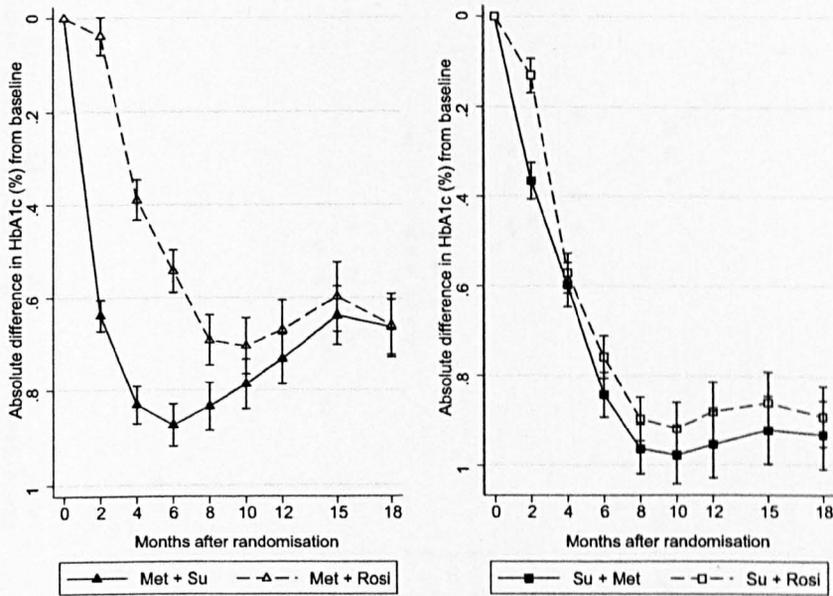


Figure B.7: The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (0.5, -0.5, 0)$

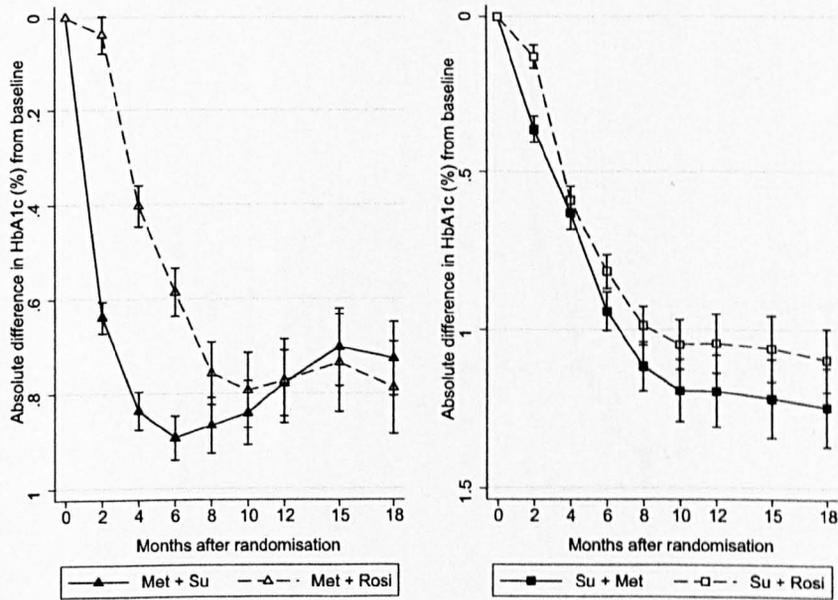


Figure B.8: The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (1, -1, 0)$

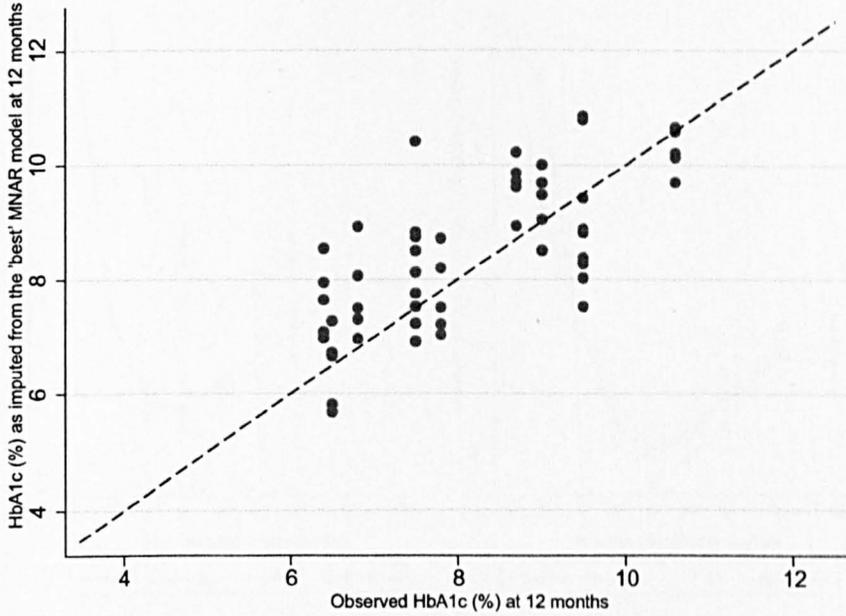


Figure B.9: HbA_{1c} at the 12-month timepoint: imputed vs. observed for the “best” combination, $(\gamma_1, \gamma_2, \gamma_3) = (0.4, 0.2, 0)$

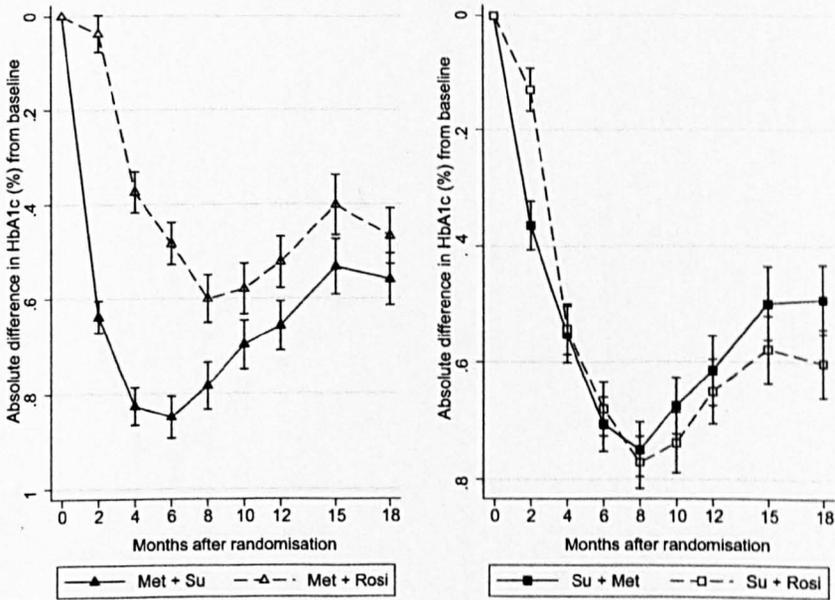


Figure B.10: The profiles (mean \pm SE) implied by the MNAR ITT analysis with $(\gamma_1, \gamma_2, \gamma_3) = (0.4, 0.2, 0)$

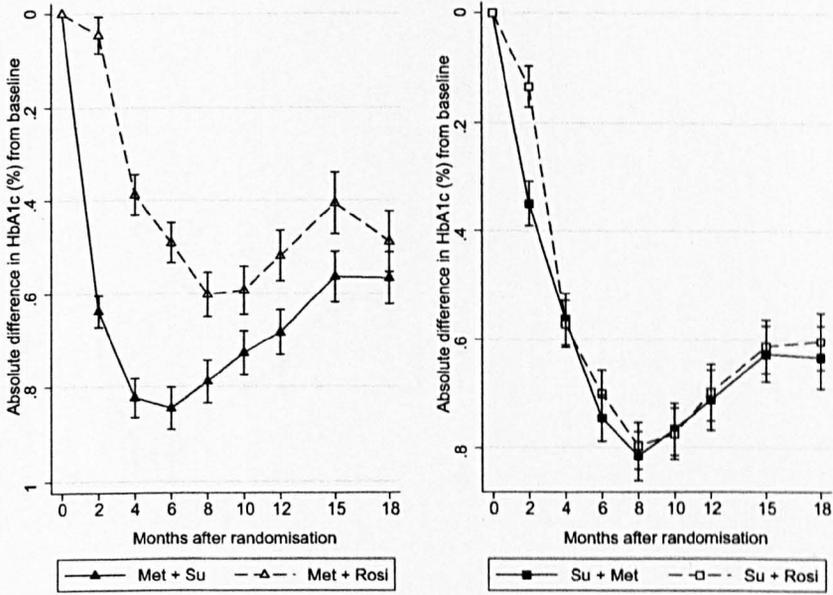


Figure B.11: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (0, 0, 0, 0)$

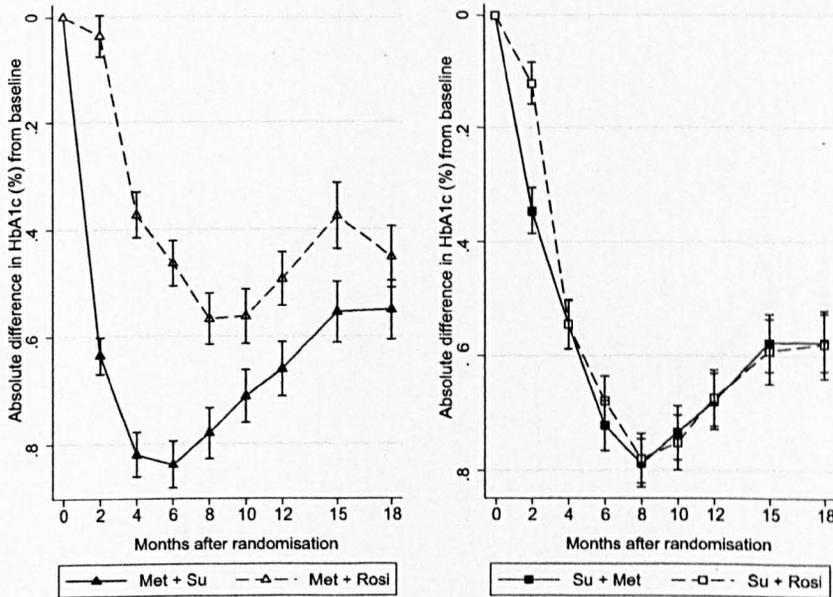


Figure B.12: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (0.25, 0.25, 0.25, 0.25)$

