

2012 | School voor Informatietechnologie  
Kennistechnologie, Informatica, Wiskunde, ICT

DOCTORAATSPROEFSCHRIFT

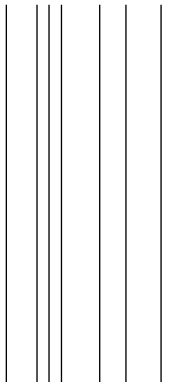
# Pseudo-likelihood and Estimating Equation Methodology for Incomplete Data

*Proefschrift voorgelegd tot het behalen van de graad van  
doctor in wetenschappen, wiskunde, te verdedigen door:*

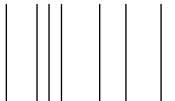
**Birhanu Teshome Ayele**

*Promotor: prof. dr. Geert Molenberghs*

*Copromotor: prof. dr. Cristina Sotito*



D/2012/2451/37



 Maastricht University

universiteit  
hasselt  
KNOWLEDGE IN ACTION



# Samenvatting

Een zekere mate van ontbrekende gegevens is onvermijdelijk in studies met gecorreleerde gegevens. Ontbrekende gegevens kunnen tot vertekening leiden en dus kunnen conclusies, gebaseerd op dit type van studie, op hun beurt foutieve interpretaties tot gevolg hebben. Een ander gevolg van dit fenomeen is het verlies aan informatie. Uiteraard is de mate waarin de informatie afneemt verwant aan de hoeveelheid ontbrekende gegevens. Er is ook de invloed van de analyse-methode. In toegepast onderzoek, zoals bijvoorbeeld in klinische studies, worden traditioneel twee paden bewandeld. Het eerste is de studie zodanig op te zetten dat het risico op ontbrekende gegevens wordt geminimaliseerd. De tweede methode, waarop we in dit werk inzoomen, richt zich op het formuleren van adequate analysemethodologie. Als dusdanig is het de bedoeling vertekening te vermijden of minstens te verminderen.

De keuze van het inferentiële paradigma hangt nauw samen met de aard van het mechanisme dat ontbrekende gegevens veroorzaakt. Het is gebruikelijk het mechanisme te classificeren in overeenstemming met de terminologie van Little and Rubin (2002). Wanneer het mechanisme afhangt van geobserveerde noch niet-geobserveerde gegevens, spreken we van *missing completely at random* (MCAR). Wanneer het afhangt van geobserveerde doch niet verder van niet-geobserveerde gegevens, wordt de term *missing at random* (MAR) gebruikt. In alle andere gevallen spreken we van *missing not at random* (MNAR).

Een belangrijk concept in dit verband is *ignorability* (Rubin, 1976). Het geeft aan dat, onder bepaalde voorwaarden, het mechanisme dat ontbrekende gegevens veroorzaakt niet hoeft gemodelleerd te worden. Dit is uiteraard alleen mogelijk wanneer we enkel aan aspecten van het respons-model geïnteresseerd zijn. Bij gebruik van *likelihood* en Bayesiaanse methoden, kan *ignorability* toegepast worden onder MCAR en MAR. Daarentegen, in een frequentistisch kader, is de strenge MCAR veronderstelling een voldoende voorwaarde voor *ignorability*.

Voor niet normaal verdeelde gegevens met middellange tot lange meetreeksen kan

zogenaamde *direct likelihood* tot moeilijk te manipuleren likelihood functies leiden. Om hieraan te verhelpen werden een reeks alternatieven voorgesteld, zoals *generalized estimating equations* (hierna aangeduid met GEE) en *pseudo-likelihood* (PL). Waar GEE de score vergelijkingen vervangen door alternatieve, eenvoudiger functies, zal PL de *likelihood* zelf vervangen door een eenvoudigere vorm. Wanneer gegevens onvolledig zijn moeten we opletten met GEE en PL, ze zijn immers frequentistisch en dus enkel onder MCAR gegarandeerd geldig. Een oplossing hiervoor is het gebruik van gewichten of het aanwenden van *multiple imputation* (bijv., Paik, 1997). We kunnen ook gebruik maken van een dubbel-robuste versie onder MAR. Robins, Rotnitzky and Zhao (1995) ontwikkelden gewogen GEE (WGEE), samen met een aantal verfijningen, om GEE toe te laten onder MAR, en zelfs onder MNAR. De methode is gebaseerd op de concepten van Horvitz en Thompson (Cochran, 1977). Metingen worden gewogen met het inverse van de kans om geobserveerd te worden. In het recente verleden werd WGEE uitgebreid met een dubbel robuuste versie (DR-GEE), waar het concept van gewichten wordt aangevuld met het gebruik van een predictief model voor de niet-geobserveerde gegevens, gegeven wat werd waargenomen.

Molenberghs and Kenward (2007) en Beunckens, Sotto and Molenberghs (2008) vergeleken WGEE met een *multiple imputation* gebaseerde versie (MI-GEE), beide zinvolle manieren om de methodologie te gebruiken onder MAR. Ook al is er heel wat werk besteed aan de vergelijking van WGEE en MI-GEE, veel minder aandacht is uitgegaan naar de dubbel robuuste versies. In Hoofdstuk 4 worden empirische vergelijkingen gemaakt tussen gewogen, dubbel robuuste en MI-gebaseerde versies. Gelijkaardige aanpassingen werden totnogtoe niet gedaan in het PL kader. In Hoofdstukken 5 en 6 worden een reeks aanpassingen aan standaard PL voorgesteld, die geldigheid onder MAR garanderen. Via simulaties wordt de performantie bestudeerd. In Hoofdstuk 7 wordt de efficiëntie en robuustheid van verscheidene GEE en PL versies bestudeerd en vergeleken, met klemtoon op marginale modellen voor niet-normaal verdeelde longitudinale gegevens met uitval. In Hoofdstuk 8 wordt een nieuwe methode voor het imputeren van ontbrekende gegevens in klinische studies voorgesteld, voor zowel de actieve als de placebo groep, daarbij gebruikmakend van gegevens uit de placebo groep. We verwijzen hier naar als *placebo multiple imputation* (pMI). De methode is ontwikkeld en de performantie ervan bestudeerd.

In wat volgt overlopen we kort de onderscheiden hoofdstukken van deze thesis.

Hoofdstukken 2 en 3 bevatten het inleidend materiaal, nodig voor de rest van de tekst. De verscheidene datasets gebruikt doorheen het werk worden ingeleid in Hoofdstuk 2. In Hoofdstuk 3 geven we een overzicht van standaard terminologie, algemene eigenschappen en een overzicht van bestaande methodologie om ontbrekende

gegevens te analyseren. De veelgebruikte methoden worden in vier delen onderveeld: eenvoudige *ad hoc* methoden, imputatie, maximum likelihood en niet-likelihood methoden.

De twee eenvoudige *ad hoc* methoden zijn: *complete case analysis* (CC), waarbij alle onvolledige reeksen worden weggelaten en *last observation carried forward* (LOCF), waarbij de laatst waargenomen meting als substituut wordt gebruikt voor alle latere metingen die niet meer werden waargenomen. Zogenaamde voordelen van deze methoden zijn computationele eenvoud, het vermijden van een volledig longitudinaal model (bijv. wanneer de onderzoeksvraag gesteld is in termen van de laatste meting) en, voor LOCF, compatibiliteit met het *Intention-to-Treat* (ITT) principe, omdat gegevens van alle gerandomiseerde patiënten kunnen gebruikt worden. Het gebruik van dergelijke methoden vermindert sterk en meer principiële methoden, geldig onder MAR, nemen toe in gebruik. Voorbeelden hiervan omvatten *multiple imputation* Rubin (1987), *maximum likelihood* en methoden niet op de likelihood gebaseerd. Het panel genaamd "*Handling of Missing Data in Clinical Trials*" werd samengebracht onder auspiciën van de *National Academy of Sciences* (NAS) van de Verenigde Staten van Amerika, op verzoek van de *U.S. Food and Drug Administration* (National Research Council, 2010). Het panel waarschuwde tegen het gebruik van de *ad hoc* methoden en pleitte voor het gebruik van principiële methoden die geldig zijn onder MAR en tegelijk makkelijk te implementeren in standaard commerciële statistische software.

GEE is een niet-likelihood methode en vereist dus aanpassingen om geldig te zijn onder MAR. Gewogen GEE (WGEE), *multiple imputation* gecombineerd met GEE (MI-GEE) en dubbel robuuste GEE (DR-GEE) zijn alle attractieve aanpassingen. Om tot consistente schattingen te leiden dient WGEE vergezeld te gaan van correct gespecificeerde gewichten, terwijl voor MI-GEE een correct gespecificeerd imputatiemodel vereist is. Voor DR-GEE dient het model voor de gewichten of het predictieve model correct te worden geformuleerd, doch niet noodzakelijk beide. In Hoofdstuk 4 vergelijken we WGEE, MI-GEE en DR-GEE via simulaties, onder een reeks correct en foutief gespecificeerde modellen. Uit deze studie blijkt dat, vooral voor kleine steekproeven, WGEE eerder inefficiënt is. De methode is ook gevoelig aan foutieve specificatie van het dropout model. Daarentegen is MI-GEE eerder robuust aan verkeerd gespecificeerde imputatiemodellen, ondanks het feit dat het imputatiemodel zich in het hart van de methodologie bevindt. We zagen ook dat MI-GEE tot eerdere preciese schatters leidt, met een lichte verbetering wanneer we overstappen op augmentatie DR-GEE (DR-GEE(Aug)). Als het dropout model correct is geformuleerd maar het imputatiemodel niet, dan gebeurt het dat MI-GEE lichtjes beter

is, maar soms is ook DR-GEE(Aug) de winnaar, afhankelijk van de specifieke setting. Daarentegen, wanneer het dropout model verkeerd werd geformuleerd terwijl de rest correct is, dan zorgt MI-GEE(Aug) voor een substantiële verbetering vergeleken met WGEE. Tenslotte, in het geval dat zowel het imputatie als het dropout model foutief werd opgezet, dan is DR-GEE(Aug) duidelijk beter dan de competitie. De methode doet het dan met name beter dan WGEE, met licht toegenomen precisie ten overstaan van MI-GEE. Samengevat, de methoden gestoeld op MI, d.w.z. MI-GEE en DR-GEE(Aug), zijn aan te bevelen voor de praktijk. Het is daarbij aangewezen de variabiliteit in de gewichten mee in rekening te brengen.

PL vormt een praktisch alternatief voor *maximum likelihood*, in het bijzonder voor toepassingen met complexe likelihood functies. In dit werk onderzoeken we PL methodologie voor het marginaal modelleren van onvolledige niet-normale (bijv. binaire) longitudinale gegevens. Aan de ene kant wordt de numerieke en computationele complexiteit van likelihood omzeild door paarsgewijze PL. Aan de andere kant wordt onvolledigheid aangepast via *inverse probability weighting* (IPW), wat tot enkelvoudig en dubbel robuuste schatters leidt.

In tegenstelling tot GEE kan PL makkelijke associatie incorporeren (Yi, Zeng and Cook, 2011; He and Yi, 2011). Men kan een onderscheid maken tussen marginale en conditionele PL. De methode is verwant aan maar verschillend van maximum likelihood. Wegens dat verschil is het niet *a priori* zeker dat de methode ook geldig is onder MAR, zelfs al is het mogelijk dat het in bepaalde gevallen wel geldt. Rubin (1976) leidde voorwaarden af voor *ignorability* die voldoende zijn doch niet altijd nodig. In Hoofdstuk 5 geven we aan dat een correctie nodig zijn om PL te kunnen gebruiken onder MAR, en dat zowel enkelvoudig als dubbel robuuste versies mogelijk zijn. In alle gevallen zijn ze praktisch toepasbaar. We schetsen een algemeen kader en leggen de klemtoon op PL. De methodologie gebruikt concepten van *inverse probability weighting* (IPW) en dubbele robuustheid (DR). Na de algemene ontwikkelingen leggen we de focus op de PL setting; zowel marginale als conditionele versies worden beschouwd. In het algemeen vereist DR een model voor de gewichten en een predictief model. In het geval van PL zijn er belangrijke speciale gevallen waar het ene, het andere, of zelfs beide overbodig worden. Dit is het geval voor *exchangeability* (EX).

Kang and Schafer (2007) toonden empirisch aan dat er omstandigheden zijn waar zware vertekening optreedt indien zowel het gewichtmodel als het predictieve model lichtjes verkeerd werden gespecificeerd. Ze toonden eveneens aan dat grote fluctuaties in de gewichten het risico op vertekening vergroten. Dus moet de analist, zoals het geval met elke statistische methoden, zich goed bewust zijn van relatieve voor- en nadelen. Het is in die zin zeer relevant dat in een aantal gevallen, beschouwd in



Hoofdstuk 5, de gewichten uit de vergelijkingen verdwijnen. Op die manier vergroot robuustheid.

We vertrekken van de uitdrukkingen in Hoofdstuk 6 en in Molenberghs *et al.* (2011); hun tegenhangers voor paarsgewijze marginale PL worden afgeleid, stoelend op IPW en DR. We bestuderen de performantie van de enkelvoudig en dubbel robuuste schatters van Molenberghs *et al.* (2011), besproken in Hoofdstuk 5 van deze thesis, via simulaties en via de analyse van gegevens. We leggen de klemtoon op binaire gegevens, maar de methodologie is uiteraard veel algemener bruikbaar. De simulaties gaven aan dat de enkelvoudige robuuste versie minstens even efficiënt is dan maximum likelihood, terwijl de efficiëntie nog toeneemt bij DR. De efficiëntie neemt beduidend toe naarmate er meer ontbrekende gegevens zijn. Onder volledige EX, of wanneer de eigenschap benaderend geldt, is de naïeve *available case* methode even efficiënt als de DR versies. Dit volgt omdat onder EX geobserveerde en niet-geobserveerde componenten uit de patiënt-geschiedenis met elkaar kunnen omgewisseld worden, waardoor de naïeve schatter consistent wordt. Uiteraard is dit een erg aantrekkelijke eigenschap, omdat we ons op DR kunnen beroepen zonder de nood aan IPW en predictieve modellen. Wanneer de correlatiestructuur verschilt van EX, verdwijnen de gewichten nog steeds onder AC, maar dienen de verwachtingswaarden berekend te worden.

GEE, PL en hun uitbreidingen zorgen voor consistente en asymptotisch normaal verdeelde schatters, op voorwaarde dat de correcte variantieschatter en een juist gespecificeerd model voor ontbrekende gegevens werden geformuleerd. In Hoofdstuk 7 evalueren we de relatieve voor- en nadelen van PL en GEE, via simulaties. Naast enkelvoudig en dubbel robuuste PL versies, ingeleid in Hoofdstukken 5 en 6, combineren we hier ook MI en PL. In dat geval wordt met MAR rekening gehouden door de ontbrekende gegevens meervoudig te imputeren via een parametrisch model. Daarna worden de vervolledigde gegevens geanalyseerd via PL. De methode wordt aangeduid door MI-PL. IPW, MI-PL, DR-PL en DR-GEE worden vergeleken onder een juist gespecificeerd model, een lichtjes foutief gespecificeerd model, en een model waarin zware specificatie fouten werden gemaakt. De fouten slaan op de dropout en de imputatiemodellen. Vergeleken bij GEE is PL ongeveer even efficiënt als maximum likelihood, terwijl de bijkomende computationele kost minimaal is. Indien minstens een deel van de wetenschappelijke interesse gericht is op de associatie, dan wordt PL nog een attractievere methoden. Het is niet verwonderlijk dat MI-GEE en MI-PL vergelijkbare resultaten leveren.