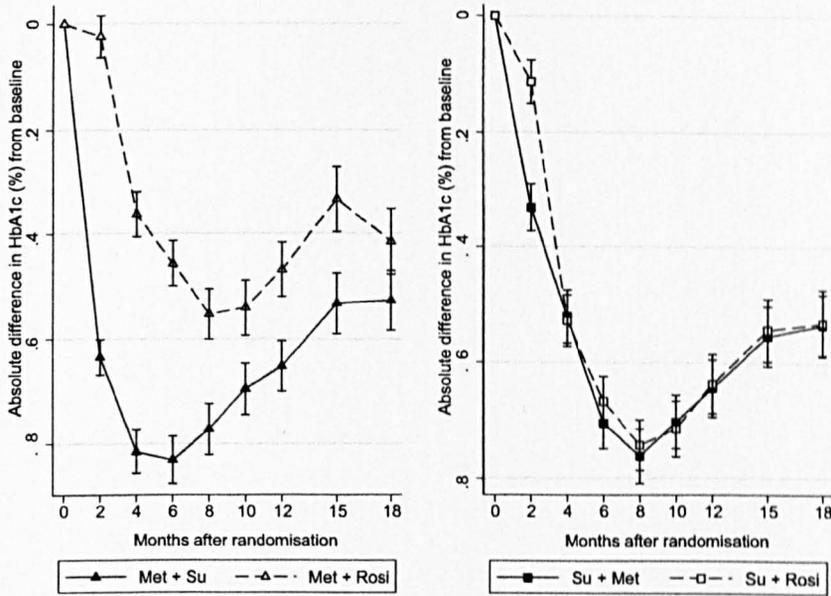


Figure B.13: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (0.5, 0.5, 0.5, 0.5)$

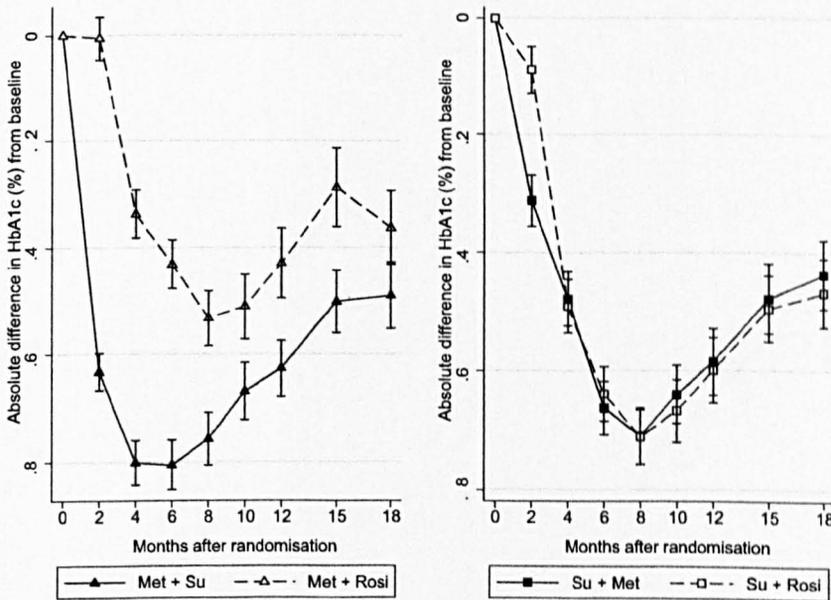


Figure B.14: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (1, 1, 1, 1)$

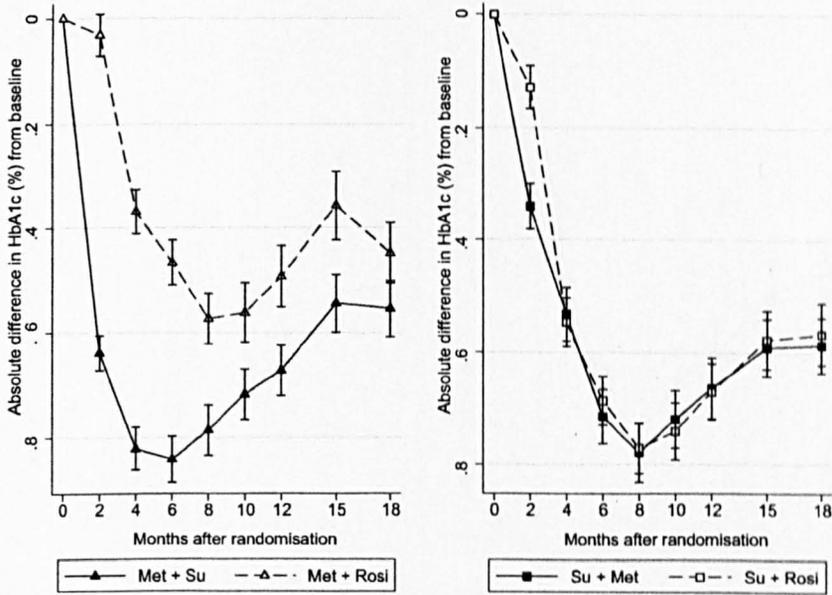


Figure B.15: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (0.25, -0.25, -0.25, 0.25)$

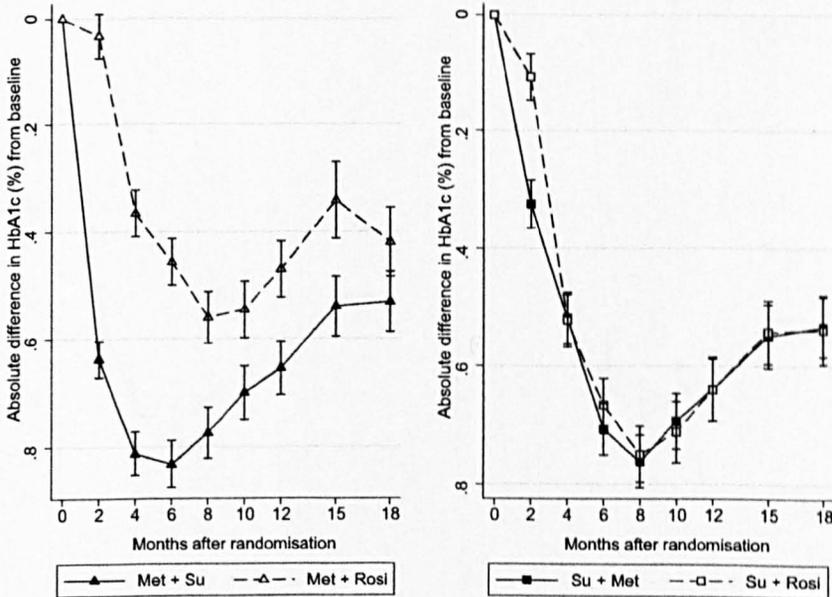


Figure B.16: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (0.5, -0.5, -0.5, 0.5)$

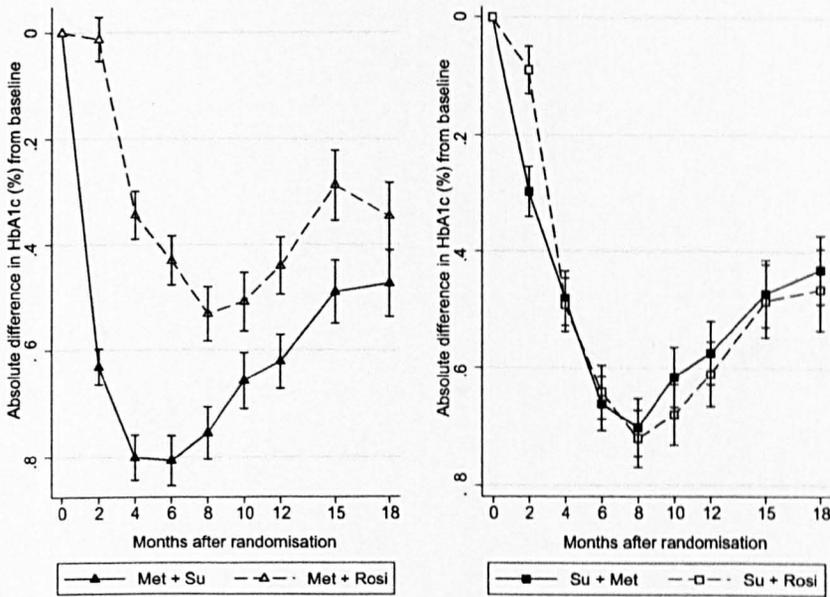


Figure B.17: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (1, -1, -1, 1)$

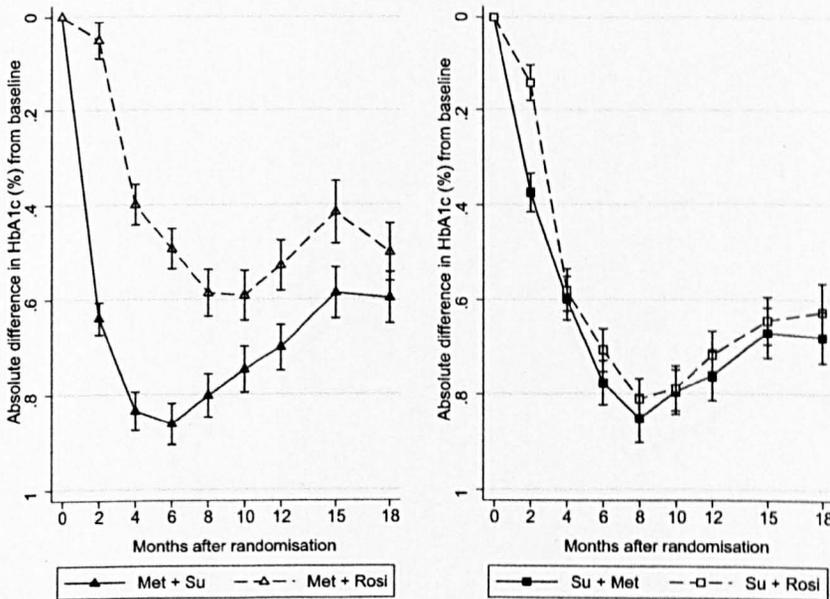


Figure B.18: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-0.25, 0.25, 0.25, -0.25)$

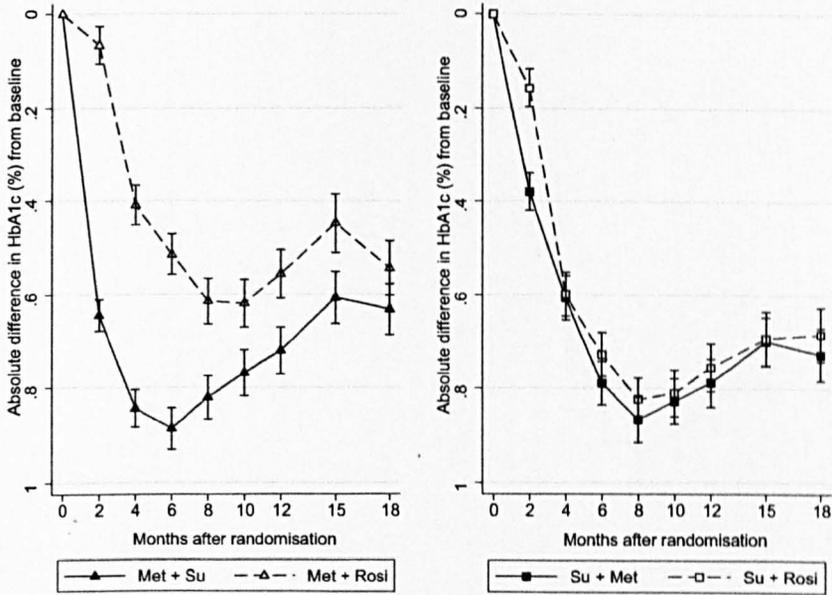


Figure B.19: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-0.5, 0.5, 0.5, -0.5)$

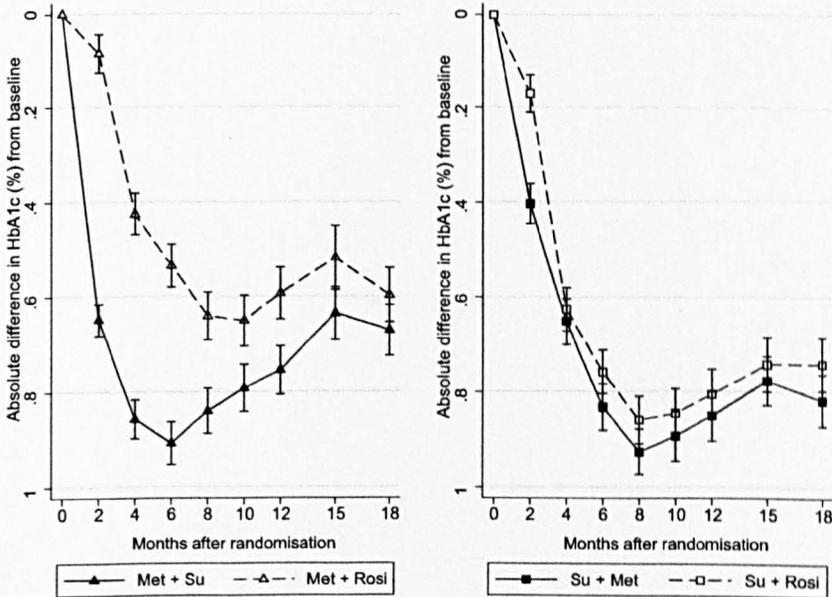


Figure B.20: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-1, 1, 1, -1)$

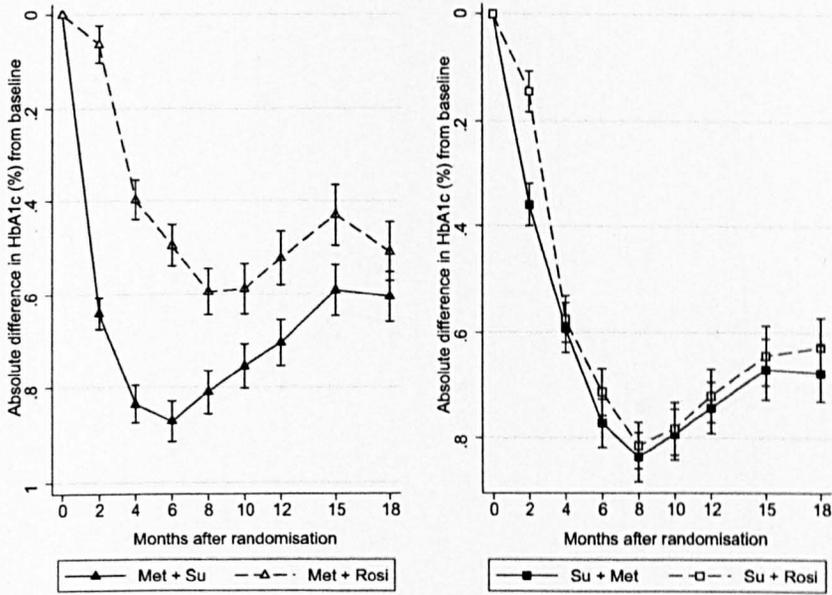


Figure B.21: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-0.25, -0.25, -0.25, -0.25)$

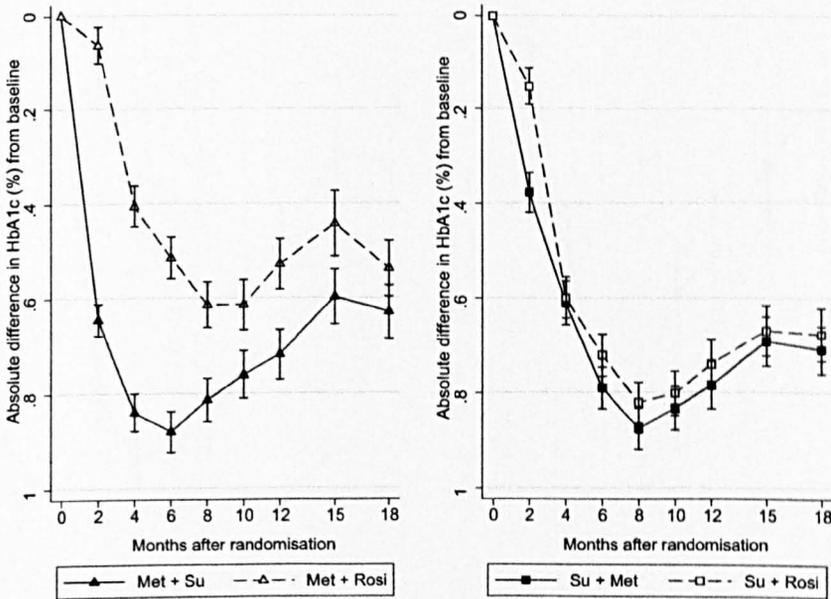


Figure B.22: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-0.5, -0.5, -0.5, -0.5)$

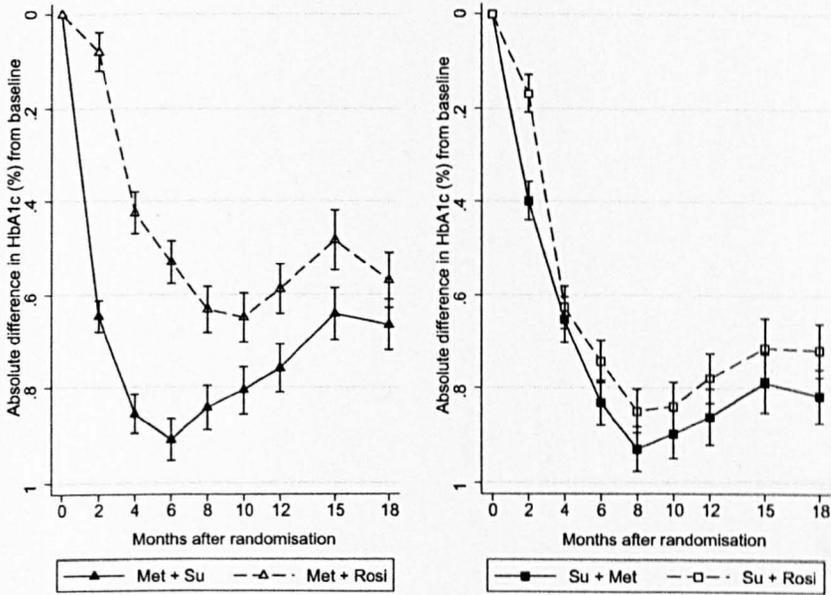


Figure B.23: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (-1, -1, -1, -1)$

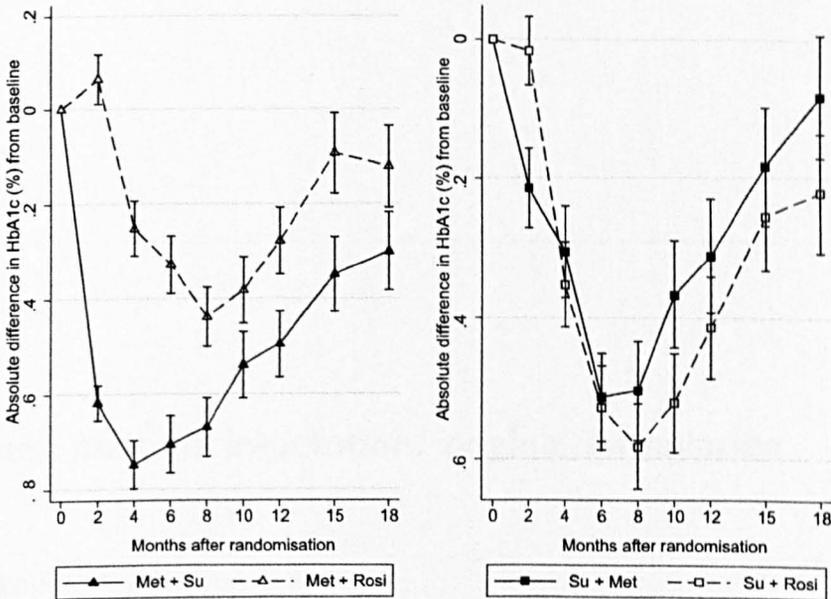


Figure B.24: The profiles (mean \pm SE) implied by the MNAR/NNAR PP analysis with $(\delta_1, \delta_2, \delta_3, \delta_4) = (3, 3, 3, 3)$

C

Computer code

C.1 Robust multiple imputation: original formulation

C.1.1 Improper

```

* Assume we have y and x, both full, and xmis, the      *
* observed portion of x, saved in a file called        *
* 'incomplete'. Also in the file 'incomplete' are     *
* the (known) probabilities, pi, and the missingness  *
* indicator, R.                                       *
* obs is a local macro containing the sample size of  *
* the full data and imps is a global macro containing *
* the number of imputed datasets                      *
*****