Het wordt vaak aangehaald, in de context van longitudinale klinische studies, dat het volgen van patiënten (ook na dropout uit de behandeling) nuttig is om de redenen voor dropout te begrijpen en te modelleren. Dergelijke data zijn nochtans

niet zo eenvoudig om te verzamelen; het risico op *confounding* ligt altijd op de loer. Inderdaad, het is bijvoorbeeld niet ongevoerd dat een patiënt naar een andere behandeling overstapt na dropout. Daarom is het imputeren vanuit de placebo groep in een dergelijk geval, zowel voor de behandelde als voor de placebogroep, een aantrekkelijk alternatief. Op die manier ontstaan gegevens om effectiviteit te schatten; men sluit daarbij de mogelijkheid dat patiënten geen voordeel doen met het geneesmiddel na dropout, niet uit. Dit is intuïtief duidelijk, omdat placebo *de facto* overeenkomt met het gebrek aan effect. De methode is daarom een nuttig alternatief voor LOCF en BOCF, die jammer genoeg in deze context als methoden worden gezien om effectiviteit te bepalen. In Hoofdstuk 8 wordt een nieuwe methode voorgesteld om gegevens te imputeren in zowel de behandelde als onbehandelde groep, gebruik makend van de placebo groep. We verwijzen hier naartoe als *placebo multiple imputation* (pMI). In de context of effectiviteit veronderstelt pMI geen farmacologisch voordeel van het geneesmiddel na dropout. In de context van efficaciteit is pMI een specifieke vorm van MNAR, die verwacht wordt van een conservatieve schatting op te leveren.

In een simulatiestudie met 18 scenario's wordt pMI vergeleken met LOCF, BOCF, DL en MI. pMI leidt over het algemeen tot vertekende schatters voor effectiviteit en conservatieve schatters voor efficaciteit. Daarnaast dient opgemerkt dat betrouwbaarheidsintervallen typisch groter zijn dan nominaal verwacht, door de vertekening in de schatter van Rubin voor de variantie. We weten dat de variantie van Rubin de werkelijke variantie in een aantal gevallen overschat (Wang and Robins, 1998; Robins and Wang, 2000). De performantie hangt af van de context en men moet voorzichtig zijn bij het gebruik, in het bijzonder als de imputatie- en analysemodellen discrepant zijn. De werkelijke variantie overschatten door pMI leidt tot minder efficiënte maar nog wel steeds geldige betrouwbaarheidsintervallen. Dit is nog altijd aanvaardbaar, omdat het gebruikelijk is in klinische studies van eerder naar het conservatieve over te hellen. Merk op dat de imputatie in beide groepen vanuit hetzelfde model leidt tot een positieve correlatie tussen de geschatte gemiddelden in beide groepen. Daardoor daalt de variantie van de schatter voor het behandelings-contrast. Dit effect wordt nochtans niet opgevangen door de schatter van Rubin. Daardoor zien we dalende verhoudingen van gesimuleerde varianties tegenover model-varianties, onder pMI, bij het schatten van het behandelingseffect.

Daarentegen zijn LOCF en BOCF conservatief in sommige scenarios en liberaal in andere, zowel voor effectiviteit als voor efficaciteit. Zoals verwacht leiden DL en gewone MI tot onvertkende schatters voor efficaciteit en lichtjes overschatte effectiviteit in gevallen waarbij een effect van behandeling werkelijk aanwezig is. In gevallen zonder een effect van behandeling, waar de effecten dus gelijk zijn aan nul,

leiden DL en MI tot overtekende schatters voor effectiviteit en efficaciteit. In onze studie werd geen onderscheid gemaakt tussen redenen voor dropout; dropout volgende uit één enkel model. In de toekomst zou het nuttig zijn van het effect van verscheidene redenen voor dropout verder te exploreren.



# Acknowledgements

This work would not have been accomplished without the help and encouragement of a number of people. These few lines are intended to thank some of them.

First and foremost, I would like to record my gratitude to my promoters Prof. Geert Molenberghs and Prof. Cristina Sotto for their inspiring guidance through out the course of the PhD work. I am very grateful to Geert, for always understanding me, his thorough guidance, and words of encouragement. Geert, I am truly fortunate to have had the opportunity to work with you. I wish to express my heartfelt gratitude and profound thanks to my co-promotor, Cristina Sotto. Tina, thanks for unconditionally willing to help me, the valuable comments on the initial draft of this thesis and being there for me.

The last section of the thesis is based on a research collaboration with Dr. Craig H. Mallinckrodt and Dr. Ilya Lipkovich from Eli Lilly. I greatly acknowledge both for many helpful discussions. The contributions of all our co-authors have been invaluable and I owe my sincere gratitude to them.

The financial support of the “Methusalem Consortium”, an interuniversity collaboration between the Vaccine & Infectious Disease Institute (Vaxinfectio) at the University of Antwerp and the CenStat at Hasselt University, is greatly acknowledged.

I extend my gratitude to the whole I-Biostat colleagues for their distinguishable support in various forms. D4 office mates: Girma, Amin, and Kim it is a pleasure to share the office with you.

Many thanks go to my close friends (in alphabetical order): *Abe*<sup>2</sup>, Alem, Abiot, Bedilu, Haile, Helen, Mihreteab (Dn.) and Yonas with their respective families. Thanks for your companion, support and being with us at the most difficult times. The weekends we’d spent with my Godson, Nathan Abiot (Fiqre Mariam), have been invaluable. Fiqre Mariam thanks for lighting our lives and we love you so much.

I am grateful for the spiritual nourishment I have received over the six years of

my stay in Belgium. My special thanks go to members of: D/M/K/Kidane-Mihret parish council and the laity, MK Belgium and St. Mary association in Hasselt.

Words fail me to express my appreciation to my wonderful wife Selamawit whose dedication, love and persistent confidence in me, has taken the load off my shoulder. I am so blessed to have you in my life and thank you for your unwavering support and encouragement over the years.

I am deeply grateful to my parents, brothers, sisters, my in-laws, friends and relatives. Mulu, Shimels, and Hagere thanks for looking after mom while I was away. I want to acknowledge my in-laws specially: Andarge, Mulugeta, Chuchu and Etete with their respective families for their tremendous support throughout my study and in life. Dad and Abu (mom), I praise God for giving you the health to see the success of your son!

Last in this list but first in my heart, I owe it all to the Almighty God for granting me the wisdom, health and strength to undertake this research task and enabling me to its completion.

Thank you all!!!  
Birhanu Teshome  
27 September, 2012  
Diepenbeek

# List of Publications

- Molenberghs, G., Kenward, M. G., Verbeke, G., and **Birhanu, T.** (2011) ‘Pseudo-likelihood Estimation for Incomplete Data’. *Statistica Sinica*, **21(1)**, 187-206.
- **Birhanu, T.**, Molenberghs, G., Sotto, C., and Kenward, M.G.(2011). ‘Doubly Robust and Multiple-Imputation-Based Generalized Estimating Equations’. *Journal of Biopharmaceutical Statistics*, **21(2)**, 202-225.
- **Birhanu, T.**, Lipkovich, I, Molenberghs, G, and Mallinckrodt,H. C.(2013). ‘A Multiple Imputation Based Approach to Sensitivity Analyses and Effectiveness Assessments in Longitudinal Clinical Trials’. Submitted to *Journal of Biopharmaceutical Statistics*, accepted.
- G. Van De Putte, J. Vlasselaer, **B. Teshome**, A. Gaddah, T. Burzykowski, J.C. Schobbens, G. Benijts, E.T.M. De Jonge (2010). ‘Outcome of per protocol best-evidence based routine breast cancer care in a large regional hospital in Belgium: the importance of a prospective database in quality assurance’. *Facts, Views and Vision in Obstetrics and Gynaecology*, **2(2)**, 119-124.
- **Birhanu, T.**, Sotto, C., Molenberghs, G., Kenward, M.G. and Verbeke, G. (2012). ‘Doubly Robust Composite Likelihood for Hierarchical Categorical Data’. *Submitted to Statistica Sinica*.
- **Birhanu, T.**, Sotto, C., Molenberghs, G., Kenward, M.G. and Verbeke, G. (2012). ‘Comparison of Pseudo-likelihood and Generalized Estimating Equations for Incomplete Data’. *In preparation*.





# Contents

|   |              |
|---|--------------|
| <b>Table of Contents</b>                            | <b>xiii</b>  |
| <b>List of Tables</b>                               | <b>xvii</b>  |
| <b>List of Figures</b>                              | <b>xxi</b>   |
| <b>List of Abbreviations</b>                        | <b>xxiii</b> |
| <b>1 Introduction</b>                               | <b>1</b>     |
| <b>2 Motivating Examples</b>                        | <b>7</b>     |
| 2.1 Onychomycosis Trial . . . . .                   | 7            |
| 2.2 Analgesic Trial . . . . .                       | 9            |
| 2.3 Depression Trial . . . . .                      | 11           |
| 2.4 National Toxicology Program Data . . . . .      | 12           |
| <b>3 Fundamental Concepts of Incomplete Data</b>    | <b>15</b>    |
| 3.1 Incomplete Data . . . . .                       | 15           |
| 3.2 Modeling Frameworks . . . . .                   | 16           |
| 3.3 Missing Data Mechanisms . . . . .               | 17           |
| 3.4 Model Families . . . . .                        | 18           |
| 3.4.1 A Classic Model for Continuous Data . . . . . | 18           |
| 3.5 Models for non-Gaussian data . . . . .          | 19           |
| 3.5.1 Generalized linear mixed models . . . . .     | 20           |
| 3.5.2 Generalized Estimating Equations . . . . .    | 21           |
| 3.5.3 Pseudo-likelihood Estimation . . . . .        | 22           |
| 3.5.4 The Bahadur Model . . . . .                   | 27           |
| 3.6 Methodology for Incomplete Data . . . . .       | 27           |

|          |  |           |
|----------|--|-----------|
| 3.6.1    | Simple <i>ad hoc</i> Methods . . . . .   | 28        |
| 3.6.2    | Multiple Imputation . . . . .  | 29        |
| 3.6.3    | Maximum Likelihood Estimation . . . . .  | 31        |
| 3.6.4    | Non-Likelihood Estimation . . . . .  | 31        |
| 3.6.4.1  | GEE and PL . . . . .   | 32        |
| 3.6.4.2  | Inverse Probability Weighting and Double Robustness . . . . .                      | 32        |
| <b>4</b> | <b>Multiple Imputation Based Doubly Robust Generalized Estimating Equations</b>    | <b>35</b> |
| 4.1      | Extensions of GEE under MAR . . . . .  | 36        |
| 4.1.1    | Weighted Generalized Estimating Equations . . . . .                                | 36        |
| 4.1.2    | Multiple Imputation based Generalized Estimating Equations . . . . .               | 38        |
| 4.1.3    | Doubly Robust Generalized Estimating Equations . . . . .                           | 38        |
| 4.1.4    | Precision Estimation . . . . .   | 42        |
| 4.2      | Simulation Study . . . . .   | 42        |
| 4.2.1    | Data-generating Models . . . . .   | 43        |
| 4.2.2    | Design of the Simulation Study . . . . .   | 45        |
| 4.2.3    | Results . . . . .  | 45        |
| 4.3      | Analysis of the Toenail Data . . . . .   | 57        |
| 4.4      | Discussion . . . . .   | 59        |
| <b>5</b> | <b>Pseudo-likelihood Estimation for Incomplete Data</b>                            | <b>61</b> |
| 5.1      | General Forms of Estimating Equations for Incomplete Data . . . . .                | 62        |
| 5.2      | Pseudo-likelihood Estimating Equations . . . . .                                   | 66        |
| 5.2.1    | Pairwise (Pseudo-)likelihood . . . . .   | 68        |
| 5.2.2    | Marginal (Pairwise) Pseudo-likelihood for Gaussian Data . . . . .                  | 70        |
| 5.2.3    | Conditional Pseudo-likelihood for Binary Data . . . . .                            | 74        |
| 5.3      | Analysis of Case Studies . . . . .   | 78        |
| 5.3.1    | The Onychomycosis Trial . . . . .  | 79        |
| 5.3.2    | The National Toxicology Program Data . . . . .                                     | 81        |
| 5.4      | Discussion . . . . .   | 83        |
| <b>6</b> | <b>Efficient Doubly Robust Pseudo-likelihood for Hierarchical Categorical Data</b> | <b>85</b> |
| 6.1      | Estimating Equations for Pairwise Likelihood . . . . .                             | 86        |
| 6.2      | Marginal Pseudo-likelihood for Binary Data . . . . .                               | 88        |
| 6.3      | Simulation Study . . . . .   | 90        |
| 6.3.1    | Results . . . . .  | 91        |

---

|          |   |            |
|----------|---|------------|
| 6.4      | Analysis of the Analgesic Trial . . . . .   | 95         |
| 6.5      | Discussion . . . . .  | 98         |
| <b>7</b> | <b>Comparison of Pseudo-likelihood and Generalized Estimating Equations for Incomplete Data</b>                                   | <b>101</b> |
| 7.1      | Inverse Probability Weighting and Multiple Imputation-based Methods   | 102        |
| 7.2      | Simulation Study . . . . .  | 104        |
| 7.2.1    | Results . . . . .   | 105        |
| 7.3      | Analysis of the Analgesic Trial . . . . .   | 113        |
| 7.4      | Discussion . . . . .  | 114        |
| <b>8</b> | <b>A Multiple Imputation Based Approach to Sensitivity Analysis and Effectiveness Assessments in Incomplete Longitudinal Data</b> | <b>117</b> |
| 8.1      | Placebo Multiple Imputation . . . . .   | 121        |
| 8.2      | Simulation Study . . . . .  | 123        |
| 8.2.1    | Results . . . . .   | 125        |
| 8.2.1.1  | Effectiveness . . . . .   | 126        |
| 8.2.1.2  | Efficacy . . . . .  | 132        |
| 8.3      | Analysis of the Depression Trial . . . . .  | 133        |
| 8.4      | Discussion . . . . .  | 134        |
| <b>9</b> | <b>General Conclusions and Future Research</b>  | <b>137</b> |
|          | <b>Bibliography</b>   | <b>143</b> |
| <b>A</b> | <b>Simulation Results for Efficacy Estimand</b>   | <b>151</b> |



# List of Tables

|     |   |    |
|-----|---|----|
| 2.1 | <i>Toenail Data. Number and percentage of patients (<math>N</math>) with severe toenail infection, for each treatment arm separately. . . . .</i>   | 8  |
| 2.2 | <i>Toenail Data. Number of available repeated measurements per subject, for each treatment arm separately. . . . .</i>  | 8  |
| 2.3 | <i>Analgesic Trial. Absolute and relative frequencies of the five GSA categories for each of the four follow up times. . . . .</i>  | 10 |
| 2.4 | <i>Analgesic Trial. Overview of missingness patterns and frequencies with which they occur. ‘O’ indicates observed and ‘M’ indicates missing. . .</i>   | 11 |
| 2.5 | <i>Depression Trials. Completion Rates and Main Reasons for Study Discontinuation . . . . .</i>   | 12 |
| 2.6 | <i>NTP data: Developmental Toxicity Study (DEHP). Summary data by dose group. . . . .</i>   | 13 |
| 4.1 | <i>Simulation study for MAR (GM I): Everything correctly specified. . . .</i>   | 47 |
| 4.2 | <i>Simulation study for MAR (GM I): Misspecified imputation model. (Misspecification in the form of omission of the previous outcome, <math>y_{i,j-1}</math>, from the imputation model). . . . .</i>                               | 48 |
| 4.3 | <i>Simulation study for MAR (GM I): Misspecified dropout model. (Misspecification in the form of omission of the previous outcome, <math>y_{i,j-1}</math>, from the dropout model). . . . .</i>                                     | 50 |
| 4.4 | <i>Simulation study for MAR (GM I): Both models misspecified. (Misspecification in the form of omission of the previous outcome, <math>y_{i,j-1}</math>, from both the dropout and imputation models). . . . .</i>                  | 51 |
| 4.5 | <i>Simulation study for MNAR (GM II): Singly misspecified dropout and imputation models. (Misspecification in the form of omission of the current outcome, <math>y_{ij}</math>, from both the dropout and imputation models). .</i> | 53 |

|     |  |    |
|-----|--|----|
| 4.6 | <i>Simulation study for MNAR (GM II): Doubly misspecified dropout and imputation models. (Misspecification in the form of omission of the current, <math>y_{ij}</math>, and previous, <math>y_{i,j-1}</math>, outcomes from both the dropout and imputation models.)</i>   | 54 |
| 4.7 | <i>Simulation study for MNAR (GM II): Singly misspecified dropout model and doubly misspecified imputation model. (Misspecification in the dropout model in the form of omission of the current outcome, <math>y_{ij}</math>; misspecification in the imputation model in the form of omission of the current, <math>y_{ij}</math>, and previous, <math>y_{i,j-1}</math>, outcomes).</i> | 56 |
| 4.8 | <i>Simulation study for MNAR (GM II): Singly misspecified imputation model and doubly misspecified dropout model. (Misspecification in the imputation model in the form of omission of the current outcome, <math>y_{ij}</math>; misspecification in the dropout model in the form of omission of the current, <math>y_{ij}</math>, and previous, <math>y_{i,j-1}</math>, outcomes).</i> | 57 |
| 4.9 | <i>Toenail Data. Severity of toenail infection. Parameter estimates (empirically corrected standard errors ignoring variability in the weights; empirically corrected standard errors acknowledging variability in the weights).</i>   | 58 |
| 5.1 | <i>Toenail Data. (Unaffected nail length outcome). Parameter estimates (purely model-based standard errors; empirically corrected standard errors) for full likelihood, and naive, singly robust, and doubly robust pairwise likelihood.</i>   | 80 |
| 5.2 | <i>Developmental Toxicity Study (DEHP). Parameter estimates (standard errors) for full likelihood, and naive, singly robust, and doubly robust pseudo-likelihood.</i>  | 82 |
| 6.1 | <i>Simulation Study (Setting 1). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for naive, singly and doubly robust pairwise likelihood and full likelihood.</i>   | 93 |
| 6.2 | <i>Simulation Study (Setting 2). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for naive, singly and doubly robust pairwise likelihood and full likelihood.</i>   | 94 |

|     |   |     |
|-----|---|-----|
| 6.3 | <i>Analgesic Trial. Parameter estimates (empirically-corrected standard errors) for naive, singly and doubly robust pairwise likelihood and for full likelihood. . . . .</i>  | 96  |
| 7.1 | <i>Simulation Study (correct imputation and dropout models). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for GEE and pairwise PL. . . . .</i>                            | 106 |
| 7.2 | <i>Simulation Study (incorrect imputation model). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for singly and doubly robust GEE and pairwise PL. . . . .</i>              | 108 |
| 7.3 | <i>Simulation Study (incorrect dropout model). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for singly and doubly robust GEE and pairwise PL. . . . .</i>                 | 110 |
| 7.4 | <i>Simulation Study (incorrect imputation and dropout models). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for singly and doubly robust GEE and pairwise PL. . . . .</i> | 112 |
| 7.5 | <i>Analgesic Trial. Parameter estimates (empirically-corrected standard errors) for singly and doubly robust versions of GEE and pairwise Pseudo-likelihood. . . . .</i>  | 114 |
| 8.1 | <i>Estimands and estimators commonly used to assess efficacy and effectiveness in clinical trials . . . . .</i>   | 119 |
| 8.2 | <i>Variance-correlation matrix of the outcome over time (variances on the diagonal, correlations off-diagonally) . . . . .</i>  | 124 |
| 8.3 | <i>Visit-wise population means in the placebo group . . . . .</i>   | 125 |
| 8.4 | <i>Bias in estimates of the effectiveness estimand for the analysis of change from baseline . . . . .</i>   | 127 |
| 8.5 | <i>Mean square error in estimates of the effectiveness estimand for the analysis of change from baseline. . . . .</i>   | 129 |
| 8.6 | <i>Confidence interval coverage for the effectiveness estimand for the analysis of change from baseline . . . . .</i>   | 130 |
| 8.7 | <i>Ratio of Simulation Variance versus Model-Based Variance of the effectiveness estimand for the analysis of change from baseline . . . . .</i>  | 131 |

|     |   |     |
|-----|---|-----|
| 8.8 | <i>Rejection rates in assessing null hypothesis of no treatment difference for the analysis of change from baseline . . . . .</i> | 132 |
| 8.9 | <i>Endpoint treatment contrasts by analytic method from the actual clinical trial dataset . . . . .</i>                           | 133 |
| A.1 | <i>Bias in estimates of the efficacy estimand for the analysis of change from baseline . . . . .</i>                              | 152 |
| A.2 | <i>Relative bias in estimates of the efficacy estimand for the analysis of change from baseline . . . . .</i>                     | 153 |
| A.3 | <i>Mean square error in estimates of the effectiveness estimand for the analysis of change from baseline . . . . .</i>            | 154 |
| A.4 | <i>Confidence Interval coverage in estimates of the efficacy estimand for the analysis of change from baseline . . . . .</i>      | 155 |



# List of Figures

|     |   |     |
|-----|---|-----|
| 4.1 | <i>Toenail Data. Evolution of the percentage of severe toenail infections in the two treatment groups separately. . . . .</i>                           | 58  |
| 5.1 | <i>Toenail Data. Individual profiles of 30 randomly selected subjects in each of the treatment groups in the toenail experiment. . . . .</i>            | 79  |
| 8.1 | <i>Relative bias in estimates of the effectiveness estimand by direction of change and dropout pattern for the 0.3 effect size of the drug. . . . .</i> | 128 |



# List of Abbreviations

|        |  |
|--------|--|
| AC     | Available Case   |
| BOCF   | Baseline Observation Carried Forward                       |
| CC     | Complete Case  |
| CCA    | Complete Case Analysis                                     |
| CI     | Confidence Interval  |
| CP     | Complete Pairs   |
| CS     | Complete Sets  |
| DL     | Direct Likelihood  |
| DR     | Doubly Robust Estimating Equation                          |
| ES     | Effect Size  |
| GEE    | Generalized Estimating Equations                           |
| GLM    | Generalized Linear Model                                   |
| GLMM   | Generalized Linear Mixed Model                             |
| GSA    | Global Satisfaction Assessment                             |
| HAMD   | Hamilton Depression Rating Scale                           |
| ITT    | Intention-to-Treat   |
| IPW    | Inverse Probability Weighting                              |
| LOCF   | Last Observation Carried Forward                           |
| MAR    | Missing At Random  |
| MCAR   | Missing Completely At Random                               |
| MCMC   | Markov Chain Monte Carlo                                   |
| MI     | Multiple Imputation  |
| MI-GEE | Multiple Imputation based Generalized Estimating Equations |
| ML     | Maximum Likelihood   |
| MNAR   | Missing Not At Random                                      |

|      |  |
|------|--|
| MSE  | Mean Square Error                        |
| NTP  | National Toxicology Program              |
| PL   | Pseudo-likelihood                        |
| pMI  | Placebo Mutiple Imputation               |
| REff | Relative Efficiency                      |
| SE   | Standard Error                           |
| TDO  | Toenail Dermatophyte Onychomycosis       |
| WGEE | Weighted Generalized Estimating Equation |

# Chapter 1

## Introduction

The applied statistician frequently encounters correlated outcome data. Common situations include multivariate, clustered, and longitudinal data. In such settings, it frequently occurs that not all of the planned measurements of subject  $i$ 's outcome vector  $\mathbf{y}_i$  are actually observed, turning the statistical analysis into a missing data problem. For example, in a longitudinal study, a subject's response vector may terminate early for a number of reasons outside of the control of the investigator. Not only do missing data lead to potentially biased results but there can be a severe loss of power if the proportion of incompleteness is high. In extreme cases this may mean that there is insufficient data to draw any useful conclusions from the study. Hence, it is almost always necessary to reflect on the nature of the missingness process and its impact on inferences.

A common taxonomy for missing data, which is explained further in Chapter 3, distinguishes between missing data that are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). A non-response process is said to be *missing completely at random* (MCAR) if missingness is independent of both unobserved and observed data and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *non-random* (MNAR).

Early work on missing values was largely concerned with the practical consequence of missing data induced imbalance. Partially as a consequence of this, *ad hoc* methods such as complete case analysis (CC) and last observation carried forward (LOCF) have become popular, a status they somewhat unfortunately retain until this day. The reasons behind this are discussed in Section 3.6.1. Over the last three decades, a number

of developments have taken place, allowing the use of *missing at random* (MAR) based methods. These include the development of the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977), multiple-imputation strategies (Rubin, 1987), and so-called direct-likelihood or direct Bayesian analysis. These rest on *ignorability*, the property ensuring that such analyses are valid under MAR, supplemented with mild regularity conditions, even without explicitly modeling the missing data mechanism, provided that all incomplete sequences are subjected to analysis (Rubin, 1976; Little and Rubin, 2002; Molenberghs and Kenward, 2007; Fitzmaurice *et al.*, 2009). The practical implication for likelihood inference is that, as soon as a module is available to handle measurement sequences of unequal length, valid inferences are obtained *without any additional work*. Thanks to the availability of flexible software, such as the SAS procedures MIXED, GLIMMIX, and NLMIXED, and their counterparts in SPlus, R, and SPSS, for example, linear and generalized linear mixed models can be fitted to incomplete sets of data. There is a variety of other likelihood methods as well, of course, as reviewed by Fahrmeir and Tutz (2002) and Molenberghs and Verbeke (2005).

For the analysis of Gaussian data, the linear mixed model (Verbeke and Molenberghs, 2000) is widely accepted as the unifying framework for a variety of correlated settings, including longitudinal data. The model contains both subject-specific and autoregressive effects at the same time. Further, this general hierarchical model marginalizes in a straightforward way to a multivariate normal model with directly interpretable mean and covariance parameters, owing to the unique property of the normal distribution that both the marginal, and in fact also the conditional, distribution of a multivariate normal is again normal. This does not hold for the non-Gaussian case, since no natural analog to the multivariate normal distribution is available. Therefore, depending on which of the three model families is chosen, that is, the marginal, random-effects, or conditional model family, different models are conceivable.

The generalized linear mixed model (GLMM), an extension of generalized linear model for univariate data to the context of correlated measurement, is the most frequently used random-effects model for non-Gaussian outcomes. For non-Gaussian outcomes, GLMMs often do not admit straight forward marginalization. That is, due to the non-linear link functions that are usually adopted for non-Gaussian outcomes. As a result, GLMMs are most useful when the scientific goal is to make inferences about individuals rather than the study population.

For non-Gaussian outcomes, apart from random-effects models, also non-likelihood marginal models have become popular. Typical marginal models for binary data in-

---

clude the Bahadur model (Bahadur, 1961) and the multivariate Dale or global odds ratio model (Molenberghs and Lesaffre, 1994, 1999). Because these models specify, in principle, the full likelihood, they can be used to analyze incomplete data as well, under MAR assumptions and making use of the ignorability property. However, marginal models for non-Gaussian data imply complex and hard to manipulate likelihoods. In many practical settings involving outcome sequences of moderate to large length, direct likelihood may be prohibitive. Some authors have voiced concern over these models' vulnerability to misspecification.

As a response to these problems, a number of alternatives have been formulated, the most popular one undoubtedly being *generalized estimating equations* (GEE; Liang and Zeger, 1986; Dale, 1986; Molenberghs and Verbeke, 2005). By transforming the score equations into estimating equations, this method essentially allows confining attention to the specification of the first moments of the outcome sequence only (i.e., the mean structure), thereby circumventing the need to address the association structure while still leading to valid inferences. A number of variations to this theme exist, such as GEE2 (also specifying the second moments; Liang, Zeger and Qaqish, 1992) and alternating logistic regressions (Carey, Zeger and Diggle, 1993). When data are incomplete, GEE suffers from its frequentist nature and is in its basic form valid only under MCAR. Therefore, Robins, Rotnitzky and Zhao (1995) have developed so-called *weighted* generalized estimating equations (WGEE), as well as a number of refinements and extensions in subsequent papers, to allow usage of GEE under not only MAR, but even under MNAR settings. The method rests on Horvitz-Thompson ideas (Cochran, 1977), weighing contributions by the inverse probability of being observed. The method is elegant and enjoys good properties, but requires specification of a model for the weights. More recently, Birhanu *et al.* (2011) extended the WGEE towards so-called doubly robust estimating equations, in which the weighting idea is supplemented with the use of a predictive model for the unobserved responses, given the observed ones. These are further discussed in Section 3.6.4.2 and the subsequent chapters.

Another alternative approach which could be combined with either likelihood or non-likelihood methods is multiple imputation (MI), developed by Rubin (Rubin, 1987). The key idea of the MI procedure is to impute missing values several times, and the resulting complete data sets are analyzed using a standard method. Afterwards, the obtained inferences are combined into a single one. Regarding the missingness process, standard multiple imputation requires MAR to hold, even though extensions exist. A more detailed review is provided in Section 3.6.2.

In instances where more than the first moment would be of scientific relevance, it

is natural to model two or even more moments, such as in GEE2 (second-order GEE) or pseudo-likelihood. Next to GEE, pseudo-likelihood methods (PL; le Cessie and van Houwelingen, 1994; Geys, Molenberghs, and Lipsitz, 1998; Geys, Molenberghs, and Ryan, 1999; Aerts *et al.*, 2002) have become popular as an alternative to full likelihood, and therefore also to GEE and GEE2. In PL, rather than replacing the score equations with alternative functions, the likelihood itself is replaced by a more tractable function. In so-called marginal pseudo-likelihood, the likelihood for an  $n_i$ -dimensional response vector is replaced by the product of all pairs, or all triples, or all  $p$ -tuples (with  $p$  a pre-specified number, corresponding to the highest order of association that is still of scientific interest) of outcomes. Pseudo-likelihood approaches have become a practical alternative to full likelihood methods, particularly for applications involving complex likelihood forms. As it stays closer to full likelihood than the score equations, pseudo-likelihood approach is expected to be more efficient than the method of moments estimates. Computational and statistical performance (e.g., efficiency) have been shown to range from acceptably good to excellent (Geys, Molenberghs, and Lipsitz, 1998). Another important advantage of the PL approach is, in contrast with GEE, that it can easily account for association. Evidently, also conditional versions of pseudo-likelihood are possible, where the contributions take the form of conditional densities of a subset of outcomes within a sequence, given another subset of outcomes.

The main objective of this thesis is to develop efficient methodologies to handle incomplete data and provide insight on existing ones, with primary interest falling on the non-Gaussian setting. Categorical (binary) outcomes are very prominent in statistical practice and techniques for this type of data are less standard, because of the lack of a simple analogue to the normal distribution. Principled methodologies for incomplete data, surrounding pseudo-likelihood and generalized estimating equations have been given special attention. While ignorability would follow under likelihood inference, this is not generally true for non-likelihood methods such as GEE and PL. As a direct consequence of this, GEE has been adapted towards a weighted or doubly-robust version, as soon as an MAR process operates. Further investigation and empirical comparisons will be made on the singly and doubly robust versions of GEE. So far, no such modification has been devised for pseudo-likelihood based strategies. We proposed a suite of corrections to the standard form of pseudo-likelihood, to ensure its validity under MAR. Our corrections follow both single and double robustness ideas. An overview of each of the chapters within the thesis now follows.

In **Chapter 2**, we present an overview of the different data sets that will be used throughout this work. Thereafter, terminology, general concepts and overview



---

of some of the existing methodologies for handling incomplete data are outlined in **Chapter 3**. Maximum likelihood, GEE, and pseudo-likelihood inferential paradigms are sketched in Section 3.6.

As discussed earlier, in its basic form GEE is not valid unless the missingness mechanism is MCAR. WGEE and multiple imputation combined with GEE (MI-GEE) are alternative routes to make it valid under the more general MAR condition. Although considerable progress has been made to robustness assessment and comparison of these approaches (Molenberghs and Kenward, 2007; Beunckens, Sotto and Molenberghs, 2008), little has been done in comparing these approaches to the recently developed doubly robust methods. In **Chapter 4**, singly robust (WGEE and MI-GEE) and doubly robust GEE are compared for non-Gaussian longitudinal data. The relative merits of the different approaches are highlighted via a simulation study conducted under various misspecifications.

Since pseudo-likelihood is not a full likelihood method, there is no *a priori* guarantee that the method would be valid under MAR. (Note that Rubin's 1976 paper provided sufficient conditions only, without claiming their necessity.) Pseudo-likelihood's behavior under the assumption of MAR was not fully understood from a methodological standpoint and further investigation was recommended by different authors including Faes *et al.* (2008). In **Chapter 5**, we will show that a correction is necessary to allow for the use of pseudo-likelihood under MAR, and that both singly robust as well as doubly robust versions of PL can be considered. General forms of estimating equations for incomplete data, as well as specific forms for the case of pseudo-likelihood, will be presented and their validity established.