```

```

qui drop if xmis==.
qui drop xmis
qui save top, replace
use incomplete, clear
egen w0=mean(pi)
egen w1=mean(1/pi)
egen w2=mean((1-pi)/pi)
egen w3=mean(1/(pi^2))
egen w4=mean(((1-pi)/(pi^2)))
egen w5=mean((((1-pi)^2)/(pi^2))
qui gen w=1/pi
qui replace w=1 if xmis==.
qui replace w=1-w if xmis!=.
qui drop x xmis
qui gen x=.
qui save bottom, replace
use top, clear
append using bottom
qui gen Rstar=(x!=.)
qui gen S=1-R
qui gen Sstar=1-Rstar
sort Sstar S, stable
qui drop S Sstar
qui gen cons=1
qui regress x [pw=w]
local muX=_b[_cons]
local sXX=e(rmse)^2
qui regress y_tamp x [pw=w]
local a=_b[_cons]
local b=_b[x]
local sYgX=e(rmse)^2
local p=('muX'*sYgX-'a'*b*sXX)/('sXX'*b^2+'sYgX')
local g=('b'*sXX)/('sXX'*b^2+'sYgX')
local e=sqrt('sYgX'*sXX/('sXX'*b^2+'sYgX'))
qui regress y_tamp x [pw=w]
mat Vq='obs'*swap*e(V)*swap
qui save robust, replace
forvalues j=1 (1) $imps {

```

```

    qui gen ximp'j'=x
    qui replace ximp'j'='p'+ 'g'*y+'e'*invnorm(uniform()) if x==.
}
global n=_N
proc1
local alphaMI=betaMI[1,1]
local betaMI=betaMI[2,1]
qui save imp, replace
qui drop if (R==1 & Rstar==0)
qui save impmod, replace
use imp, clear
qui drop if Rstar==1
forvalues j=1 (1) $imps {
    rename ximp'j' ximpb'j'
}
qui drop x
merge using impmod
qui drop w Rstar _merge
mat IFinv=(0,0\0,0)
forvalues j=1 (1) $imps {
    qui regress y ximp'j'
    mat Var'j'='obs'*swap*e(V)*swap
    mat IFinv=IFinv+Var'j'
}
mat IFinv=IFinv/$imps
mata: proc2('obs')
mata: proc3('obs')
local w0=w0
local w1=w1
local w2=w2
local w3=w3
local w4=w4
local w5=w5
mata: proc4('w0','w1','w2','w3','w4','w5')
local SEalphaMI=sqrt(VarImp[1,1]/'obs')
local SEbetaMI=sqrt(VarImp[2,2]/'obs')
local uba='alphaMI'+invttail('obs'-2,0.025)*'SEalphaMI'
local lba='alphaMI'-invttail('obs'-2,0.025)*'SEalphaMI'
local ubb='betaMI'+invttail('obs'-2,0.025)*'SEbetaMI'
local lbb='betaMI'-invttail('obs'-2,0.025)*'SEbetaMI'
if 'uba'<1 | 'lba'>1 {
    local covb=0
}
else {
    local covb=1
}
if 'ubb'<2 | 'lbb'>2 {
    local covb=0
}

```

```

}
else {
    local covb=1
}

capture program drop procl
program define procl
mat betaMI=(0,0,0)'
forvalues j=1(1)$imps {
    gen wx=w*ximp'j'
    gen wxx=w*(ximp'j'^2)
    gen wy=w*y
    gen wxy=w*y*ximp'j'
    egen sw=sum(w)
    egen swx=sum(wx)
    egen swxx=sum(wxx)
    egen swy=sum(wy)
    egen swxy=sum(wxy)
    local sw=sw
    local swx=swx
    local swxx=swxx
    local swy=swy
    local swxy=swxy
    mat A=('sw', 'swx' \ 'swx', 'swxx')
    mat B=('swy', 'swxy')
    mat beta=invsym(A)*B
    local al=beta[1,1]
    local be=beta[2,1]
    gen y_m_yhat=y-'al'-'be'*ximp'j'
    gen wy_m_yhat_sq=w*(y_m_yhat^2)
    egen swy_m_yhat_sq=sum(wy_m_yhat_sq)
    local sYgX=($n/($n-2))*swy_m_yhat_sq/sw
    mat beta'j'=(beta', 'sYgX')
    mat betaMI=betaMI+beta'j'
    drop wx-swy_m_yhat_sq
}
mat betaMI=betaMI/$imps
end

mata:
real matrix proc2(obs)
{
    beta1=st_matrix("beta1")
    beta2=st_matrix("beta2")
    beta3=st_matrix("beta3")
    beta4=st_matrix("beta4")
    beta5=st_matrix("beta5")
    beta6=st_matrix("beta6")
}

```

```
beta7=st_matrix("beta7")
beta8=st_matrix("beta8")
beta9=st_matrix("beta9")
beta10=st_matrix("beta10")
betaMI=st_matrix("betaMI")
st_view(X1=.,.,("cons","ximp1"))
st_view(X2=.,.,("cons","ximp2"))
st_view(X3=.,.,("cons","ximp3"))
st_view(X4=.,.,("cons","ximp4"))
st_view(X5=.,.,("cons","ximp5"))
st_view(X6=.,.,("cons","ximp6"))
st_view(X7=.,.,("cons","ximp7"))
st_view(X8=.,.,("cons","ximp8"))
st_view(X9=.,.,("cons","ximp9"))
st_view(X10=.,.,("cons","ximp10"))
st_view(x1=.,.,("ximp1"))
st_view(x2=.,.,("ximp2"))
st_view(x3=.,.,("ximp3"))
st_view(x4=.,.,("ximp4"))
st_view(x5=.,.,("ximp5"))
st_view(x6=.,.,("ximp6"))
st_view(x7=.,.,("ximp7"))
st_view(x8=.,.,("ximp8"))
st_view(x9=.,.,("ximp9"))
st_view(x10=.,.,("ximp10"))
st_view(Y=.,.,"y")
V1=(0,0,1)*beta1
V2=(0,0,1)*beta2
V3=(0,0,1)*beta3
V4=(0,0,1)*beta4
V5=(0,0,1)*beta5
V6=(0,0,1)*beta6
V7=(0,0,1)*beta7
V8=(0,0,1)*beta8
V9=(0,0,1)*beta9
V10=(0,0,1)*beta10
VMI=(0,0,1)*betaMI
beta1=(1,0,0\0,1,0)*beta1
beta2=(1,0,0\0,1,0)*beta2
beta3=(1,0,0\0,1,0)*beta3
beta4=(1,0,0\0,1,0)*beta4
beta5=(1,0,0\0,1,0)*beta5
beta6=(1,0,0\0,1,0)*beta6
beta7=(1,0,0\0,1,0)*beta7
beta8=(1,0,0\0,1,0)*beta8
beta9=(1,0,0\0,1,0)*beta9
beta10=(1,0,0\0,1,0)*beta10
betaMI=(1,0,0\0,1,0)*betaMI
```

```

S1=(1/V1)*(Y-X1*beta1)
S2=(1/V2)*(Y-X2*beta2)
S3=(1/V3)*(Y-X3*beta3)
S4=(1/V4)*(Y-X4*beta4)
S5=(1/V5)*(Y-X5*beta5)
S6=(1/V6)*(Y-X6*beta6)
S7=(1/V7)*(Y-X7*beta7)
S8=(1/V8)*(Y-X8*beta8)
S9=(1/V9)*(Y-X9*beta9)
S10=(1/V10)*(Y-X10*beta10)
T1=x1:*S1
T2=x2:*S2
T3=x3:*S3
T4=x4:*S4
T5=x5:*S5
T6=x6:*S6
T7=x7:*S7
T8=x8:*S8
T9=x9:*S9
T10=x10:*S10
S1=(S1,T1)'
S2=(S2,T2)'
S3=(S3,T3)'
S4=(S4,T4)'
S5=(S5,T5)'
S6=(S6,T6)'
S7=(S7,T7)'
S8=(S8,T8)'
S9=(S9,T9)'
S10=(S10,T10)'
SavA=(S1+S2+S3+S4+S5+S6+S7+S8+S9+S10)/10
IFmI1A=(S1-SavA)*(S1-SavA)'
IFmI2A=(S2-SavA)*(S2-SavA)'
IFmI3A=(S3-SavA)*(S3-SavA)'
IFmI4A=(S4-SavA)*(S4-SavA)'
IFmI5A=(S5-SavA)*(S5-SavA)'
IFmI6A=(S6-SavA)*(S6-SavA)'
IFmI7A=(S7-SavA)*(S7-SavA)'
IFmI8A=(S8-SavA)*(S8-SavA)'
IFmI9A=(S9-SavA)*(S9-SavA)'
IFmI10A=(S10-SavA)*(S10-SavA)'
IFmIA=(1/9)*(1/obs)*(IFmI1A+IFmI2A+IFmI3A+IFmI4A+IFmI5A+IFmI6A+IFmI7A
+IFmI8A+IFmI9A+IFmI10A)

return(st_matrix("IFmI",IFmIA))
}
end

mata:

```

```
real matrix proc3(obs)
{
  beta1=st_matrix("beta1")
  beta2=st_matrix("beta2")
  beta3=st_matrix("beta3")
  beta4=st_matrix("beta4")
  beta5=st_matrix("beta5")
  beta6=st_matrix("beta6")
  beta7=st_matrix("beta7")
  beta8=st_matrix("beta8")
  beta9=st_matrix("beta9")
  beta10=st_matrix("beta10")
  betaMI=st_matrix("betaMI")
  st_view(X1=.,.,("cons","ximpb1"))
  st_view(X2=.,.,("cons","ximpb2"))
  st_view(X3=.,.,("cons","ximpb3"))
  st_view(X4=.,.,("cons","ximpb4"))
  st_view(X5=.,.,("cons","ximpb5"))
  st_view(X6=.,.,("cons","ximpb6"))
  st_view(X7=.,.,("cons","ximpb7"))
  st_view(X8=.,.,("cons","ximpb8"))
  st_view(X9=.,.,("cons","ximpb9"))
  st_view(X10=.,.,("cons","ximpb10"))
  st_view(x1=.,.,("ximpb1"))
  st_view(x2=.,.,("ximpb2"))
  st_view(x3=.,.,("ximpb3"))
  st_view(x4=.,.,("ximpb4"))
  st_view(x5=.,.,("ximpb5"))
  st_view(x6=.,.,("ximpb6"))
  st_view(x7=.,.,("ximpb7"))
  st_view(x8=.,.,("ximpb8"))
  st_view(x9=.,.,("ximpb9"))
  st_view(x10=.,.,("ximpb10"))
  st_view(Y=.,.,"y")
  V1=(0,0,1)*beta1
  V2=(0,0,1)*beta2
  V3=(0,0,1)*beta3
  V4=(0,0,1)*beta4
  V5=(0,0,1)*beta5
  V6=(0,0,1)*beta6
  V7=(0,0,1)*beta7
  V8=(0,0,1)*beta8
  V9=(0,0,1)*beta9
  V10=(0,0,1)*beta10
  VMI=(0,0,1)*betaMI
  beta1=(1,0,0\0,1,0)*beta1
  beta2=(1,0,0\0,1,0)*beta2
  beta3=(1,0,0\0,1,0)*beta3
```

```
beta4=(1,0,0\0,1,0)*beta4
beta5=(1,0,0\0,1,0)*beta5
beta6=(1,0,0\0,1,0)*beta6
beta7=(1,0,0\0,1,0)*beta7
beta8=(1,0,0\0,1,0)*beta8
beta9=(1,0,0\0,1,0)*beta9
beta10=(1,0,0\0,1,0)*beta10
betaMI=(1,0,0\0,1,0)*betaMI
S1=(1/V1)*(Y-X1*beta1)
S2=(1/V2)*(Y-X2*beta2)
S3=(1/V3)*(Y-X3*beta3)
S4=(1/V4)*(Y-X4*beta4)
S5=(1/V5)*(Y-X5*beta5)
S6=(1/V6)*(Y-X6*beta6)
S7=(1/V7)*(Y-X7*beta7)
S8=(1/V8)*(Y-X8*beta8)
S9=(1/V9)*(Y-X9*beta9)
S10=(1/V10)*(Y-X10*beta10)
T1=x1:*S1
T2=x2:*S2
T3=x3:*S3
T4=x4:*S4
T5=x5:*S5
T6=x6:*S6
T7=x7:*S7
T8=x8:*S8
T9=x9:*S9
T10=x10:*S10
S1=(S1,T1)'
S2=(S2,T2)'
S3=(S3,T3)'
S4=(S4,T4)'
S5=(S5,T5)'
S6=(S6,T6)'
S7=(S7,T7)'
S8=(S8,T8)'
S9=(S9,T9)'
S10=(S10,T10)'
SavA=(S1+S2+S3+S4+S5+S6+S7+S8+S9+S10)/10
IFmI1A=(S1-SavA)*(S1-SavA)'
IFmI2A=(S2-SavA)*(S2-SavA)'
IFmI3A=(S3-SavA)*(S3-SavA)'
IFmI4A=(S4-SavA)*(S4-SavA)'
IFmI5A=(S5-SavA)*(S5-SavA)'
IFmI6A=(S6-SavA)*(S6-SavA)'
IFmI7A=(S7-SavA)*(S7-SavA)'
IFmI8A=(S8-SavA)*(S8-SavA)'
IFmI9A=(S9-SavA)*(S9-SavA)'
```

```

IFmI10A=(S10-SavA)*(S10-SavA)'
IFmIA=(1/9)*(1/obs)*(IFmI1A+IFmI2A+IFmI3A+IFmI4A+IFmI5A+IFmI6A+IFmI7A
+IFmI8A+IFmI9A+IFmI10A)

return(st_matrix("IFmIY",IFmIA))
}
end

mata:
real matrix proc4(real scalar w0, real scalar w1, real scalar w2,
real scalar w3, real scalar w4, real scalar w5)
{
IFinv=st_matrix("IFinv")
IFmI=st_matrix("IFmI")
IFmIY=st_matrix("IFmIY")
IF=invsym(IFinv)
I=IF-IFmI
IY=IF-IFmIY
Vq=st_matrix("Vq")
VarImp=IFinv*(w3*((1/10)*IFmI+I)+w5*((1/10)*IFmIY+IY)+(IFmI*Vq*IFmI)
+(w2^2)*(IFmIY*Vq*IFmIY)-2*w4*((1/10)*(IF-w0*IFmIY)
+(9/10)*IY)+2*w1*(IFmI)-2*w1*w2*(IFmIY)
-w2*(IY*IFinv*IFmI)
-w2*(IFmI*IFinv*IY)+2*(w2^2)*(IY*IFinv*IFmIY)
-w2*IFmI*Vq*IFmIY-w2*IFmIY*Vq*IFmI)*IFinv
return(st_matrix("VarImp",VarImp))
}
end

```

C.1.2 Proper

```

use robust, clear
qui regress x [pw=w]
local muX=_b[_cons]
local semuX=_se[_cons]
local sXX=e(rmse)^2
qui regress y_tamp x [pw=w]
local a=_b[_cons]
local b=_b[x]
local sYgX=e(rmse)^2
qui regress y_tamp x [pw=w]
mat Vq='obs'*swap*e(V)*swap
mat Vq2=Vq/'obs'
forvalues j=1 (1) $imps {
local a'j'='a'+sqrt(Vq2[1,1])*invnorm(uniform())

```

```

local b'j'='b'+(Vq2[1,2]/Vq2[1,1])*('a'j''-'a')
              +sqrt(Vq2[2,2]-((Vq2[1,2]^2)/Vq2[1,1]))
              *invnorm(uniform())
local sYgX'j'='sYgX'*invchi2('completers'-2,uniform())/('completers'-2)
local muX'j'='muX'+'semuX'*invnorm(uniform())
local sXX'j'='sXX'*invchi2('completers'-2,uniform())/('completers'-2)
local p'j'=('muX'j''*'sYgX'j''-'a'j''*'b'*'sXX')/('sXX'*'b'^2+'sYgX')
local g'j'=('b'*'sXX')/('sXX'*'b'^2+'sYgX')
local e'j'=sqrt('sYgX'*'sXX')/('sXX'*'b'^2+'sYgX')
qui gen ximp'j'=x
qui replace ximp'j'='p'j''+'g'j''*y+'e'j''*invnorm(uniform()) if x==.
}
global n=_N
proc1
local alphaMI=betaMI[1,1]
local betaMI=betaMI[2,1]
qui save imp, replace
qui drop if (R==1 & Rstar==0)
qui save impmod, replace
use imp, clear
qui drop if Rstar==1
forvalues j=1 (1) $imps {
    rename ximp'j' ximpb'j'
}
qui drop x
merge using impmod
qui drop w Rstar _merge
mat IFinv=(0,0\0,0)
forvalues j=1 (1) $imps {
    qui regress y ximp'j'
    mat Var'j'='obs'*swap*e(V)*swap
    mat IFinv=IFinv+Var'j'
}
mat IFinv=IFinv/$imps
mata: proc2('obs')
mata: proc3('obs')
local w0=w0
local w1=w1
local w2=w2
local w3=w3
local w4=w4
local w5=w5
mata: proc7('w0','w1','w2','w3','w4','w5')
local SEalphaMI=sqrt(VarImp[1,1]/'obs')
local SEbetaMI=sqrt(VarImp[2,2]/'obs')
local uba='alphaMI'+invttail('obs'-2,0.025)*'SEalphaMI'
local lba='alphaMI'-invttail('obs'-2,0.025)*'SEalphaMI'
local ubb='betaMI'+invttail('obs'-2,0.025)*'SEbetaMI'

```

```

local lbb='betaMI'-invttail('obs'-2,0.025)*'SEbetaMI'
if 'uba'<1 | 'lba'>1 {
    local covb=0
}
else {
    local covb=1
}
if 'ubb'<2 | 'lbb'>2 {
    local covb=0
}
else {
    local covb=1
}

mata:
real matrix proc7(real scalar w0, real scalar w1, real scalar w2,
                 real scalar w3, real scalar w4, real scalar w5)
{
    IFinv=st_matrix("IFinv")
    IFmI=st_matrix("IFmI")
    IFmIY=st_matrix("IFmIY")
    IF=invsym(IFinv)
    I=IF-IFmI
    IY=IF-IFmIY
    Vq=st_matrix("Vq")
    VarImp=IFinv*(w3*((1/10)*IFmI+I)+w5*((1/10)*IFmIY+IY)
              +(11/10)*(IFmI*Vq*IFmI)
              +(11/10)*(w2^2)*(IFmIY*Vq*IFmIY)
              -2*w4*((1/10)*(IF-w0*IFmIY)
              +(9/10)*IY)+2*w1*(IFmI)-2*w1*w2*(IFmIY)
              -w2*(IY*IFinv*IFmI)-w2*(IFmI*IFinv*IY)
              +2*(w2^2)*(IY*IFinv*IFmIY)
              -(11/10)*w2*IFmI*Vq*IFmIY
              -(11/10)*w2*IFmIY*Vq*IFmI)*IFinv
    return(st_matrix("VarImp",VarImp))
}
end

```

C.2 Robust multiple imputation: alternative formulation

Here is the Stata command for DRMI (both models correct) used in the simulation study discussed in §9.3.3.

```

#delimit ;
ice x xsq p1 p2 p12a p12b w1 y1 y1sq p12 w2 y2,
    eq(y1:x xsq y2 w1,y2:x y1 w2)
    passive(y1sq:y1^2 \ p12:(exp(p12a+p12b*y1sq)/(1+exp(p12a+p12b*y1sq)))
    \ tw2:(1/(p2+p1*p12)))
    m(10) cycles(10) saving(DRMicc_RMM_longit) replace orderasis
;

```

C.3 Binary simulation study: scenario 1

```

clear
set mem 745m
set obs 500
local y1y2_int=1
qui gen full_11=.
qui gen iee_11=.
qui gen gee_11=.
qui gen cwgee_11=.
qui gen owgee_11=.
qui gen mige_11=.
qui gen br_11=.
qui gen rr_11=.
qui gen drmi_11=.
qui gen full_10=.
qui gen iee_10=.
qui gen gee_10=.
qui gen cwgee_10=.
qui gen owgee_10=.
qui gen mige_10=.
qui gen br_10=.
qui gen rr_10=.
qui gen drmi_10=.
qui gen full_01=.
qui gen iee_01=.
qui gen gee_01=.
qui gen cwgee_01=.
qui gen owgee_01=.
qui gen mige_01=.
qui gen br_01=.
qui gen rr_01=.
qui gen drmi_01=.
qui gen full_00=.
qui gen iee_00=.