We take off from the expressions in Chapter 5 and present their counterparts for the case of pairwise *marginal* pseudo-likelihood in **Chapter 6**. The statistical performance of the methods introduced in Chapter 5 will be investigated by a modest simulation study and real data analysis focusing on marginal models for binary data.

GEE and PL are commonly encountered approaches to non-Gaussian data. Both GEE, and its extensions, and PL yield consistent and asymptotically normal estimators, provided an empirically corrected variance estimator and the appropriate missing data mechanism assumptions are used. In **Chapter 7** we will discuss the relative merits of Pseudo-likelihood (PL) and Generalized estimating equations (GEE) for incomplete data and illustrate them using simulation studies and a practical case study. A further numerical study of the effect of misspecification of weights and/or predictive models will be reported.

In longitudinal clinical trials, often times it is said that follow up data are useful in helping to understand dropout because patients will not continue to benefit from

a drug if they don't continue to take it. But follow up data are both difficult to obtain and often fraught with confounders as patients may switch to a new drug after dropping out. Hence, imputing from the placebo group for both the drug and placebo group may be a useful way to artificially create follow up data to estimate the effectiveness of a drug after allowing for the fact that patients get no benefit if they stop taking drug. This has intuitive appeal as placebo is the *de facto* estimate of no benefit. This approach is a useful substitute for LOCF and BOCF, which are often interpreted in this effectiveness context. In **Chapter 8**, a novel method of imputing missing data for both the drug and placebo groups from the placebo group, referred to as placebo multiple imputation (pMI), will be assessed as an estimator of effectiveness and as a worst reasonable case sensitivity analysis in assessing efficacy. pMI is compared with the LOCF, BOCF, DL and MI under different scenarios.

Finally, in **Chapter 9** general conclusions and recommendations for future research regarding the different methods considered within the thesis are presented. Simulation studies results excluded from the main text are provided in the Appendix.

## Chapter 2

# Motivating Examples

This chapter introduces the data sets which will be used as key examples throughout this thesis. A two-armed clinical trial in patients treated for toenail infection is introduced in section 2.1. A single-arm clinical trial conducted in patients with chronic pain, the analgesic trial, is introduced in Section 2.2. Section 2.3 is devoted to a multi-center clinical trial, in which the Hamilton depression rating scale total score (HAMD17 score) from patients with major depressive disorder treated with an experimental drug and patients treated with placebo are compared. While the previous case studies are longitudinal in nature, a clustered data example from the development of toxicology area, conducted under the U.S. National Toxicology Program (NTP) is presented in Section 2.4. These data sets, which contain missing observations, will be used in illustrating methodologies in the subsequent chapters.

### 2.1 Onychomycosis Trial

The data introduced in this section were obtained from a randomized, double-blind, parallel group, multicenter study for the comparison of two oral treatments (in what follows coded as  $A$  and  $B$ ) for toenail dermatophyte onychomycosis (TDO), described in full detail by De Backer *et al* (1996). TDO is a common toenail infection, difficult to treat, affecting more than 2 out of 100 persons (Roberts 1992). Antifungal compounds, classically used for treatment of TDO, need to be taken until the whole nail has grown out healthy. The development of new compounds, however, has reduced the treatment duration to 3 months. Interest is on the comparison of the efficacy and safety of 12 weeks of continuous therapy with treatment  $A$  or with treatment  $B$ .

Table 2.1: *Toenail Data. Number and percentage of patients ( $N$ ) with severe toenail infection, for each treatment arm separately.*

|           | Group A  |     |       | Group B  |     |       |
|-----------|----------|-----|-------|----------|-----|-------|
|           | # Severe | $N$ | %     | # Severe | $N$ | %     |
| Baseline  | 54       | 146 | 37.0% | 55       | 148 | 37.2% |
| 1 month   | 49       | 141 | 34.7% | 48       | 147 | 32.6% |
| 2 months  | 44       | 138 | 31.9% | 40       | 145 | 27.6% |
| 3 months  | 29       | 132 | 22.0% | 29       | 140 | 20.7% |
| 6 months  | 14       | 130 | 10.8% | 8        | 133 | 6.0%  |
| 9 months  | 10       | 117 | 8.5%  | 8        | 127 | 6.3%  |
| 12 months | 14       | 133 | 10.5% | 6        | 131 | 4.6%  |

Table 2.2: *Toenail Data. Number of available repeated measurements per subject, for each treatment arm separately.*

| # Obs. | Group A |        | Group B |        |
|--------|---------|--------|---------|--------|
|        | $N$     | %      | $N$     | %      |
| 1      | 4       | 2.74%  | 1       | 0.68%  |
| 2      | 2       | 1.37%  | 1       | 0.68%  |
| 3      | 4       | 2.74%  | 3       | 2.03%  |
| 4      | 2       | 1.37%  | 4       | 2.70%  |
| 5      | 2       | 1.37%  | 8       | 5.41%  |
| 6      | 25      | 17.12% | 14      | 9.46%  |
| 7      | 107     | 73.29% | 117     | 79.05% |
| Total: | 146     | 100%   | 148     | 100%   |

In total,  $2 \times 189$  patients were randomized, distributed over 36 centers. Subjects were followed during 12 weeks (3 months) of treatment and followed further, up to a total of 48 weeks (12 months). Measurements were taken at baseline, every month during treatment, and every 3 months afterwards, resulting in a maximum of 7 measurements per subject. At the first occasion, the treating physician indicates one of the affected toenails as the target nail, the nail which will be followed over time. Subsequent analyses will be restricted to only those patients for which the target nail was one of the two big toenails. This reduces the sample under consideration to 146 and 148 subjects, in group  $A$  and group  $B$ , respectively.

One of the responses of interest was the unaffected nail length, measured from the nail bed to the infected part of the nail, which is always at the free end of the nail, expressed in  $mm$ . This outcome has been studied extensively in Verbeke and

Molenberghs (2000). Another important outcome in this data set was the severity of the infection, coded as 0 (not severe) or 1 (severe). The question of interest was whether the percentage of severe infections decreased over time, and whether that evolution was different for the two treatment groups. This outcome has been studied extensively in Molenberghs and Verbeke (2005) and Molenberghs and Kenward (2007). A summary of the number of patients in the study at each time-point, and the number of patients with severe infections is given in Table 2.1.

Due to a variety of reasons, the outcome has been measured at all 7 scheduled time points, for only 224 (76%) out of the 298 participants. Table 2.2 summarizes the number of available repeated measurements per subject, for both treatment groups separately. We see that the occurrence of missingness is similar in both treatment groups.

The data are subjected to analyses in Chapter 4 and Chapter 5.

## 2.2 Analgesic Trial

These data, studied extensively in Molenberghs and Verbeke (2005) and Molenberghs and Kenward (2007), come from a single-arm clinical trial in 395 patients who are given analgesic treatment for pain caused by chronic nonmalignant disease. Treatment was to be administered for 12 months and assessed by means of a ‘Global Satisfaction Assessment’ (GSA) scale, rated on a five-point scale:

$$\text{GSA} = \begin{cases} 1 : \text{very good,} \\ 2 : \text{good,} \\ 3 : \text{indifferent,} \\ 4 : \text{bad,} \\ 5 : \text{very bad.} \end{cases} \quad (2.1)$$

Some analyses have been done on a dichotomized version:

$$\text{GSABIN} = \begin{cases} 1 : \text{if GSA} \leq 3 \text{ ('Very Good' to 'Moderate')}, \\ 0 : \text{otherwise.} \end{cases} \quad (2.2)$$

Apart from the outcome of interest, a number of covariates are available, such as age, sex, weight, duration of pain in years prior to the start of the study, type of pain, physical functioning, psychiatric condition, respiratory problems, etc.

GSA was rated by each person four times during the trial, at months 3, 6, 9, and 12. An overview of the frequencies per follow up time is given in Table 2.3. Inspecting Table 2.3 reveals that the total per column is variable. This is due to missingness.

At three months, 10 subjects lack a measure, with these numbers being 93, 168, and 172 at subsequent times. Not only monotone missingness or dropout occurs, there are also subjects with intermittent values.

Table 2.3: *Analgesic Trial. Absolute and relative frequencies of the five GSA categories for each of the four follow up times.*

| GSA | Month 3 |       | Month 6 |       | Month 9 |       | Month 12 |       |
|-----|---------|-------|---------|-------|---------|-------|----------|-------|
| 1   | 55      | 14.3% | 38      | 12.6% | 40      | 17.6% | 30       | 13.5% |
| 2   | 112     | 29.1% | 84      | 27.8% | 67      | 29.5% | 66       | 29.6% |
| 3   | 151     | 39.2% | 115     | 38.1% | 76      | 33.5% | 97       | 43.5% |
| 4   | 52      | 13.5% | 51      | 16.9% | 33      | 14.5% | 27       | 12.1% |
| 5   | 15      | 3.9%  | 14      | 4.6%  | 11      | 4.9%  | 3        | 1.4%  |
| Tot | 385     |       | 302     |       | 227     |       | 223      |       |

An overview of the extent of missingness is shown in Table 2.4. Note that only around 40% of the subjects have complete data. The dropout sequences amount to roughly another 40%, with close to 20% of the patterns showing intermittent missingness. This example underscores that a satisfactory longitudinal analysis will oftentimes have to address the missing data problem.

Table 2.4: *Analgesic Trial. Overview of missingness patterns and frequencies with which they occur. ‘O’ indicates observed and ‘M’ indicates missing.*

| Measurement occasion     |         |         |          | Number | %     |
|--------------------------|---------|---------|----------|--------|-------|
| Month 3                  | Month 6 | Month 9 | Month 12 |        |       |
| Completers               |         |         |          |        |       |
| O                        | O       | O       | O        | 163    | 41.2  |
| Dropouts                 |         |         |          |        |       |
| O                        | O       | O       | M        | 51     | 12.91 |
| O                        | O       | M       | M        | 51     | 12.91 |
| O                        | M       | M       | M        | 63     | 15.95 |
| Non-monotone missingness |         |         |          |        |       |
| O                        | O       | M       | O        | 30     | 7.59  |
| O                        | M       | O       | O        | 7      | 1.77  |
| O                        | M       | O       | M        | 2      | 0.51  |
| O                        | M       | M       | O        | 18     | 4.56  |
| M                        | O       | O       | O        | 2      | 0.51  |
| M                        | O       | O       | M        | 1      | 0.25  |
| M                        | O       | M       | O        | 1      | 0.25  |
| M                        | O       | M       | M        | 3      | 0.76  |

## 2.3 Depression Trial

These data come from a clinical trial in major depressive disorder originally reported by Diggle *et al.* (2002). The primary objective in that trial was to compare the efficacy of an experimental antidepressant with placebo to support a new drug application. As such, this was a phase III (confirmatory) trial. Patients were randomly assigned (1:1 ratio) to placebo (n=139) or the experimental drug (n=128), with the double-blind treatment period lasting 9 weeks. Study visits were scheduled once a week for the first 3 weeks after randomization, and every two weeks thereafter. The Hamilton Depression Rating Scale (HAMD<sub>17</sub>) was used to measure the depression status of the patients. The completion rates and dropout rates due to adverse events and lack of efficacy are presented in Table 2.5. Adverse events and lack of efficacy are the main reasons for dropout in the experimental and placebo treatment arm respectively.

Table 2.5: *Depression Trials. Completion Rates and Main Reasons for Study Discontinuation*

| Reason for discontinuation | Experimental Drug | Placebo |
|----------------------------|-------------------|---------|
| Protocol completion rate   | 60.9%             | 64.7%   |
| Adverse Events             | 12.5%             | 4.3%    |
| Lack of Efficacy           | 5.5%              | 13.7%   |

The experimental drug was found to be significantly superior to placebo on the a priori declared primary efficacy analysis (direct likelihood-based repeated measures) of mean change to endpoint on the HAMD<sub>17</sub> total score. Therefore, even though results were significant based on the primary analysis (Diggle *et al.*, 2002), it is reasonable to wonder how effective the medication would be in actual practice given the rates of dropout in the trial, and to what degree missing data might have biased the estimate of treatment efficacy. The data are used as a motivating example for the study to be discussed in Chapter 8.

## 2.4 National Toxicology Program Data

This developmental toxicity study investigates the dose-response relationship in mice of the potentially hazardous chemical compound di(2-ethylhexyl)phthalate (DEHP), used in vacuum pumps (Windholz, 1983) and as plasticizers for numerous plastic devices made of polyvinyl chloride. The developmental toxicity study, conducted in timed-pregnant mice during the period of major organogenesis and described by Tyl *et al.* (1988), has attracted much interest in the toxicity of DEHP. The doses selected for the study were 0, 0.025, 0.05, 0.1, and 0.15%, corresponding to a DEHP consumption of 0, 44, 91, 191, and 292 mg/kg/day, respectively. The dams were sacrificed, slightly prior to normal delivery, and the status of uterine implantation sites recorded. A total of 1082 live fetuses were dissected from the uterus, anesthetized, and examined for ‘malformation,’ defined as 0 if none of the three malformations occurs, and 1 otherwise.

Evidently, fetuses are clustered within mothers; hence the implied association needs to be accommodated in the analysis. Table 2.6 summarizes the data. Detailed summary of the data and analyses can be found in Aerts *et al.* (2002) and Molenberghs and Verbeke (2005). Our focus will be on the binary malformation. The data are analyzed in Chapter 5.



Table 2.6: *NTP data: Developmental Toxicity Study (DEHP). Summary data by dose group.*

| Dose | # Dams, $\geq 1$ |       | Live | Litter      |               |       |      |
|------|------------------|-------|------|-------------|---------------|-------|------|
|      | Impl.            | Viab. |      | Size (mean) | Malformations |       |      |
|      |                  |       |      | Ext.        | Visc.         | Skel. |      |
| 0    | 30               | 30    | 330  | 13.2        | 0.0           | 1.5   | 1.2  |
| 44   | 26               | 26    | 288  | 11.1        | 1.0           | 0.4   | 0.4  |
| 91   | 26               | 26    | 277  | 10.7        | 5.4           | 7.2   | 4.3  |
| 191  | 24               | 17    | 137  | 8.1         | 17.5          | 15.3  | 18.3 |
| 292  | 25               | 9     | 50   | 5.6         | 54.0          | 50.0  | 48.0 |



## Chapter 3

# Fundamental Concepts of Incomplete Data

This chapter reviews basic terminology, fundamental concepts and existing methodologies that are used in the area of incomplete (longitudinal) data analysis and that will be used throughout the thesis.

### 3.1 Incomplete Data

The following terminology is based on the standard framework of Rubin (1976) and Little and Rubin (2002). Let the random variable  $Y_{ij}$  denote the response for the  $i^{\text{th}}$  study subject at the  $j^{\text{th}}$  occasion ( $i = 1, \dots, N$ ,  $j = 1, \dots, n_i$ ). Independence across subjects is assumed. In clustered-data setting, such as the NTP data (Section 2.4),  $Y_{ij}$  indicates the  $j^{\text{th}}$  individual in the  $i^{\text{th}}$  cluster. We group the outcomes into a vector  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})'$  and define a further vector of missingness indicators  $\mathbf{R}_i = (R_{i1}, \dots, R_{in_i})'$  with  $R_{ij} = 1$  if  $Y_{ij}$  is observed and 0 otherwise. The set of measurements, along with the missingness indicators,  $(\mathbf{Y}_i, \mathbf{R}_i)$ , comprise what is called the *full* data. Typically, the vector  $\mathbf{Y}_i$  is divided into *observed* ( $\mathbf{Y}_i^o$ ) and *missing* ( $\mathbf{Y}_i^m$ ) components, respectively. For incomplete data, only  $(\mathbf{Y}_i^o, \mathbf{R}_i)$  is available.

The structure of the missingness vector admits two basic types of missingness: *monotone* and *non-monotone*. When the missingness is *monotone* or of a *dropout* nature, the unobserved measurement within the longitudinal series all occur after a particular measurement occasion, and in that sense, the subject is said to have “dropped out” of the study. In such cases, the missingness indicator  $\mathbf{R}_i$  consists of a very particular form, with all  $R_{ij}$  equal to one up to a particular time point

$j$  and zero thereafter. This structure allows the missingness indicators in  $\mathbf{R}_i$  to be collapsed into a single variate,  $D_i$ , defined as  $D_i = 1 + \sum_{j=1}^{n_i} R_{ij}$  denoting the time point at which subject  $i$  drops out. *Non-monotone* missingness on the other hand, occurs when missing values arise intermittently within the series, leading to no distinct configuration of the missingness indicators, and thus, simplifications of  $\mathbf{R}_i$  into a lower-dimensional form is not straightforward.

## 3.2 Modeling Frameworks

In principle, one would need to consider the density of the full data  $f(\mathbf{y}_i, \mathbf{r}_i | \boldsymbol{\theta}, \boldsymbol{\psi})$ , where the parameter vectors  $\boldsymbol{\theta}$  and  $\boldsymbol{\psi}$  describe the measurement and missingness processes, respectively. When appropriate,  $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')$  will be used to split into the mean regression parameters  $\boldsymbol{\beta}$  and the association parameters  $\boldsymbol{\alpha}$ . Covariates are assumed to be measured and grouped in a design matrix  $\mathbf{X}_i$ , although, for notational simplicity, this is sometimes dropped from the notation in later sections. This full density function can be factored in different ways, each leading to a different framework. Under a *selection model* framework (Rubin, 1976; Little and Rubin, 2002), the joint distribution is factored into a marginal density of the measurement process and a conditional model for missingness process given the outcomes, that is,

$$f(\mathbf{y}_i, \mathbf{r}_i | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}) f(\mathbf{r}_i | \mathbf{x}_i, \mathbf{y}_i, \boldsymbol{\psi}). \quad (3.1)$$

Selection models are an obvious choice for clinicians, for instance, who are often interested in the marginal effect,  $\boldsymbol{\theta}$ , of the independent variables (e.g., treatment) on the response.

Alternatively, one can consider so-called *pattern-mixture models* (Little, 1993, 1994a; Molenberghs *et al.*, 1997), using the reversed factorization:

$$f(\mathbf{y}_i, \mathbf{r}_i | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i | \mathbf{r}_i, \mathbf{x}_i, \boldsymbol{\theta}) f(\mathbf{r}_i | \mathbf{x}_i, \boldsymbol{\psi}). \quad (3.2)$$

This density can be seen as a mixture of different populations, each of which is defined conditionally on the observed pattern of missingness. The parameters  $\boldsymbol{\theta}$  then denote pattern-specific effects of the independent variables on the response.

Instead of using the selection or pattern-mixture model frameworks, the measurement and the dropout process can be jointly modeled using a *shared-parameter model* (Wu and Carroll, 1988; Wu and Bailey, 1989). In such a model the measurement and dropout process are assumed to be independent, conditional upon a certain set of shared parameters. This shared-parameter model is formulated by way of the

following factorization:

$$f(\mathbf{y}_i, \mathbf{r}_i | \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{x}_i, \boldsymbol{\theta}) f(\mathbf{r}_i | \mathbf{b}_i, \mathbf{x}_i, \boldsymbol{\psi}). \quad (3.3)$$

Here,  $\mathbf{b}_i$  are the shared parameters, often considered to be random effects and following a specific parametric distribution.  $\boldsymbol{\theta}$  denotes the effects of the independent variables, conditional on the random effects.

### 3.3 Missing Data Mechanisms

To obtain valid inferences from incomplete (longitudinal) data, we must consider the nature of the “missing data mechanism”. Ordinarily, the missing data mechanism is not under the control of the investigators; consequently, it is often not well understood. Instead, assumptions are made about the missing data mechanism, and the validity of the analyses will depend on whether these assumptions hold for the data at hand.

The general missing data taxonomy described in this section is fully presented in Rubin (1976) and Little and Rubin (2002). Within the selection model framework, Rubin (1976) developed a taxonomy to classify the missingness process based on its dependence (or lack thereof) on the measurement process. This classification is based on the structure of the second term in the right hand side of (3.1), which upon partitioning of the response vector into its observed and missing components, can be expressed as  $f(\mathbf{r}_i | \mathbf{x}_i, \mathbf{y}_i, \boldsymbol{\psi}) = f(\mathbf{r}_i | \mathbf{x}_i, \mathbf{y}_i^o, \mathbf{y}_i^m, \boldsymbol{\psi})$ . When there is independence of the measurement and missingness process, conditionally of the covariates, the mechanism is be *missing completely at random* (MCAR). A less rigid assumption would be one of *missing at random* (MAR), for which the missingness may depend on the observed outcomes and covariates but, given these, no further on the unobserved outcomes. If the cause of missing data is neither MCAR nor MAR, the data is *missing not at random* (MNAR).

To help motivate the different missing data mechanisms, consider a longitudinal clinical trial to assess the efficacy of a new treatment for a particular disease or condition. When a patient does not show up for a visit because his car broke down while driving to the hospital or couldn’t go to the hospital because of a terrible weather, this is most probably fall within the category of MCAR, since the missingness process and the outcome are independent. Alternatively, if a patient missed a visit because in previous visits her/his condition stabilized and s/he is convinced that continuing the visits to the hospital are of no value, the nature of the missingness is related to the previously observed (improving) outcomes, and the process most plausibly is MAR. In general, if dropping out is known to be unrelated to changes in

health status, an MAR assumption for the missing values seems justified; however, if dropping out is related to health status (e.g., a move to live with and be cared for by a parent or offspring), then the MAR assumption is not justified, and the missing data are likely MNAR.

Throughout this thesis, we basically focus on longitudinal data with monotone missingness on the one hand and on incomplete clustered data on the other, each time under MAR.

### 3.4 Model Families

In this section we discuss methods to model (longitudinal) data both in the Gaussian and non-Gaussian setting. For Gaussian longitudinal data the linear mixed model, undoubtedly the most commonly used, is considered. Then, we focus on the situation of non-Gaussian outcomes, for which we distinguish between three model families: marginal, random-effects and conditional models.

#### 3.4.1 A Classic Model for Continuous Data

The most widely used methodology for continuous longitudinal data within the likelihood framework is the general *linear mixed-effects model* (Laird and Ware, 1982), which takes the form

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \quad (3.4)$$

where  $\mathbf{Y}_i$  is the  $n_i$ -dimensional (longitudinal) response vector for subject  $i$ ,  $\mathbf{X}_i$  and  $\mathbf{Z}_i$  are, respectively, the  $(n_i \times p)$  and  $(n_i \times q)$  known design matrices,  $\boldsymbol{\beta}$  is the  $p$ -dimensional vector containing the fixed effects,  $\mathbf{b}_i \sim N(\mathbf{0}, D)$  is the  $q$ -dimensional vector containing random effects, and  $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \Sigma_i)$  is an  $n_i$ -dimensional vector of residual components, combining measurement error and serial correlation. Further,  $\mathbf{b}_1, \dots, \mathbf{b}_N, \boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_N$  are assumed to be independent.  $D$  and  $\Sigma_i$  are general covariance matrices of size  $(q \times q)$  and  $(n_i \times n_i)$ , respectively. In the case of no serial correlation,  $\Sigma_i$  reduces to  $\sigma^2 \mathbf{I}_{n_i}$ .

It follows from (3.4) that, conditional on  $\mathbf{b}_i$ ,  $\mathbf{Y}_i$  is normally distributed with mean vector  $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$  and with covariance matrix  $\Sigma_i$ . Upon integration over the random effects,  $\mathbf{b}_i \sim N(\mathbf{0}, D)$ , the resulting marginal (i.e., averaged over the random effects) model for the response can be expressed as:

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i D \mathbf{Z}_i' + \Sigma_i). \quad (3.5)$$

In line with Diggle *et al.* (2002), Aerts *et al.* (2002) and Molenberghs and Verbeke (2005), we distinguish between three model families. Next, an example of each for the case of Gaussian outcomes, or more generally for models with a linear mean structure will be given.

A *marginal model*, often referred as population-averaged model, is characterized by a marginal mean function of the form

$$E(Y_{ij}|\mathbf{x}_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}, \quad (3.6)$$

where  $\mathbf{x}_{ij}$  is a vector of covariates for subject  $i$  at time point  $j$  and  $\boldsymbol{\beta}$  is a vector of regression parameters.

*Random-effects models* or *cluster-specific models*, on the other hand, further condition on a vector of random effects  $\mathbf{b}_i$ , over and above conditioning on covariates, that is,

$$E(Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i. \quad (3.7)$$

Finally, a third family of models, *conditional model*, describes the distribution of the components of the outcome vector, conditional on the predictor variables but also conditional on (a subset of) the other components of the response vector. A simple first-order stationary transition model focuses on expectations of the form:

$$E(Y_{ij}|Y_{i,j-1}, \dots, Y_{i1}, \mathbf{x}_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \alpha Y_{i,j-1}. \quad (3.8)$$

Alternatively, one might condition upon all outcomes except the one being modeled.

As shown by Verbeke and Molenberghs (2000), random-effects model imply a simple marginal model in the linear mixed model case. This is due to the elegant properties of the multivariate normal distribution. In particular, expectation (3.6) follows from (3.7) either by (a) marginalizing over the random effects or by (b) conditioning on the random-effects vector  $\mathbf{b}_i = \mathbf{0}$ . For instance, the linear mixed model (3.4), pointed out in Section 3.4.1, which belongs to the random-effects model family, implies marginal model (3.5). Thus, the fixed-effects parameters  $\boldsymbol{\beta}$  have both a marginal and hierarchical model interpretation. Certain auto-regressive models, in which later-time residuals are expressed in terms of earlier ones, can also lead to particular instances for which the general linear mixed-effects model implies some marginal function of the form (3.6).

### 3.5 Models for non-Gaussian data

Whereas the linear mixed model is seen as a unifying parametric model for Gaussian repeated measures (Verbeke and Molenberghs (2000)), there are a variety of methods

in common use in the non-Gaussian setting.

### 3.5.1 Generalized linear mixed models

When the response of interest is discrete, linear mixed-effects models are not appropriate for at least two main reasons. First, with a discrete response there is intrinsic dependence of the variability on the mean. Second, the range of the mean response (e.g., a proportion or rate for a response that is binary or a count, respectively) is constrained (Fitzmaurice *et al.*, 2009). Instead extension of the linear mixed-effects model to non-Gaussian (e.g., binary) longitudinal responses classified as *generalized linear mixed models* (Molenberghs and Verbeke, 2005) are used.

*Generalized linear mixed models* (GLMMs) expand the generalized linear model framework (McCullagh and Nelder, 1989) to the case of correlated responses by (1) including subject-specific regression parameters  $\mathbf{b}_i$  in the linear predictor to address correlations among the repeated measures, and (2) assuming that conditional on the random effects  $\mathbf{b}_i$ , the elements of  $\mathbf{Y}_i$  are independent. A typical GLMM assumes that all  $Y_{ij}$  have densities of the form  $f_i(y_{ij})$  in the exponential family, and the mean  $\mu_{ij}$  is modeled through a linear predictor containing fixed regression parameters  $\boldsymbol{\beta}$ , as well as subject-specific parameters  $\mathbf{b}_i$ , with some known link function  $\eta(\cdot)$ . It is further assumed that the random effects follow a normal distribution.

As indicated earlier, models for correlated repeated measures present in any longitudinal (or multivariate) can be grouped in to three broad, but quite distinct, families. These models differ not only in how the correlation among the repeated outcomes is accounted for, but also have different parameter interpretations.

*Marginal models* evaluate the overall risk as a function of covariates only. The correlation among the components of the outcome vector can be captured either by adopting a full parametric approach or by means of working assumptions, such as in GEE (Liang and Zeger, 1986). Typical marginal models for binary data, include the Bahadur (1961) model and the multivariate Dale or global odds ratio model (Molenberghs and Lesaffre, 1994, 1999). For an overview, see Molenberghs and Verbeke (2005). Such models, however, can involve complex and hard-to-manipulate likelihoods; they can be prohibitive in some settings. Moreover, various authors have voiced concern over these models' vulnerability to misspecification. As a consequence, generalized estimating equation and pseudo-likelihood have been used as alternative methods. Pseudo-likelihood can be used with both marginal and conditional models.

In *random-effects models* or *cluster-specific models*, the response is modeled as a function of covariates and parameters, specific to the the subject. In such mod-



els, interpretation of fixed-effects parameters is conditional on a constant level of the random-effects parameter. Under such models, correlations among the repeated measures are addressed by the inclusion of parameters that are specific to a subject, so that given the collection of these subject-specific parameters or random effects, the responses within the longitudinal series are assumed to be independent.

In a *conditional model*, the parameter describe a feature (expectation, odds, logits, ...) of a (set of) responses, given values for the other responses. The best known example is the log-linear model. Correlations within the longitudinal responses can be dealt with by considering a particular outcome in the series and modeling it conditionally on the other outcomes (or subsets thereof).

### Marginal versus Subject-Specific Models

Unlike for the Gaussian case, the parameters of the subject-specific and marginal models for correlated binary data describe different types of effects of the covariates on the response probabilities. The consequence of this disparity between random-effects and marginally specified models in the non-Gaussian setting is an obvious distinction between the parameters under each.

We illustrate this in line with Molenberghs and Verbeke (2005) by considering a binary response. Assume a random-intercept logistic model with linear predictor logit  $P(Y_{ij} = 1|t_{ij}, b_i) = \beta_0 + \beta_1 t_{ij} + b_i$  where  $t_{ij}$  represent the time covariate. The conditional mean response are given by:

$$E(Y_{ij}|t_{ij}, b_i) = P(Y_{ij} = 1|t_{ij}, b_i) = \frac{\exp(\beta_0 + \beta_1 t_{ij} + b_i)}{1 + \exp(\beta_0 + \beta_1 t_{ij} + b_i)}, \quad (3.9)$$

whereas the marginal average evolution,  $E(Y_{ij}|t_{ij}) \equiv E(Y_{ij})$ , is obtained from averaging the random effects (3.9):

$$E(Y_{ij}) = E[E(Y_{ij}|b_i)] = E\left[\frac{\exp(\beta_0 + \beta_1 t_{ij} + b_i)}{1 + \exp(\beta_0 + \beta_1 t_{ij} + b_i)}\right] \neq \frac{\exp(\beta_0 + \beta_1 t_{ij})}{1 + \exp(\beta_0 + \beta_1 t_{ij})}. \quad (3.10)$$

This demonstrates that the implied marginal model from a random effects specification does not necessarily reduce to the marginally specified model. The inherent differences across the model families, particularly for non-Gaussian responses, warrants careful consideration regarding the type of model to be employed.

### 3.5.2 Generalized Estimating Equations

Liang and Zeger (1986) proposed so-called *Generalized Estimating Equations* (GEE), useful to circumvent the computational complexity of the full likelihood, can be con-

sidered whenever scientific interest is restricted to the mean parameters. GEE requires only the correct specification of the univariate marginal distributions, provided one is willing to adopt so-called *working assumptions* about the association structure of the vector of repeated measurements. When inferences focus on population averages, one can directly model all of the marginal expectations  $E(Y_{ij}) = \mu_{ij}$  in terms of covariates of interest. This is typically done via  $h(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}$ , with  $h(\cdot)$  some known link function, such as the logit link for binary responses.

The marginal variance depends on the marginal mean according to  $\text{Var}(Y_{ij}) = v(\mu_{ij})\phi$ , where  $v(\cdot)$  is a known variance function and  $\phi$  is a scale (overdispersion) parameter. The correlation between  $Y_{ij}$  and  $Y_{ik}$  is expressed via a correlation matrix  $R_i(\boldsymbol{\alpha})$  where  $\boldsymbol{\alpha}$  is a vector of nuisance parameters. The covariance matrix  $V_i$  of  $\mathbf{Y}_i$  can then be written as  $V_i = V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \phi A_i^{1/2} R_i A_i^{1/2}$ , with  $A_i$  the matrix with the marginal variances on the main diagonal and zeros elsewhere.

Generalized estimating equations take the form

$$U(\boldsymbol{\beta}) = \sum_{i=1}^N \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} V_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \quad (3.11)$$

The nuisance parameter  $\boldsymbol{\alpha}$  needs to be replaced by a consistent estimate. Liang and Zeger (1986) proposed a moment-based estimator to this effect.