```

```

qui gen gee_00=.
qui gen cwgee_00=.
qui gen owgee_00=.
qui gen mige_00=.
qui gen br_00=.
qui gen rr_00=.
qui gen drmi_00=.
forvalues wei=0(1)1 {
  local weights=1-'wei'
  forvalues cond=0(1)1 {
    local conditional=1-'cond'
    forvalues sim=1(1)1000 {
      di 'sim' " ." _cont
      set seed 'sim'
      keep full_* iee_* gee_* cwgee_* owgee_* mige_* br_* rr_* drmi_*
      gen x1=uniform()<0.5
      gen x2=uniform()<0.25
      gen piy1=exp(x1-0.5*x1*x2)/(1+exp(x1-0.5*x1*x2))
      gen piy2=exp(-1+0.25*x1+0.25*x2-x1*x2)/(1+exp(-1+0.25*x1+0.25*x2-x1*x2))
      gen piy3=exp(-x1+0.5*x2-x1*x2)/(1+exp(-x1+0.5*x2-x1*x2))
      if 'y12_int'==1 {
        gen g12=0.3
        gen g13=-0.15
        gen g23=0.3
        gen g123=-0.1
      }
      else {
        gen g12=0.2
        gen g13=0
        gen g23=0.2
        gen g123=-.173
      }
      gen y1=uniform()<piy1
      gen e1=(y1-piy1)/(sqrt(piy1*(1-piy1)))
      gen piy2gy1=piy2*(1+g12*e1*(1-piy2)/sqrt(piy2*(1-piy2)))
      gen y2=uniform()<piy2gy1
      gen e2=(y2-piy2)/(sqrt(piy2*(1-piy2)))
      #delimit ;
      gen piy3gy1y2=piy3*(1+g12*e1*e2+(g13*e1+g23*e2+g123*e1*e2)*((1-piy3)
/sqrt(piy3*(1-piy3)))));
      #delimit cr
      gen y3=uniform()<piy3gy1y2

      *full
      gen x1x2=x1*x2
      gen id=_n
      qui reshape long y, i(id) j(t)
      gen t1=(t==1)

```

```

gen t2=(t==2)
gen t3=(t==3)
gen t1x1=t1*x1
gen t1x2=t1*x2
gen t1x1x2=t1*x1*x2
gen t2x1=t2*x1
gen t2x2=t2*x2
gen t2x1x2=t2*x1*x2
gen t3x1=t3*x1
gen t3x2=t3*x2
gen t3x1x2=t3*x1x2
#delimit ;
qui xtgee y t1 t1x1 t1x2 t1x1x2 t2 t2x1 t2x2 t2x1x2 t3 t3x1 t3x2 t3x1x2,
i(id) t(t) family(binomial) link(logit) corr(unstr) vce(robust) nocons;
local full=_b[t3x1];
keep full_* iee_* gee_* cwgee_* owgee_* migegee_* br_*
rr_* drmi_* id t y x1 x2 x1x2;
#delimit cr
qui reshape wide y, i(id) j(t)
qui replace full_ 'conditional' 'weights'='full' in 'sim'

gen pir2=exp(0.5*x1-0.5*x2+3*y1)/(1+exp(0.5*x1-0.5*x2+3*y1))
#delimit ;
gen pir3g2=exp(-0.5-0.5*x1+0.5*x2+x1*y1-y1*y2+4*y2*x1)/
(1+exp(-0.5-0.5*x1+0.5*x2+x1*y1-y1*y2+4*y2*x1));
#delimit cr
gen r1=1
gen r2=uniform()<pir2
qui gen r3=uniform()<pir3g2 if r2==1
qui replace r3=0 if r2==0
qui replace y2=. if r2==0
qui replace y3=. if r3==0

*iee
qui reshape long y r, i(id) j(t)
gen t1=(t==1)
gen t2=(t==2)
gen t3=(t==3)
gen t1x1=t1*x1
gen t1x2=t1*x2
gen t1x1x2=t1*x1*x2
gen t2x1=t2*x1
gen t2x2=t2*x2
gen t2x1x2=t2*x1*x2
gen t3x1=t3*x1
gen t3x2=t3*x2
gen t3x1x2=t3*x1x2
#delimit ;

```

```

capture qui xtgee y t1 t1x1 t1x2 t1x1x2 t2 t2x1 t2x2 t2x1x2 t3 t3x1
t3x2 t3x1x2, i(id) t(t) family(binomial) link(logit) corr(ind)
vce(robust) nocons;
#delimit cr
if abs(e(dif))<e(tol) {
  local check=1
}
else {
  local check=0
}
local iee=_b[t3x1]
#delimit ;
keep full_* iee_* gee_* cwgee_* owgee_* migegee_* br_* rr_* drmi_*
id t y r x1 x2 x1x2;
#delimit cr
qui reshape wide y r, i(id) j(t)
if 'check'==1 {
  qui replace iee_'conditional' 'weights'='iee' in 'sim'
}

qui gen y1y2=y1*y2
qui gen x1y1=x1*y1
qui gen x1y2=x1*y2
qui gen x2y1=x2*y1
if 'weights'==1 {
  qui logit r3 x1 x2 x1y1 y1y2 x1y2 if r2==1, asis
  qui predict p3g2
  qui gen lp3g2a=_b[_cons]+_b[x1]*x1+_b[x2]*x2+_b[x1y1]*x1*y1
  qui gen lp3g2b=_b[y1y2]*y1+_b[x1y2]*x1
  qui logit r2 x1 x2 y1, asis
  qui predict p2
  qui gen w2=1/p2
  qui gen p3=p3g2*p2
  qui gen w3=1/p3
  qui gen cw=w3 if r3==1
  qui replace cw=w2/(1-p3g2) if r3==0
  qui replace cw=1/(1-p2) if r2==0
}
if 'weights'==0 {
  qui logit r3 x1 x1y1 if r2==1, asis
  qui gen lp3g2a=_b[_cons]+_b[x1]*x1+_b[x1y1]*x1*y1
  qui gen lp3g2b=0
  qui predict p3g2
  qui logit r2 x1 x2y1, asis
  qui predict p2
  qui gen w2=1/p2
  qui gen p3=p3g2*p2
  qui gen w3=1/p3
}

```

```

    qui gen cw=w3 if r3==1
    qui replace cw=w2/(1-p3g2) if r3==0
    qui replace cw=1/(1-p2) if r2==0
}

*gee
qui reshape long y r, i(id) j(t)
gen t1=(t==1)
gen t2=(t==2)
gen t3=(t==3)
gen t1x1=t1*x1
gen t1x2=t1*x2
gen t1x1x2=t1*x1*x2
gen t2x1=t2*x1
gen t2x2=t2*x2
gen t2x1x2=t2*x1*x2
gen t3x1=t3*x1
gen t3x2=t3*x2
gen t3x1x2=t3*x1*x2
#delimit ;
capture qui xtgee y t1 t1x1 t1x2 t1x1x2 t2 t2x1 t2x2 t2x1x2 t3 t3x1
t3x2 t3x1x2, i(id) t(t) family(binomial) link(logit) corr(unstr)
vce(robust) nocons;
#delimit cr
if abs(e(dif))<e(tol) {
    local check1=1
    local gee=_b[t3x1]
    qui predict mu
    mat b=e(b)
    mat b=b'
    mat corr=e(R)
}
else {
    local check1=0
    qui gen mu=.
}

*cluster-level weighted gee
#delimit ;
capture qui xtgee y t1 t1x1 t1x2 t1x1x2 t2 t2x1 t2x2 t2x1x2 t3 t3x1
t3x2 t3x1x2 [pw=cw], i(id) t(t) family(binomial) link(logit) corr(unstr)
vce(robust) nocons;
#delimit cr
if abs(e(dif))<e(tol) {
    local check2=1
    local cwgee=_b[t3x1]
}
else {

```

```

    local check2=0
}
#delimit ;
keep full_* iee_* gee_* cwgee_* owgee_* migegee_* br_* rr_* drmi_*
id t y x1 x2 x1x2 r w2 w3 cw mu y1y2 p3 p2 p3g2 lp3g2a lp3g2b;
#delimit cr
qui reshape wide y r mu, i(id) j(t)
if 'check1'==1 {
    qui replace gee_ 'conditional' 'weights'='gee' in 'sim'
}
if 'check2'==1 {
    qui replace cwgee_ 'conditional' 'weights'='cwgee' in 'sim'
}

*observation-level weighted gee
if 'check1'==1 {
    qui gen py1=y1-mu1
    qui gen py2=w2*(y2-mu2)
    qui gen py3=w3*(y3-mu3)
    qui replace py2=0 if py2==.
    qui replace py3=0 if py3==.
    qui gen d1=mu1*(1-mu1)
    qui gen d2=mu2*(1-mu2)
    qui gen d3=mu3*(1-mu3)
    qui replace d2=0 if d2==.
    qui replace d3=0 if d3==.
    qui gen r12=corr[1,2]
    qui gen r13=corr[1,3]
    qui gen r23=corr[2,3]
    mat r=(1,r12[1],r13[1]\r12[1],1,r23[1]\r13[1],r23[1],1)
    qui gen nu1=0
    qui gen nu2=0
    qui gen nu3=0
    local absdiff=1
    local count=0
    local check_ow=1
    local check_ow2=1
    while 'absdiff'>1e-5 & 'count'<500 & 'check_ow'==1 & 'check_ow2'==1 {
        mat ss1=J(12,12,0)
        mat ss2=J(12,1,0)
        forvalues sub=1(1)500 {
            if 'check_ow'==1 {
                #delimit ;
                mat x=(1,0,0\x1['sub'],0,0\x2['sub'],0,0\x1x2['sub'],0,0\
                    0,1,0\0,x1['sub'],0\0,x2['sub'],0\0,x1x2['sub'],0\
                    0,0,1\0,0,x1['sub']\0,0,x2['sub']\0,0,x1x2['sub'])';
                #delimit cr
                mat d=(d1['sub'],0,0\0,d2['sub'],0\0,0,d3['sub'])
            }
        }
    }
}

```

```

mat vhalf=(sqrt(d1['sub']),0,0\0,sqrt(d2['sub']),0\0,0,sqrt(d3['sub']))
mat w=vhalf*r*vhalf
if matmissing(w)==0 {
  mat invw=invsym(w)
}
else {
  local check_ow=0
}
}
if 'check_ow'==1 {
mat py=(py1['sub']\py2['sub']\py3['sub'])
mat s2=x'*d*invw*py
mat s1=s2*s2'
mat ss1=ss1+s1
mat ss2=ss2+s2
}
}
if 'check_ow'==1 {
if matmissing(ss1)==0 {
mat diff=invsym(ss1)*ss2
}
else {
local check_ow2=0
}
if 'check_ow2'==1 {
local absdiff=abs(diff[1,1])
forvalues j=2(1)12 {
  if 'absdiff'<abs(diff[['j'],1]) {
    local absdiff=abs(diff[['j'],1])
  }
}
}
mat b=b+diff
forvalues sub=1(1)500 {
  #delimit ;
  mat x=(1,0,0\x1['sub'],0,0\x2['sub'],0,0\x1x2['sub'],0,0\
    0,1,0\0,x1['sub'],0\0,x2['sub'],0\0,x1x2['sub'],0\
    0,0,1\0,0,x1['sub']\0,0,x2['sub']\0,0,x1x2['sub'])';
  #delimit cr
  mat nu=x*b
  qui replace nu1=nu[1,1] in 'sub'
  qui replace nu2=nu[2,1] in 'sub'
  qui replace nu3=nu[3,1] in 'sub'
}
qui replace mu1=exp(nu1)/(1+exp(nu1))
qui replace mu2=exp(nu2)/(1+exp(nu2))
qui replace mu3=exp(nu3)/(1+exp(nu3))
qui replace py1=y1-mu1
qui replace py2=w2*(y2-mu2)