Assuming that the marginal mean  $\boldsymbol{\mu}_i$  has been correctly modeled, it can be shown that, under mild regularity conditions, the estimator  $\hat{\boldsymbol{\beta}}$  obtained from solving (3.11) is asymptotically normally distributed with mean  $\boldsymbol{\beta}$  and with covariance matrix

$$\text{var}(\hat{\boldsymbol{\beta}}) = I_0^{-1} I_1 I_0^{-1}, \quad (3.12)$$

where

$$I_0 = \sum_{i=1}^N \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}, \quad I_1 = \sum_{i=1}^N \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \text{Var}(\mathbf{y}_i) V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}. \quad (3.13)$$

In practice,  $\text{Var}(\mathbf{y}_i)$  in (3.13) is replaced by  $(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)'$ , which is unbiased on the sole condition, again, that the mean was correctly specified. We will refer to  $I_0^{-1}$  as the model based variance estimator (which should not be used as it overestimates the precision), to  $I_1$  as the empirical correction, and to  $I_0^{-1} I_1 I_0^{-1}$  as the empirically corrected variance estimator (sandwich estimator).

### 3.5.3 Pseudo-likelihood Estimation

Maximum likelihood estimation could be unattractive due to extensive computational requirements especially for non-Gaussian outcomes and when measurement sequences

are of moderate to large length. As discussed in the previous section, this is one of the reasons why *generalized estimating equations* (GEE) have become popular. One way to view the genesis of GEE is by modifying the score equations to simpler estimating equations, thereby preserving consistency and asymptotic normality, upon using an appropriately corrected variance-covariance matrix. Alternatively, the (log-)likelihood itself can be simplified to a more manageable form. This is, broadly speaking, the idea behind *pseudo-likelihood* (PL). Its principal idea is to replace a numerically challenging joint density (and hence likelihood) by a simpler function assembled from suitable factors. As a simple illustration, a three-way density

$$L_i = f(y_{i1}, y_{i2}, y_{i3} | \boldsymbol{\theta}_i) \quad (3.14)$$

would be replaced by the product

$$L_i^* = f(y_{i1}, y_{i2} | \boldsymbol{\theta}_i^*) \cdot f(y_{i1}, y_{i3} | \boldsymbol{\theta}_i^*) \cdot f(y_{i2}, y_{i3} | \boldsymbol{\theta}_i^*) \quad (3.15)$$

Such a change is computationally advantageous, asymptotics can be rescued, and modeling (3.15) is equally simple, if not simpler, than modeling (3.14), as the parameter vector  $\boldsymbol{\theta}_i^*$  in (3.15) typically is a sub-vector of  $\boldsymbol{\theta}_i$  in (3.14) (Molenberghs and Verbeke, 2005).

While the method achieves important computational economies by changing the method of estimation, it fortunately does not affect model interpretation. Model parameters can be chosen in the same way as with full likelihood and retain their meaning. Because the above product does not lead to a likelihood function, appropriate modifications will be needed to guarantee correct inferences. Pseudo-likelihood (PL) for incomplete data is not a full likelihood method and hence in its basic form valid only under MCAR. To introduce pseudo-likelihood formally, we will use the convenient general definition given by Arnold and Strauss (1991).

**Definition and Asymptotic Properties:** Define  $S$  as the set of all  $2^n - 1$  vectors of length  $n$ , consisting solely of zeros and ones, with each vector having at least one non-zero entry. Denote by  $\mathbf{y}_i^{(s)}$  the subvector of  $\mathbf{y}_i$  corresponding to the components of  $s$  that are non-zero. The associated joint density is  $f_s(\mathbf{y}_i^{(s)}; \boldsymbol{\theta}_i)$ . To define a pseudo-likelihood function, one chooses a set  $\delta = \{\delta_s | s \in S\}$  of real numbers, with at least one non-zero component. The log of the pseudo-likelihood is then defined as

$$p\ell = \sum_{i=1}^N \sum_{s \in S} \delta_s \ln f_s(\mathbf{y}_i^{(s)}; \boldsymbol{\theta}_i). \quad (3.16)$$

Adequate regularity conditions have to be assumed to ensure that (3.16) can be maximized by solving the pseudo-likelihood (score) equations, the latter obtained by differentiating the logarithmic pseudo-likelihood and by equating its derivative to zero. The classical log-likelihood function is found by setting  $\delta_s = 1$  if  $s$  is the vector consisting solely of ones, and 0 otherwise. The required regularity conditions (Arnold and Strauss, 1991; Geys, Molenberghs, and Ryan, 1999; Aerts *et al.*, 2002) on the density functions  $f_s(\mathbf{y}^{(s)}; \boldsymbol{\theta})$  are:

- A0** The densities  $f_s(\mathbf{y}^{(s)}; \boldsymbol{\theta})$  are distinct for different values of the parameter  $\boldsymbol{\theta}$ .
- A1** The densities  $f_s(\mathbf{y}^{(s)}; \boldsymbol{\theta})$  have common support, which does not depend on  $\boldsymbol{\theta}$ .
- A2** The parameter space  $\Omega$  contains an open region  $\omega$  of which the true parameter value  $\boldsymbol{\theta}_0$  is an interior point.
- A3**  $\omega$  is such that for all  $s$ , and almost all  $\mathbf{y}^{(s)}$  in the support of  $\mathbf{Y}^{(s)}$ , the densities admit all third derivatives

$$\frac{\partial^3 f_s(\mathbf{y}^{(s)}; \boldsymbol{\theta})}{\partial \theta_j \partial \theta_k \partial \theta_\ell}.$$

- A4** The first and second logarithmic derivatives of  $f_s$  satisfy

$$E_{\boldsymbol{\theta}} \left( \frac{\partial \ln f_s(\mathbf{y}^{(s)}; \boldsymbol{\theta})}{\partial \theta_k} \right) = 0, \quad k = 1, \dots, q,$$

and

$$0 < E_{\boldsymbol{\theta}} \left( \frac{-\partial^2 \ln f_s(\mathbf{y}^{(s)}; \boldsymbol{\theta})}{\partial \theta_k \partial \theta_\ell} \right) < \infty, \quad k, \ell = 1, \dots, q.$$

- A5** The matrix  $I_0$ , defined in (3.18), is positive definite.

- A6** There exist functions  $M_{klr}$  such that

$$\sum_{s \in S} \delta_s E_{\boldsymbol{\theta}} \left| \frac{\partial^3 \ln f_s(\mathbf{y}^{(s)}; \boldsymbol{\theta})}{\partial \theta_k \partial \theta_\ell \partial \theta_r} \right| < M_{klr}(\mathbf{y})$$

for all  $\mathbf{y}$  in the support of  $f$  and for all  $\boldsymbol{\theta} \in \omega$  and  $m_{klr} = E_{\boldsymbol{\theta}_0}(M_{klr}(Y)) < \infty$ .

Let  $\boldsymbol{\theta}_0$  be the true parameter. Under the suitable regularity conditions, it can be shown that maximizing the function (3.16) produces a consistent and asymptotically normal estimator  $\tilde{\boldsymbol{\theta}}_0$  so that  $\sqrt{N}(\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$  converges in distribution to

$$N_p[\mathbf{0}, I_0(\boldsymbol{\theta}_0)^{-1} I_1(\boldsymbol{\theta}_0) I_0(\boldsymbol{\theta}_0)^{-1}]. \quad (3.17)$$

Precise statements are as shown in Theorem 1. Theorem 1, proven by Arnold and Strauss (1991), guarantees the existence of at least one solution to the pseudo-likelihood equations, which is consistent and asymptotically normal. Without loss of generality, we can assume  $\boldsymbol{\theta}$  is constant. Replacing it by  $\boldsymbol{\theta}_i$ , and modeling it as a function of covariates is straightforward.

**Theorem 1 (Consistency and Asymptotic Normality)** *Assume that  $(\mathbf{Y}_1, \dots, \mathbf{Y}_N)$  are i.i.d. with common density that depends on  $\boldsymbol{\theta}_0$ . Then under regularity conditions (A1)–(A6):*

1. *the pseudo-likelihood estimator  $\tilde{\boldsymbol{\theta}}_N$ , defined as the maximizer of (3.16), converges in probability to  $\boldsymbol{\theta}_0$ .*
2.  *$\sqrt{N}(\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$  converges in distribution to  $N_p(\mathbf{0}, I_0(\boldsymbol{\theta}_0)^{-1}I_1(\boldsymbol{\theta}_0)I_0(\boldsymbol{\theta}_0)^{-1})$  with  $I_0(\boldsymbol{\theta})$  defined by*

$$I_{0,k\ell}(\boldsymbol{\theta}) = - \sum_{s \in S} \delta_s E_{\boldsymbol{\theta}} \left( \frac{\partial^2 \ln f_s(\mathbf{y}^{(s)}; \boldsymbol{\theta})}{\partial \theta_k \partial \theta_\ell} \right) \quad (3.18)$$

and  $I_1(\boldsymbol{\theta})$  by

$$I_{1,k\ell}(\boldsymbol{\theta}) = \sum_{s,t \in S} \delta_s \delta_t E_{\boldsymbol{\theta}} \left( \frac{\partial \ln f_s(\mathbf{y}^{(s)}; \boldsymbol{\theta})}{\partial \theta_k} \frac{\partial \ln f_t(\mathbf{y}^{(t)}; \boldsymbol{\theta})}{\partial \theta_\ell} \right). \quad (3.19)$$

Similar to GEE (Section 3.5.2), this result provides an easy way to consistently estimate the asymptotic covariance. The matrix  $I_0$  arises from evaluating the second derivative of  $p\ell$  in (3.16) at the PL estimate. The expectation in  $I_1$  can be replaced by the cross-products of the observed scores.

As discussed by Arnold and Strauss (1991), and exactly the same as with GEE, the Cramèr-Rao inequality implies that  $I_0^{-1}I_1I_0^{-1}$  is greater than the inverse of  $I$  (the Fisher information matrix for the maximum likelihood case), in the sense that  $I_0^{-1}I_1I_0^{-1} - I^{-1}$  is positive semi-definite. Strict inequality holds if the PL estimator fails to be a function of a minimal sufficient statistic. Geys, Molenberghs, and Ryan (1999) have shown that, in realistic clustered-data settings in toxicology experiments, efficiency loss is often negligible and is certainly justified in view of computational convenience and speed.

**Marginal Pseudo-likelihood** As stated earlier, marginal models for non-Gaussian data can become prohibitive when subjected to full maximum likelihood inference, especially with large within-unit replication. In such a situation both GEE and PL are viable alternatives (Molenberghs and Verbeke, 2005). Marginal PL methodology

has been proposed, among other, by le Cessie and van Houwelingen (1994) and Geys, Molenberghs, and Lipsitz (1998). le Cessie and van Houwelingen (1994) replace the true contribution of a vector of correlated binary data to the full likelihood, written as  $f(y_{i1}, \dots, y_{in_i})$ , by the product of all pairwise contributions  $f(y_{ij}, y_{ik}), 1 \leq j < k \leq n_i$ , to obtain a pseudo-likelihood function. Also the term composite likelihood is encountered in this context. Renard, Molenberghs, and Geys (2004) refer to this particular instance of pseudo-likelihood as *pairwise likelihood*. The contribution of the  $i^{\text{th}}$  subject or cluster to the log pseudo-likelihood then specializes to

$$p\ell_i = \sum_{j < k} \ln f(y_{ij}, y_{ik}), \quad (3.20)$$

if it contains more than one observation. Otherwise,  $p\ell_i = f(y_{i1})$ . Extension to three-way and higher-order pseudo-likelihood is straightforward. All of these are special cases of (3.16).

Marginal models should be chosen whenever the scientific interest is on population level, e.g., the time evolution of a response in a study. They are also useful when there is interest in quantification of strength of association between occasions or clusters.

**Conditional Pseudo-likelihood** The example given in (3.15) is of a marginal form. Here conditional forms will be discussed. Some models lend themselves more easily to conditioning than to marginalization, such as log-linear models (Molenberghs and Verbeke, 2005, Ch. 12). Upon noting that

$$f(y_{ij}|y_{ik}, k \neq j) = \frac{f(y_{i1}, \dots, y_{in_i})}{f(y_{i1}, \dots, y_{i,j-1}, y_{i,j+1}, \dots, y_{in_i})} = \frac{f_1(\mathbf{y}_i^{(1)})}{f_{s_j}(\mathbf{y}_i^{(s_j)})},$$

a full conditional likelihood contribution becomes:

$$p\ell_i = n_i \cdot \ln f_1(\mathbf{y}_i^{(1)}) - \sum_{j=1}^{n_i} \ln f_{s_j}(\mathbf{y}_i^{(s_j)}).$$

For example, when a joint density contains a computationally intractable normalizing constant, one might calculate a suitable product of conditional densities which does not involve such a complicated function. A bivariate distribution  $f(y_1, y_2)$ , for instance, can be replaced by the product of both conditionals  $f(y_1|y_2)f(y_2|y_1)$ , even though this is not the correct factorization.

Evidently, alternative versions of conditional pseudo-likelihood are possible. For example, one could consider all pairs, conditioning upon the remaining  $n_i - 2$  outcomes. This setting has been considered by Geys, Molenberghs, and Ryan (1999) for

the analysis of the NTP data (Section 2.4). This particular setting, but then with attention for the missing data aspect, will be taken up in Section 5.2.3.

Next, we introduce a fully perimetrically specified marginal model, for correlated or longitudinal binary responses which will be considered in the subsequent chapters.

### 3.5.4 The Bahadur Model

Bahadur (1961) proposed a marginal model for binary outcomes, accounting for the association via marginal correlations. Define the marginal probability as  $\pi_{ij} = E(Y_{ij}) = P(Y_{ij} = 1)$  and standardized deviations as

$$\varepsilon_{ij} = \frac{Y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1 - \pi_{ij})}} \quad \text{and} \quad e_{ij} = \frac{y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1 - \pi_{ij})}}, \quad (3.21)$$

where  $y_{ij}$  is an actual value of the binary response variable  $Y_{ij}$ . Further, let  $\rho_{ij_1j_2} = E(\varepsilon_{ij_1}\varepsilon_{ij_2})$ ,  $\rho_{ij_1j_2j_3} = E(\varepsilon_{ij_1}\varepsilon_{ij_2}\varepsilon_{ij_3}), \dots$ , and  $\rho_{i12\dots J} = E(\varepsilon_{i1}\varepsilon_{i2} \dots \varepsilon_{iJ})$ . Then, the general Bahadur model can be represented by the expression

$$f(\mathbf{y}_i) = f_1(\mathbf{y}_i) \cdot c(\mathbf{y}_i), \quad (3.22)$$

where

$$f_1(\mathbf{y}_i) = \prod_{j=1}^{n_i} \nu_{ij}^{y_{ij}} (1 - \nu_{ij})^{1-y_{ij}}, \quad (3.23)$$

$$\begin{aligned} c(\mathbf{y}_i) = & 1 + \sum_{j_1 < j_2} \rho_{ij_1j_2} e_{ij_1} e_{ij_2} + \sum_{j_1 < j_2 < j_3} \rho_{ij_1j_2j_3} e_{ij_1} e_{ij_2} e_{ij_3} + \\ & \dots + \rho_{ij_1j_2 \dots j_{n_i}} e_{ij_1} e_{ij_2} \dots e_{ij_{n_i}}, \end{aligned} \quad (3.24)$$

Thus, the probability mass function  $f(\mathbf{y}_i)$  in (3.22) is the product of the independence model  $f_1(\mathbf{y}_i)$  and the correction factor  $c(\mathbf{y}_i)$ . One viewpoint is to consider the factor  $c(\mathbf{y}_i)$  as a model for overdispersion.

Besides the Bahadur model, a broad set of marginal models have been proposed by, for example, Dale (1986), Plackett (1965), Lang and Agresti (1994), Molenberghs and Lesaffre (1994), and Molenberghs and Lesaffre (1999). Even though a variety of flexible full likelihood models exist, maximum likelihood can be unattractive due to excessive computational requirements, especially when high-dimensional vectors of correlated data arise, as alluded to in the context of the Bahadur model.

## 3.6 Methodology for Incomplete Data

In this section, we review methods commonly used for handling incomplete data in longitudinal data analysis in four parts: simple methods, imputation methods,

maximum likelihood estimation methods and non-likelihood estimation methods. We also discuss the assumptions about missingness data mechanism required for each method to yield valid inferences in the longitudinal data setting.

### 3.6.1 Simple *ad hoc* Methods

Two simple, common methods to analyze incomplete data are *complete case analysis* (CC), which discards subjects with incomplete sequences, and simple imputation. *Last observation carried forward* (LOCF), for which the last observed measurement is substituted for values at later points in time that are not observed, is among the commonly used simple imputation methods. Still recently, clinical trial practice has put a strong emphasis on such methods. Claimed advantages include computational simplicity, no need for a full longitudinal model (e.g., when the research question is in terms of the last observed measurement occasion only) and, for LOCF, compatibility with the Intention-to-Treat (ITT) principle, since data on all patients randomized can be used.

The recent report, prepared by the panel on “Handling of Missing Data in Clinical Trials” convened by the National Academy of Sciences (NAS) at the request of the U.S. Food and Drug Administration (National Research Council, 2010), provide thorough details and recommendations on the prevention and treatment of missing data in clinical trials. The panel recommended avoiding the use of simple *ad hoc* methods, such as CC and LOCF, and advocated the use of more appropriate and more principled methods which are valid under the weaker MAR assumption and easy to implement in existing statistical software.

#### Complete Case Analysis

A *complete case analysis* (CCA) includes only those cases for analysis, for which all measurements (covariates and outcomes) were recorded (Verbeke and Molenberghs, 2000; Little and Rubin, 2002; Molenberghs and Verbeke, 2005). This method has an obvious advantage: simplicity, although the wide availability of more sophisticated methods of analysis minimizes the significance of this. It is also an inefficient use of information, with adverse effects on precision and power, even if the frequency of missing data for single variables low. Further, such an analysis will only be representative for patients who remain on study and have complete data. In addition, and very important, severe bias can result when the missingness mechanism is MAR. This method is valid under MCAR. From an intuitive point of view, if the completers in a longitudinal study are generally the “better” patients, CCA would lead to overly



optimistic results for the general population.

### **Last observation Carried forward**

*Last observation carried forward* (LOCF) is a common single imputation method where the most recent observation replaces any subsequent missing ones. It can be applied to both monotone and non-monotone missingness. The idea of LOCF is based on a very strong and unrealistic assumption that a subject's measure stays at the same level until the end of the trial or during the period they are unobserved in the case of intermittent missingness. In most clinical trial settings, the assumption that patients' condition would remain at the response level is questionable as study effects, placebo effects, and natural time evolution also influence outcomes. Molenberghs and Kenward (2007) showed, using hypothetical data, that, even under the unrealistically strong assumption of MCAR, while CCA produces unbiased estimates, the bias in the LOCF estimator does not vanish, and can even induce an apparent treatment effect when there is none. Under MAR, they showed that both can be biased and bias can go in either direction. The same authors further examined the nature of the resulting missing data mechanism implied by using LOCF (Kenward and Molenberghs, 2009). They determined that LOCF effects a missing data mechanism that is forced to depend on future, unobserved measurements— a sharp contradiction and incompatibility with MCAR, under which LOCF has been thought, apparently incorrectly, to be valid.

### **3.6.2 Multiple Imputation**

One approach for handling incomplete data, that is widely used in practice, is some form of imputation. The basic idea behind imputation is very simple: substitute or fill in the values that were not recorded with the imputed values. Methods that impute or fill in the missing values have the advantage that, unlike CCA, the information from the observed values in the incomplete cases is retained and once a filled-in data set has been constructed, standard methodology for complete data can be applied. However, single imputation methods, creating only a single filled-in data set, fail to acknowledge the uncertainty inherent in the imputation of the unobserved responses. Multiple imputation (MI) circumvents this difficulty.

MI was formally introduced by Rubin (1978). The key idea of the procedure is to first replace each missing value with a set of  $M$  plausible values drawn from the conditional distribution of the unobserved values, given the observed ones. This conditional distribution represents the uncertainty about the right value to impute.

In this way,  $M$  imputed data sets are generated (imputation stage), which are then analyzed using standard complete data methods (analysis stage). Finally, the results from the  $M$  analyses have to be combined into a single inference by means of the method laid out in Rubin (1978). In its basic form, multiple imputation requires the missingness mechanism to be MAR, even though versions under MNAR have been proposed (Rubin, 1978; Molenberghs *et al.*, 1997).

In line with notation already introduced, suppose the parameter vector of the distribution of  $\mathbf{Y}_i = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)$  is denoted by  $\boldsymbol{\theta}$ . MI uses the observed data  $\mathbf{Y}^o$  to estimate the conditional distribution of  $\mathbf{Y}^m$  given  $\mathbf{Y}^o$ . The missing data are sampled several times from this conditional distribution and augmented to the observed data. The resulting completed data are then used to estimate  $\boldsymbol{\theta}$ . If the distribution of  $\mathbf{Y}_i = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)$  were known, with parameter vector  $\boldsymbol{\theta}$ , then  $\mathbf{Y}_i^m$  could be imputed by drawing a value of  $\mathbf{Y}_i^m$  from the conditional distribution  $f(\mathbf{y}_i^m | \mathbf{y}_i^o, \boldsymbol{\theta})$ . The objective of the imputation phase is to sample from this true predictive distribution. However,  $\boldsymbol{\theta}$  in the imputation model is unknown, and therefore needs to be estimated from the data first, say  $\hat{\boldsymbol{\theta}}$ , after which  $f(\mathbf{y}_i^m | \mathbf{y}_i^o, \hat{\boldsymbol{\theta}})$  is used to impute the missing data. Precisely, this implies that one first generates draws from the distribution of  $\hat{\boldsymbol{\theta}}$ , thereby taking sampling uncertainty into account. Generally, the vector  $\boldsymbol{\theta}$  in the imputation model differs from the parameter vector  $\boldsymbol{\beta}$  that governs the analysis model. Alternatively, a Bayesian approach, in which uncertainty about  $\boldsymbol{\theta}$  is incorporated by means of some prior distribution for  $\boldsymbol{\theta}$ , can also be adopted. In the context of multiple imputation, a random  $\boldsymbol{\theta}^*$  is first drawn from this prior distribution, which is then put into the distribution of  $\mathbf{Y}_i$ , and then a random  $\mathbf{Y}_i^m$  is selected from  $f(\mathbf{y}_i^m | \mathbf{y}_i^o, \boldsymbol{\theta}^*)$ . The estimate of  $\boldsymbol{\beta}$  and its estimated variance are calculated using the completed data and a potentially different, analysis model,  $(\mathbf{Y}^o, \mathbf{Y}^{m*})$ :  $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Y}) = \hat{\boldsymbol{\beta}}(\mathbf{Y}^o, \mathbf{Y}^{m*})$ , and the *within* imputation variance is  $\mathbf{W} = \widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$ . These steps are repeated a number of times ( $M$ ), producing  $\hat{\boldsymbol{\beta}}^m$  and  $\mathbf{W}^m$ , for  $m = 1, \dots, M$ . In the last phase of multiple imputation, the results of the analyses for the  $M$  imputed data sets are pooled into a single inference. The combined point estimate for the parameter of interest  $\boldsymbol{\beta}$  from the multiple imputation is simply the average of the  $M$  complete-data point estimates Schafer (1999). That is, the estimate and its estimated variance are given by:

$$\bar{\boldsymbol{\beta}} = \frac{1}{M} \sum_{m=1}^M \hat{\boldsymbol{\beta}}^m \quad \text{and} \quad \mathbf{V} = \mathbf{W} + \left( \frac{M+1}{M} \right) \mathbf{B}, \quad (3.25)$$

where

$$\mathbf{W} = \sum_{m=1}^M \frac{\mathbf{W}^m}{M} \quad \text{and} \quad \mathbf{B} = \sum_{m=1}^M \frac{(\hat{\boldsymbol{\beta}}^m - \bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^m - \bar{\boldsymbol{\beta}})'}{M-1}, \quad (3.26)$$

with  $\mathbf{W}$  denoting the average *within* imputation variance and  $\mathbf{B}$  the *between* imputation variance (Rubin, 1987).

### 3.6.3 Maximum Likelihood Estimation

When data are incomplete and under a selection model framework, subject  $i$ 's observed-data likelihood contribution takes the form:

$$L_i = \int f(\mathbf{y}_i|\boldsymbol{\theta})f(\mathbf{r}_i|\mathbf{y}_i^o, \mathbf{y}_i^m, \boldsymbol{\psi}) d\mathbf{y}_i^m. \quad (3.27)$$

In general, (3.27) does not simplify, but under MCAR (or MAR), we obtain respectively:

$$L_i = f(\mathbf{y}_i^o|\boldsymbol{\theta})f(\mathbf{r}_i|\boldsymbol{\psi}). \quad (3.28)$$

or

$$L_i = f(\mathbf{y}_i^o|\boldsymbol{\theta})f(\mathbf{r}_i|\mathbf{y}_i^o, \boldsymbol{\psi}). \quad (3.29)$$

Hence, likelihood and Bayesian inferences for the measurement model parameters  $\boldsymbol{\theta}$  can be made without explicitly formulating the missing data mechanism, provided the parameters  $\boldsymbol{\theta}$  and  $\boldsymbol{\psi}$  are distinct, meaning that their joint parameter space is the Cartesian product of the two component parameter spaces (Rubin, 1976). For Bayesian inferences, additionally the priors need to be independent (Little and Rubin, 2002). It is precisely this result which makes so-called direct likelihood analyses, valid under MCAR and MAR, appealing in a variety of settings (Molenberghs *et al.*, 2004), even though sensitivities to model assumption and non-intuitive aspects of model formulation of verification have been documented (Molenberghs, Verbeke and Beunckens, 2008).

Although maximum likelihood enjoy desirable properties, such as efficiency under appropriate regularity and the ability to calculate functions of interest, specification of the likelihood and estimation of the parameters are computationally intensive and the parameter space are restricted, especially for correlated non-Gaussian data with medium to large measurement sequences.

### 3.6.4 Non-Likelihood Estimation

When the response of interest is non-Gaussian, a first issue that arises is the lack of a non-Gaussian analogue to the multivariate normal distribution. Specification of the full likelihood becomes more problematic and fully likelihood-based methods are generally awkward, especially when high dimensional vector of correlated data arises. As a consequence, alternative methods have been in demand. Next, the two

most popular non-likelihood estimation methods, generalized estimating equations and pseudo-likelihood, will be discussed.

#### 3.6.4.1 GEE and PL

Generalized estimating equations (GEE) and pseudo-likelihood (PL) methods, introduced in Section 3.5.2 and 3.5.3 respectively, are popular missing data analysis approaches. As stated in the introduction chapter, both GEE and PL suffer from their frequentist nature and are in their basic form valid only under the MCAR assumption. A direct consequence of this is that GEE and PL need to be adapted towards a weighted, multiple imputation-based (e.g., Paik, 1997) or doubly-robust (DR) version when an MAR process operates. This will be discussed further in Chapter 4 and Chapter 5.

#### 3.6.4.2 Inverse Probability Weighting and Double Robustness

When data are missing, desirable properties of GEE such as consistency and asymptotic normality no longer holds. Instead, Inverse probability weighting (IPW) methods can be used to obtain consistent estimates (Robins, Rotnitzky and Zhao, 1995). Inverse probability weighting (IPW) methods were first proposed by Horvitz-Thompson (Cochran, 1977) in sample survey literature, where the weights are known and based on survey design. In incomplete data analysis, the general idea behind IPW method is to base estimation on the observed responses but to weight them to account for the probability of dropping out. Under MAR, the weights can be estimated as a function of the observed measurements and also as a function of the covariates and any additional variables that could help predict the unobserved measurements. In practice, a logistic model is used. The use of inverse probability weighting methods in incomplete data analysis has been increased (Robins, Rotnitzky and Zhao, 1995; Schafer, 1999; Carpenter, Kenward and Vansteelandt, 2006; Molenberghs and Kenward, 2007; Fitzmaurice *et al.*, 2009, to mention a few).

In contrast to the sample survey, where the weights are known by design and fixed, the weights in the IPW methods are not ordinarily known but estimated from the observed data. Therefore, the variance of IPW estimators must account for estimation of the weights. The sandwich estimators to be discussed in Section 4.1.4 can be used in practice. In general, weighting methods are elegant and enjoy good properties, but requires correct specification of a model for the weights.

Recently, doubly robust estimating equations (DR) has been designed to improve the efficiency of IPW. In doubly robust estimating equations (DR), the weighting

---

idea is supplemented with the use of a predictive model for the missing observations given the observed ones. Excellent reviews of this topic can be found in (Scharfstein, Rotnitzky and Robins, 1999; Van der Laan and Robins, 2003; Bang and Robins, 2005; Tsiatis, 2006; Carpenter, Kenward and Vansteelandt, 2006), and Rotnitzky (2009). Double robustness will be taken up in Chapter 4 and Chapter 5.



## Chapter 4

# Multiple Imputation Based Doubly Robust Generalized Estimating Equations

While full likelihood methods are appealing because of their flexible ignorability properties, their use for non-Gaussian outcomes can be problematic due to prohibitive computational requirements. Therefore, GEE is an attractive alternative within the marginal model family. As we have already mentioned, GEE is only valid under MCAR. One method to ensure validity of GEE under MAR is *Weighted* GEE (Section 4.1.1). *Weighted* GEE (WGEE) involves weighting observations by their inverse probability of being observed, according to an assumed dropout model. In Section 3.6.2, Multiple Imputation (MI) was described as an alternative method to handle MAR missingness. It consists of multiply imputing the missing outcomes using a parametric model, followed by analyzing the resulting complete data sets using a standard method. When GEE is considered as the standard method, the combination of MI and GEE is usually referred to as “MI-GEE”. Finally, the obtained inferences are combined into a single one. Standard multiple imputation requires MAR to hold, even though extensions exist.

In both methods (WGEE and MI-GEE), missingness needs to be addressed by means of a dropout model for WGEE or by an imputation model for MI-GEE. WGEE are consistent when the dropout model is correctly specified, while imputation methods are consistent when the imputation model is correctly specified. Recently Birhanu *et al.* (2011), extended WGEE towards so-called doubly robust estimating equations, where the weighting idea is supplemented with the use of a predictive model for the

unobserved responses, given the observed ones (Section 4.1.3). Doubly robust (DR) methods need correct specification of either the weight or predictive (imputation) model, but not necessarily both.

The focus of this chapter is to compare the efficiency and robustness of various GEE versions for incomplete data: WGEE, MI-GEE and doubly robust GEE (DR-GEE). Comparisons are made by means of a small-sample simulation study, as well as analysis of a case study, Toenail data (Section 2.1). In the simulation study, the behavior of the methods is studied under correctly specified and misspecified models. In this way, efficiency and robustness of the methods under misspecification of either the dropout model, the imputation model, or both, can be explored.

The outline of this chapter is as follows. In Section 4.1, we discuss methods used for analyzing incomplete non-Gaussian longitudinal data that are valid under the MAR assumption, with main attention to WGEE, MI-GEE and doubly robust (DR) estimation methods. A description and result of the small-sample simulation is provided in Section 4.2. Results of the Toenail data are presented in Section 4.3. Finally, a brief discussion and some concluding remarks are provided in Section 4.4. The contribution of this chapter has been published in Birhanu *et al.* (2011).

## 4.1 Extensions of GEE under MAR

As mentioned in the previous chapters, GEE is an attractive approach for non-Gaussian data within the marginal model family. However, it is based on frequentist methods and thus requires the missingness to be MCAR. The extensions of GEE to WGEE to ensure its validity under MAR and to doubly robust (DR) GEE to improve efficiency of IPW that have been proposed by Robins, Rotnitzky and Zhao (1995) and Scharfstein, Rotnitzky and Robins (1999) respectively, and MI-GEE will be the focus of this section.