```

```

qui replace py3=w3*(y3-mu3)
qui replace py2=0 if py2==.
qui replace py3=0 if py3==.
qui replace d1=mu1*(1-mu1)
qui replace d2=mu2*(1-mu2)
qui replace d3=mu3*(1-mu3)
qui replace d2=0 if d2==.
qui replace d3=0 if d3==.
}
}
local count='count'+1
}
if 'count'>499 {
  qui replace owgee_ 'conditional' 'weights'=. in 'sim'
}
else {
  if 'check_ow'==1 & 'check_ow2'==1 {
    qui replace owgee_ 'conditional' 'weights'=b[10,1] in 'sim'
  }
}
}
}

```

```

*Robins & Rotnitzky (1995)
qui save thesis_ch12_sat, replace
mat b=(1,-.5,-.75,-.75,-1,-.5)'
qui drop if x1==0
local tosub=_N
local check_rr=1
local check_rr2=1
if 'check1'==1 & 'check_ow'==1 & 'check_ow2'==1 {
  if 'weights'==1 {
    qui logit r3 y1 x2
  }
  if 'weights'==0 {
    qui logit r3 y1
  }
}
qui predict p3g1
qui gen k21=w2*(y2-mu2)
qui gen k31=w3*(y3-mu3)
qui gen k32=p2*w3*(y3-mu3)
if 'conditional'==1 {
  qui regress k21 y1 x2
  qui predict kap21
  qui regress k31 y1 x2
  qui predict kap31
  qui regress k32 y1 y2 y1y2 x2
  qui predict kap32
}
}

```

```

if 'conditional'==0 {
  qui regress k21 y1
  qui predict kap21
  qui regress k31 y1 x2
  qui predict kap31
  qui regress k32 y1 y2 x2
  qui predict kap32
}
qui gen g22=p2*kap21
qui gen g32=p3g1*kap31
qui gen g33=p3g2*kap32
qui gen q22=w2*g22
qui gen q32=w2*g32
qui gen q33=w3*g33
qui replace q22=0 if q22==.
qui replace q32=0 if q32==.
qui replace q33=0 if q33==.
qui gen P2=(r2-p2)*q22
qui gen P3=(r2-p2)*q32+(r3-r2*p3g2)*q33
qui gen UmP1=py1
qui gen UmP2=py2-P2
qui gen UmP3=py3-P3
qui gen UmPUmP11=UmP1^2
qui gen UmPUmP12=UmP1*UmP2
qui gen UmPUmP13=UmP1*UmP3
qui gen UmPUmP22=UmP2^2
qui gen UmPUmP23=UmP2*UmP3
qui gen UmPUmP33=UmP3^2
qui regress UmPUmP11 x2
qui predict l11
qui regress UmPUmP12 x2
qui predict l12
qui regress UmPUmP13 x2
qui predict l13
qui regress UmPUmP22 x2
qui predict l22
qui regress UmPUmP23 x2
qui predict l23
qui regress UmPUmP33 x2
qui predict l33
qui gen s12=d1*l12*q22+d1*l13*q32
qui gen s22=d2*l22*q22+d2*l23*q32
qui gen s32=d3*l23*q22+d3*l33*q32
qui gen s13=d1*l13*q33
qui gen s23=d2*l23*q33
qui gen s33=d3*l33*q33
qui gen logit3g2=log(p3g2/(1-p3g2))
qui gen logit2=log(p2/(1-p2))

```



```

if 'check_rr2'==1 {
local absdiff=abs(diff[1,1])
forvalues j=2(1)6 {
  if 'absdiff'<abs(diff['j',1]) {
    local absdiff=abs(diff['j',1])
  }
}
mat b=b+diff
forvalues sub=1(1)'totsub' {
  #delimit ;
  mat x=(1,0,0\x2['sub'],0,0\
    0,1,0\x2['sub'],0\
    0,0,1\x2['sub'])';
  #delimit cr
  mat nu=x*b
  qui replace nu1=nu[1,1] in 'sub'
  qui replace nu2=nu[2,1] in 'sub'
  qui replace nu3=nu[3,1] in 'sub'
}
qui replace mu1=exp(nu1)/(1+exp(nu1))
qui replace mu2=exp(nu2)/(1+exp(nu2))
qui replace mu3=exp(nu3)/(1+exp(nu3))
qui replace py1=y1-mu1
qui replace py2=w2new*(y2-mu2)
qui replace py3=w3new*(y3-mu3)
qui replace py2=0 if py2==.
qui replace py3=0 if py3==.
qui replace d1=mu1*(1-mu1)
qui replace d2=mu2*(1-mu2)
qui replace d3=mu3*(1-mu3)
qui replace d2=0 if d2==.
qui replace d3=0 if d3==.
}
}
local count='count'+1
}
if 'count'>99 {
  local rr_1check=0
}
else {
  if 'check_rr'==1 & 'check_rr2'==1 {
    local rr_1check=1
    local rr_1=b[5,1]
  }
  else {
    local rr_1check=0
  }
}
}

```

```
}

qui use thesis_ch12_sat, replace
mat b=(0,0,-1,.25,0,.5)'
qui drop if x1==1
local tosub=_N
local check_rr=1
local check_rr2=1
if 'check1'==1 & 'check_ow'==1 & 'check_ow2'==1 {
  if 'weights'==1 {
    qui logit r3 y1 x2
  }
  if 'weights'==0 {
    qui logit r3 y1
  }
}
qui predict p3g1
qui gen k21=w2*(y2-mu2)
qui gen k31=w3*(y3-mu3)
qui gen k32=p2*w3*(y3-mu3)
if 'conditional'==1 {
  qui regress k21 y1 x2
  qui predict kap21
  qui regress k31 y1 x2
  qui predict kap31
  qui regress k32 y1 y2 y1y2 x2
  qui predict kap32
}
if 'conditional'==0 {
  qui regress k21 y1
  qui predict kap21
  qui regress k31 y1 x2
  qui predict kap31
  qui regress k32 y1 y2 x2
  qui predict kap32
}
qui gen g22=p2*kap21
qui gen g32=p3g1*kap31
qui gen g33=p3g2*kap32
qui gen q22=w2*g22
qui gen q32=w2*g32
qui gen q33=w3*g33
qui replace q22=0 if q22==.
qui replace q32=0 if q32==.
qui replace q33=0 if q33==.
qui gen P2=(r2-p2)*q22
qui gen P3=(r2-p2)*q32+(r3-r2*p3g2)*q33
qui gen UmP1=py1
qui gen UmP2=py2-P2
```

```

qui gen UmP3=py3-P3
qui gen UmPUmP11=UmP1^2
qui gen UmPUmP12=UmP1*UmP2
qui gen UmPUmP13=UmP1*UmP3
qui gen UmPUmP22=UmP2^2
qui gen UmPUmP23=UmP2*UmP3
qui gen UmPUmP33=UmP3^2
qui regress UmPUmP11 x2
qui predict l11
qui regress UmPUmP12 x2
qui predict l12
qui regress UmPUmP13 x2
qui predict l13
qui regress UmPUmP22 x2
qui predict l22
qui regress UmPUmP23 x2
qui predict l23
qui regress UmPUmP33 x2
qui predict l33
qui gen s12=d1*112*q22+d1*113*q32
qui gen s22=d2*122*q22+d2*123*q32
qui gen s32=d3*123*q22+d3*133*q32
qui gen s13=d1*113*q33
qui gen s23=d2*123*q33
qui gen s33=d3*133*q33
qui gen logit3g2=log(p3g2/(1-p3g2))
qui gen logit2=log(p2/(1-p2))
qui logit r3 s13 s23 s33 if r2==1, nocons offset(logit3g2)
qui predict p3g2new
qui logit r2 s12 s22 s32, nocons offset(logit2)
qui predict p2new
qui gen w2new=1/p2new
qui gen p3new=p2new*p3g2new
qui gen w3new=1/p3new
qui replace py2=w2new*(y2-mu2)
qui replace py3=w3new*(y3-mu3)
qui replace py2=0 if py2==.
qui replace py3=0 if py3==.
local absdiff=1
local count=0
while 'absdiff'>1e-5 & 'count'<100 & 'check_rr'==1 & 'check_rr2'==1 {
  mat ss1=J(6,6,0)
  mat ss2=J(6,1,0)
  forvalues sub=1(1)'totsub' {
    #delimit ;
    mat x=(1,0,0\x2['sub'],0,0\
      0,1,0\0,x2['sub'],0\
      0,0,1\0,0,x2['sub'])';
  }
}

```

```

#delimit cr
mat d=(d1['sub'],0,0\0,d2['sub'],0\0,0,d3['sub'])
#delimit ;
    mat l=(l11['sub'],l12['sub'],l13['sub']\l12['sub'],l22['sub'],l23['sub']
    \l13['sub'],l23['sub'],l33['sub']);
    #delimit cr
if matmissing(l)==0 {
mat invl=invsym(l)
}
else {
local check_rr=0
}
if 'check_rr'==1 {
mat py=(py1['sub']\py2['sub']\py3['sub'])
mat s2=x*d*invl*py
mat s1=s2*s2'
mat ss1=ss1+s1
mat ss2=ss2+s2
}
}
if 'check_rr'==1 {
if matmissing(ss1)==0 {
mat diff=invsym(ss1)*ss2
}
else {
local check_rr2=0
}
if 'check_rr2'==1 {
local absdiff=abs(diff[1,1])
forvalues j=2(1)6 {
    if 'absdiff'<abs(diff['j',1]) {
        local absdiff=abs(diff['j',1])
    }
}
}
mat b=b+diff
forvalues sub=1(1)'totsub' {
    #delimit ;
    mat x=(1,0,0\x2['sub'],0,0\
    0,1,0\0,x2['sub'],0\
    0,0,1\0,0,x2['sub'])';
    #delimit cr
    mat nu=x*b
    qui replace nu1=nu[1,1] in 'sub'
    qui replace nu2=nu[2,1] in 'sub'
    qui replace nu3=nu[3,1] in 'sub'
}
}
qui replace mu1=exp(nu1)/(1+exp(nu1))
qui replace mu2=exp(nu2)/(1+exp(nu2))

```

```
qui replace mu3=exp(nu3)/(1+exp(nu3))
qui replace py1=y1-mu1
qui replace py2=w2new*(y2-mu2)
qui replace py3=w3new*(y3-mu3)
qui replace py2=0 if py2==.
qui replace py3=0 if py3==.
qui replace d1=mu1*(1-mu1)
qui replace d2=mu2*(1-mu2)
qui replace d3=mu3*(1-mu3)
qui replace d2=0 if d2==.
qui replace d3=0 if d3==.
}
}
local count='count'+1
}
if 'count'>99 {
local rr_0check=0
}
else {
if 'check_rr'==1 & 'check_rr2'==1 {
local rr_0check=1
local rr_0=b[5,1]
}
else {
local rr_0check=0
}
}
}
qui use thesis_ch12_sat, clear
if 'rr_1check'==1 & 'rr_0check'==1 {
qui replace rr_`conditional'`weights'='rr_1'-'rr_0' in `sim'
}

*MI-gee
qui save thesis_ch12_sat, replace
qui drop if x1==0
if 'conditional'==1 {
#delimit ;
qui ice y1 y2 y1y2 y3 x2, eq(y2:x2 y1, y3:x2 y1 y2 y1y2)
passive(y1y2:y1*y2) m(10) cycles(10)
saving(thesis_ch12_sat_x1e1_MIGEE) replace;
#delimit cr
}
if 'conditional'==0 {
#delimit ;
qui ice y1 y2 y3 x2, eq(y2:y1, y3:x2 y1 y2) m(10) cycles(10)
saving(thesis_ch12_sat_x1e1_MIGEE) replace;
#delimit cr
}
```

```

}
qui use thesis_ch12_sat, clear
qui drop if x1==1
if 'conditional'==1 {
    #delimit ;
    qui ice y1 y2 y1y2 y3 x2, eq(y2:x2 y1, y3:x2 y1 y2 y1y2)
    passive(y1y2:y1*y2) m(10) cycles(10)
    saving(thesis_ch12_sat_x1e0_MIGEE) replace;
    #delimit cr
}
if 'conditional'==0 {
    #delimit ;
    qui ice y1 y2 y3 x2, eq(y2:y1, y3:y1 y2) m(10) cycles(10)
    saving(thesis_ch12_sat_x1e0_MIGEE) replace;
    #delimit cr
}
use thesis_ch12_sat_x1e1_MIGEE, clear
gen idnew=_n
drop y1y2
capture drop py1-nu3
capture drop k21-w3new
qui reshape long y r, i(idnew) j(t)
gen t1=(t==1)
gen t2=(t==2)
gen t3=(t==3)
gen t1x2=t1*x2
gen t2x2=t2*x2
gen t3x2=t3*x2
#delimit ;
capture qui micombine xtgee y t1 t1x2 t2 t2x2 t3 t3x2,
i(idnew) t(t) family(binomial) link(logit) corr(unstr) vce(robust) nocons;
#delimit cr
if abs(e(dif))<e(tol) {
    local migeel=_b[t3]
    local check_migeel=1
}
else {
    local check_migeel=0
}
use thesis_ch12_sat_x1e0_MIGEE, clear
gen idnew=_n
drop y1y2
capture drop py1-nu3
capture drop k21-w3new
qui reshape long y r, i(idnew) j(t)
gen t1=(t==1)
gen t2=(t==2)
gen t3=(t==3)

```

```

gen t1x2=t1*x2
gen t2x2=t2*x2
gen t3x2=t3*x2
#delimit ;
capture qui micombine xtgee y t1 t1x2 t2 t2x2 t3 t3x2,
i(idnew) t(t) family(binomial) link(logit) corr(unstr) vce(robust) nocons;
#delimit cr
if abs(e(dif))<e(tol) {
    local mige0=_b[t3]
    local check_mige0=1
}
else {
    local check_mige0=0
}
use thesis_ch12_sat, clear
if 'check_mige1'==1 & 'check_mige0'==1 {
    qui replace mige_ 'conditional' 'weights'='mige1'-'mige0' in 'sim'
}

*Bang & Robins (2005)
qui gen h3=y3
if 'conditional'==1 {
    qui logit h3 x2 y1 y2 y1y2 w3 if r3==1 & x1==1
}
if 'conditional'==0 {
    qui logit h3 x2 y1 y2 w3 if r3==1 & x1==1
}
qui predict h2 if r2==1 & x1==1
if 'conditional'==1 {
    qui logit h3 x2 y1 y2 y1y2 w3 if r3==1 & x1==0
}
if 'conditional'==0 {
    qui logit h3 x2 y1 y2 w3 if r3==1 & x1==0
}
qui predict h20 if r2==1 & x1==0
qui replace h2=h20 if x1==0
drop h20
if 'conditional'==1 {
    qui regress h2 y1 x2 w2 if r2==1 & x1==1
}
if 'conditional'==0 {
    qui regress h2 y1 w2 if r2==1 & x1==1
}
qui predict h1 if x1==1
if 'conditional'==1 {
    qui regress h2 y1 x2 w2 if r2==1 & x1==0
}

```

```

if 'conditional'==0 {
  qui regress h2 y1 w2 if r2==1 & x1==0
}
qui predict h10 if x1==0
qui replace h1=h10 if x1==0
drop h10
qui regress h1 x2 if x1==1
local m1=_b[_cons]
qui regress h1 x2 if x1==0
local m0=_b[_cons]
#delimit ;
qui replace br_'conditional','weights'=log((('m1'*(1-'m0'))/
((1-'m1')*'m0')) in 'sim';
#delimit cr

*Doubly robust MI
qui save thesis_ch12_sat, replace
qui drop if x1==0
qui save thesis_ch12_sat_x1e1, replace
if 'conditional'==1 {
  if 'weights'==1 {
    #delimit ;
    qui ice y1 x2 p2 w2 lp3g2a lp3g2b y2 y1y2 p3g2 p3 w3 y3,
    eq(y2:x2 y1 w2, y3:x2 y1 y2 y1y2 w3)
    passive(y1y2:y1*y2 \ p3g2:(exp(lp3g2a+lp3g2b*y2)/
      (1+exp(lp3g2a+lp3g2b*y2))) \ p3:p2*p3g2 \ w3:1/p3)
    m(10) cycles(10) saving(thesis_ch12_sat_x1e1_DRMI) replace;
    #delimit cr
  }
  else {
    #delimit ;
    qui ice y1 x2 w2 y2 y1y2 w3 y3,
    eq(y2:x2 y1 w2, y3:x2 y1 y2 y1y2 w3)
    passive(y1y2:y1*y2)
    m(10) cycles(10) saving(thesis_ch12_sat_x1e1_DRMI) replace;
    #delimit cr
  }
}
if 'conditional'==0 {
  if 'weights'==1 {
    #delimit ;
    qui ice y1 x2 p2 w2 lp3g2a lp3g2b y2 p3g2 p3 w3 y3,
    eq(y2:y1 w2, y3:y1 y2 x2 w3)
    passive(p3g2:(exp(lp3g2a+lp3g2b*y2)/(1+exp(lp3g2a+lp3g2b*y2)))
      \ p3:p2*p3g2 \ w3:1/p3)
    m(10) cycles(10) saving(thesis_ch12_sat_x1e1_DRMI) replace;
    #delimit cr
  }
}

```

```

else {
  #delimit ;
  qui ice y1 x2 w2 y2 w3 y3,
  eq(y2:y1 w2, y3:y1 y2 x2 w3)
  m(10) cycles(10) saving(thesis_ch12_sat_x1e1_DRMI) replace;
  #delimit cr
}
}
qui use thesis_ch12_sat, clear
qui drop if x1==1
qui save thesis_ch12_sat_x1e0, replace
if 'conditional'==1 {
  if 'weights'==1 {
    #delimit ;
    qui ice y1 x2 p2 w2 lp3g2a lp3g2b y2 y1y2 p3g2 p3 w3 y3,
    eq(y2:x2 y1 w2, y3:x2 y1 y2 y1y2 w3)
    passive(y1y2:y1*y2 \ p3g2:(exp(lp3g2a+lp3g2b*y2)/
      (1+exp(lp3g2a+lp3g2b*y2))) \ p3:p2*p3g2 \ w3:1/p3)
    m(10) cycles(10) saving(thesis_ch12_sat_x1e0_DRMI) replace;
    #delimit cr
  }
  else {
    #delimit ;
    qui ice y1 x2 w2 y2 y1y2 w3 y3,
    eq(y2:x2 y1 w2, y3:x2 y1 y2 y1y2 w3)
    passive(y1y2:y1*y2)
    m(10) cycles(10) saving(thesis_ch12_sat_x1e0_DRMI) replace;
    #delimit cr
  }
}
if 'conditional'==0 {
  if 'weights'==1 {
    #delimit ;
    qui ice y1 x2 p2 w2 lp3g2a lp3g2b y2 p3g2 p3 w3 y3,
    eq(y2:y1 w2, y3:x2 y1 y2 w3)
    passive(p3g2:(exp(lp3g2a+lp3g2b*y2)/(1+exp(lp3g2a+lp3g2b*y2)))
      \ p3:p2*p3g2 \ w3:1/p3)
    m(10) cycles(10) saving(thesis_ch12_sat_x1e0_DRMI) replace;
    #delimit cr
  }
  else {
    #delimit ;
    qui ice y1 x2 w2 y2 w3 y3,
    eq(y2:y1 w2, y3:x2 y1 y2 w3)
    m(10) cycles(10) saving(thesis_ch12_sat_x1e0_DRMI) replace;
    #delimit cr
  }
}
}

```

```

use thesis_ch12_sat_x1e1_DRMI, clear
gen idnew=_n
drop y1y2
capture drop py1-nu3
capture drop k21-w3new
qui reshape long y r, i(idnew) j(t)
gen t1=(t==1)
gen t2=(t==2)
gen t3=(t==3)
gen t1x2=t1*x2
gen t2x2=t2*x2
gen t3x2=t3*x2
#delimit ;
capture qui micombine xtgee y t1 t1x2 t2 t2x2 t3 t3x2,
i(idnew) t(t) family(binomial) link(logit) corr(unstr) vce(robust) nocons;
#delimit cr
if abs(e(dif))<e(tol) {
    local drmi1=_b[t3]
    local check_drmi1=1
}
else {
    local check_drmi1=0
}
use thesis_ch12_sat_x1e0_DRMI, clear
gen idnew=_n
drop y1y2
capture drop py1-nu3
capture drop k21-w3new
qui reshape long y r, i(idnew) j(t)
gen t1=(t==1)
gen t2=(t==2)
gen t3=(t==3)
gen t1x2=t1*x2
gen t2x2=t2*x2
gen t3x2=t3*x2
#delimit ;
capture qui micombine xtgee y t1 t1x2 t2 t2x2 t3 t3x2, i(idnew) t(t)
family(binomial) link(logit) corr(unstr) vce(robust) nocons;
#delimit cr
if abs(e(dif))<e(tol) {
    local drmi0=_b[t3]
    local check_drmi0=1
}
else {
    local check_drmi0=0
}
use thesis_ch12_sat, clear
if 'check_drmi1'==1 & 'check_drmi0'==1 {

```

```
    qui replace drmi_ 'conditional' 'weights'='drmi1'-'drmi0' in 'sim'  
  }  
}  
}  
}  
qui save thesis_ch12_sat, replace
```