### 4.1.1 Weighted Generalized Estimating Equations

Due to the fact that they are based on frequentist considerations, inferences under GEE are valid only under the strong assumption that the missing data are MCAR. In response to this, and to allow for MAR missingness, Robins, Rotnitzky and Zhao (1995) proposed a class of so-called *weighted* estimating equations (WGEE). The idea of WGEE is to weigh each subject's contribution in the GEEs by the inverse probability, either of being fully observed, or of being observed up to a certain time. Thus, anyone staying in the study is considered representative of himself as well as of a



number of similar subjects that did drop out from the study. The method is elegant and enjoys good properties, but requires specification of a model for the weights. Let  $\pi_i$  be the probability for subject  $i$  to be completely observed and  $\pi'_i$  the probability for subject  $i$  to drop out at occasion  $d_i$ . These can be written as

$$\pi_i = \prod_{\ell=2}^{n_i} (1 - p_{i\ell}), \quad (4.1)$$

$$\pi'_i = \left[ \prod_{\ell=2}^{d_i-1} (1 - p_{i\ell}) \right] \cdot p_{id_i}, \quad (4.2)$$

where  $p_{i\ell} = P(D_i = \ell | D_i \geq \ell, Y_{i\bar{\ell}}, X_{i\bar{\ell}})$  are the component probabilities of dropping out at occasion  $\ell$ , given the subject is still in the study, the covariate history  $X_{i\bar{\ell}}$  and the outcome history  $Y_{i\bar{\ell}}$ . In such a case, one can opt either for WGEE based on the completers only.

$$U(\boldsymbol{\beta}) = \sum_{i=1}^N \frac{\tilde{R}_i}{\pi_i} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} V_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \quad (4.3)$$

with  $\tilde{R}_i = 1$  if a subject is fully observed and 0 otherwise, or, upon using (4.2), for WGEE using all subjects:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^N \frac{1}{\pi'_i} \frac{\partial \boldsymbol{\mu}_i^o}{\partial \boldsymbol{\beta}'} (V_i^o)^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o) = \mathbf{0}. \quad (4.4)$$

Here the subscript ‘o’ indicates the portion corresponding to the observed data in the corresponding matrix or vector. Of course, with (4.3), the incomplete subjects also contribute through the model for the dropout probabilities  $\pi_i$ . When assuming the weights are fixed and known, expressions (3.25) and (3.26) can be used for variance estimation. In practice, (4.3) and (4.4) use weights estimated from the observed data that will impact the variance estimation and have to be taken into account. This will be taken up in Section 4.1.4. The above development focuses on dropout but can be generalized to encompass non-monotone missingness as well (Vansteelandt, Rotnitzky, and Robins (2007).

WGEE enjoy robustness properties similar to the ones for regular GEE, i.e., the correlation structure does not need to be correctly specified. Applying WGEE is technically feasible and can be conducted using the SAS procedure GENMOD. Of course, some extra programming is needed to construct the weights.

### 4.1.2 Multiple Imputation based Generalized Estimating Equations

Because in WGEE all subjects are given weights, calculated using the hypothesized dropout model, any misspecification of this dropout model will affect all subjects, and thus the results. Alternatively, one can consider multiple imputation (MI) together with GEE, referred as MI-GEE. In essence, this method comes down to first using the predictive distribution of the unobserved outcomes given the observed ones and perhaps covariates. After this step, the missing data mechanism can be further ignored, provided the missing data mechanism is MAR. Then, misspecification occurring at the imputation step will only affect the unobserved (i.e., imputed) but not the observed part of the data. Meng's (1994) results show that, as long as the imputation model is not grossly misspecified, this approach will perform well. Simulation studies done by Beunckens, Sotto and Molenberghs (2008) showed that MI-GEE has good robustness properties against model misspecification, in comparison with WGEE.

### 4.1.3 Doubly Robust Generalized Estimating Equations

Recently IPW has been extended towards the so-called doubly robust estimating equations (DR), where the weighting idea is supplemented with the use of a predictive model for the unobserved responses, given the observed ones. Excellent reviews can be found in Scharfstein, Rotnitzky and Robins (1999), Van der Laan and Robins (2003), Bang and Robins (2005), Carpenter, Kenward and Vansteelandt (2006), and Rotnitzky (2009).

Let us introduce doubly robust estimating equations starting from conventional generalized estimating equations:

$$\mathbf{U} = \sum_{i=1}^N \mathbf{U}_i(\boldsymbol{\beta} | \mathbf{Y}_i, \mathbf{x}_i) \quad (4.5)$$

Assume that  $E(\mathbf{U}) = \mathbf{0}$ .

For this setup the weighted estimating equation (IPW) could be written as

$$\mathbf{U}_{\text{IPWCC}} = \sum_{i=1}^N \frac{\tilde{R}_i}{\pi_i} \mathbf{U}_i(\mathbf{Y}_i), \quad (4.6)$$

where  $\tilde{R}_i = 1$  if subject  $i$  is fully observed and 0 otherwise, and  $\pi_i$  be the probability for subject  $i$  to be completely observed.

The expression in equation (4.6) only includes fully observed subjects, and this is where the information is lost relative to likelihood methods, which include information

on partially observed subjects. Any term with expectation zero can be added to (4.6) without changing its property of unbiasedness and consistency of the resulting parameter estimators.

Hence, doubly robust versions could be

$$\mathbf{U}_{\text{IPWCC,dr}} = \sum_{i=1}^N \left[ \frac{\tilde{R}_i}{\pi_i} \mathbf{U}_i(\mathbf{Y}_i) + \left( 1 - \frac{\tilde{R}_i}{\pi_i} \right) E_{Y^m|y^o} \mathbf{U}_i(\mathbf{Y}_i) \right], \quad (4.7)$$

where the second term in (4.7) refers to predictive terms of the unobserved outcomes given the observed ones. Remarkably, appropriate choices of these functions, like the one in (4.7), offers not only efficiency improvements over standard weighted GEE (IPW), but also bias protection against misspecification of the model for the dropout probabilities.

The predictive model may show varying degrees of complexity, depending on the type of estimating equations. Molenberghs *et al.* (2011) considered situations where a fully analytic approach is possible (a thorough discussion is given in Chapter 5). A sufficiently rich predictive model could be used, such as logistic regression for example. Evidently, such a predictive model would, strictly speaking, be incompatible with the actual model under consideration but, as Bang and Robins (2005) point out, virtually all parametric models are misspecified to some extent. In this sense, a reasonable predictive model, coupled with a sensible missingness model for the weights, often considerably increases efficiency and reduces bias. Bang and Robins' simulation results were encouraging in this respect. In a similar vein, with multiple imputation, Meng (1994) shows that so-called *uncongenial imputation models* can still lead to inferences that are practically acceptable. Daniel (2008) usefully combined doubly robust estimating equations with multiple imputation ideas. Next, we will consider an approach similar to this.

### Data Augmentation for MI based Doubly Robust Estimation

Like with all multiple imputation strategies, the principle is to first augment the data by re-ordering subset of the observed sequences and then apply the estimation method of choice on the augmented data as if they were the fully observed data set. The method uses multiple imputation in a weighted analysis on the augmented data.

In general, assume that the observed data consist of  $(Y_{i1}, \dots, Y_{it})$ . We then replace the observed data by  $t$  patterns  $(Y_{i1}, \dots, Y_{ij})$ , for  $j = 1, \dots, t$ , with weight contributions

$$\left( \prod_{k=2}^j \pi_i^{k-1} \right) \cdot \left( 1 - \pi_i^{j+1-1} \right)$$

for new patterns  $j = 1, \dots, t-1$  and

$$\prod_{k=2}^t \pi_i^{k-1}$$

for the last one, pattern  $t$ .

Next, we will illustrate the data augmentation for a hypothetical longitudinal data set measured at two time points. Assume that the observed data consist of the following two observed sequences:  $(Y_{i1}, \cdot)$  and  $(Y_{i1}, Y_{i2})$  with weight contributions. For the first two subjects:

|   |  |
|---|--|
| Original data   | Weight   |
| $\begin{pmatrix} Y_{1,i1} & \cdot \\ Y_{2,i1} & Y_{2,i2} \end{pmatrix}$ | $\begin{matrix} \pi_i^{(2)-1} \\ \pi_i^{(2)-1} \end{matrix}$ |

where  $\pi_i^{(2)}$  denotes the probability to dropout at time 2.

We then create another data set by partially “wiping out” the second outcome:

|  |  |
|--|--|
| Augmentation matrix  | Weight   |
| $\begin{pmatrix} Y_{1,i1} & \cdot \\ Y_{2,i1} & \cdot \end{pmatrix}$ | $\begin{matrix} 1 - \pi_i^{(2)-1} \\ 1 - \pi_i^{(2)-1} \end{matrix}$ |

Augmenting the newly formed data set to the original data set produces:

|   |   |
|---|---|
| Augmented data  | Weight  |
| $\begin{pmatrix} Y_{1,i1} & \cdot \\ Y_{2,i1} & \cdot \\ Y_{2,i1} & Y_{2,i2} \end{pmatrix}$ | $\begin{matrix} 1 \\ 1 - \pi_i^{(2)-1} \\ \pi_i^{(2)-1} \end{matrix}$ |

Let us consider another longitudinal data set with three outcomes. Assuming no dropout at the first time point, there are three possible dropout patterns: dropout at time 2, dropout at time 3 and no dropout.

|   |                 |                 |
|---|-----------------|-----------------|
| Original data   | Weight          |                 |
| $\begin{pmatrix} Y_{1,i1} & \cdot & \cdot \\ Y_{2,i1} & Y_{2,i2} & \cdot \\ Y_{3,i1} & Y_{3,i2} & Y_{3,i3} \end{pmatrix}$ | time 2          | time 3          |
|   | $\pi_i^{(2)-1}$ |                 |
|   | $\pi_i^{(2)-1}$ | $\pi_i^{(3)-1}$ |
|   | $\pi_i^{(2)-1}$ | $\pi_i^{(3)-1}$ |

where  $\pi_i^{(2)}$  and  $\pi_i^{(3)}$  denotes the probability to dropout at time two and three respectively.

**Step 1:** Create another data set by partially “wiping out” the second and third outcomes:

$$\begin{array}{cc} \text{Augmentation matrix} & \text{Weight} \\ \left( \begin{array}{ccc} Y_{1,i1} & \cdot & \cdot \\ Y_{2,i1} & \cdot & \cdot \\ Y_{3,i1} & \cdot & \cdot \end{array} \right) & \begin{array}{l} 1 - \pi_i^{(2)-1} \\ 1 - \pi_i^{(2)-1} \\ 1 - \pi_i^{(2)-1} \end{array} \end{array}$$

This leads to:

$$\begin{array}{cc} \text{Augmentation matrix} & \text{Weight} \\ \left( \begin{array}{ccc} Y_{1,i1} & \cdot & \cdot \\ Y_{2,i1} & \cdot & \cdot \\ Y_{3,i1} & \cdot & \cdot \\ Y_{2,i1} & Y_{2,i2} & \cdot \\ Y_{3,i1} & Y_{3,i2} & Y_{3,i3} \end{array} \right) & \begin{array}{cc} \text{time 2} & \text{time 3} \\ 1 & \\ 1 - \pi_i^{(2)-1} & \\ 1 - \pi_i^{(2)-1} & \\ \pi_i^{(2)-1} & \pi_i^{(3)-1} \\ \pi_i^{(2)-1} & \pi_i^{(3)-1} \end{array} \end{array}$$

**Step 2:** Focusing on time 3:

$$\begin{array}{cc} \text{Augmentation matrix} & \text{Weight} \\ \left( \begin{array}{ccc} Y_{2,i1} & Y_{2,i2} & \cdot \\ Y_{3,i1} & Y_{3,i2} & \cdot \end{array} \right) & \begin{array}{cc} \text{time 2} & \text{time 3} \\ \pi_i^{(2)-1} & 1 - \pi_i^{(3)-1} \\ \pi_i^{(2)-1} & 1 - \pi_i^{(3)-1} \end{array} \end{array}$$

This finally produces:

$$\begin{array}{cc} \text{Augmented data} & \text{Weight} \\ \left( \begin{array}{ccc} Y_{1,i1} & \cdot & \cdot \\ Y_{2,i1} & \cdot & \cdot \\ Y_{3,i1} & \cdot & \cdot \\ Y_{2,i1} & Y_{2,i2} & \cdot \\ Y_{3,i1} & Y_{3,i2} & \cdot \\ Y_{3,i1} & Y_{3,i2} & Y_{3,i3} \end{array} \right) & \begin{array}{l} 1 \\ 1 - \pi_i^{(2)-1} \\ 1 - \pi_i^{(2)-1} \\ \pi_i^{(2)-1} \\ \pi_i^{(2)-1} \times (1 - \pi_i^{(3)-1}) \\ \pi_i^{(2)-1} \times (\pi_i^{(2)-1}) \end{array} \end{array}$$

Based on this, we proceed in two distinct ways:

**MI-GEE(Aug):** Apply Multiple Imputation (MI) on the so-augmented data. In this way, due to the ‘wiping out’ of a fraction of the sequence, leading to the need of using expectations, the derivation of precision is not straightforward. However, the precision estimation we are proposing as shown in Section 4.1.4, especially

when the variability in the weights are considered, is a viable way forward. In contrast to sample surveys, here the weights are not fixed and known but estimated from the data. This estimation of the weights induce additional variability that needs to be taken into account.

**MI-WGEE:** In this way, one simply multiply imputes the missing outcomes multiply, without ‘wiping out’ the observed data. As such, classical MI results, combined with sandwich estimators will provide standard errors.

Next, we turn to precision estimation for this case.

#### 4.1.4 Precision Estimation

A general expression for the precision of the estimates obtained using generalized estimating equations is given in (3.12). When the single or doubly robust versions of GEE (WGEE or DR-GEE) are used, with a parametric model for dropout, then the uncertainty induced by estimation of the  $\psi$  parameters needs to be accommodated. As shorthand for any of the forms (4.3), we write  $\mathbf{U} = \sum_{i=1}^N \mathbf{V}_i(\boldsymbol{\beta})$ , and the parameters  $\psi$  are estimated from score or estimating equations  $\mathbf{W} = \sum_{i=1}^N \mathbf{W}_i(\psi)$ . The entire score for subject  $i$  is  $\mathbf{S}_i = (\mathbf{V}'_i, \mathbf{W}'_i)'$ . The asymptotic variance-covariance matrix can then be consistently estimated by  $\widehat{I}_0^{-1} \widehat{I}_1 \widehat{I}_0^{-1}$ , with

$$I_0 = \sum_{i=1}^N \begin{pmatrix} \frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\beta}} & \frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\psi}} \\ 0 & \frac{\partial \mathbf{W}_i}{\partial \boldsymbol{\psi}} \end{pmatrix}, \quad (4.8)$$

$$I_1 = \sum_{i=1}^N \mathbf{S}_i(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}) \mathbf{S}'_i(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}). \quad (4.9)$$

See also Bang and Robins (2005), Molenberghs and Kenward (2007), and Rotnitzky (2009).

When ignoring variability in the weights, one first uses a conventional sandwich estimator, by assuming all weights are fixed and known. Then, expressions (3.25) and (3.26) can be used. Of course, this will underestimate the true variability. To acknowledge the variability in weights, the sandwich estimators have to be computed using the full expressions (4.8) and (4.9).

## 4.2 Simulation Study

In what precedes, various approaches to overcome the bias occurring in GEE under MAR have been proposed. WGEE is unbiased for a correctly specified dropout and

mean structure of the measurement model. MI-GEE requires compatibility between the imputation and estimation model to be correctly specified. On the other hand, doubly robust GEEs (DR-GEE) need correct specification of either of the models but not necessarily both. It is of interest to quantify the bias and precision under various scenarios for misspecification. To this end, a small-sample simulation study was conducted.

In the simulations, we distinguish between the data-generating and the analysis stage. In the first stage, a data-generating model is defined. Under the selection model framework, this generating model consists of a measurement model on the one hand, and a dropout model, given the measurement model on the other. In the analysis stage, a distinction should be made among three types of models: a measurement model, a dropout model and an imputation model. For the WGEE approach, only a marginal measurement model and a dropout model need to be specified. In contrast, the analysis stage for MI-GEE would entail the specification of an imputation model, rather than a dropout model, as well as a marginal measurement model. For the standard doubly robust versions of GEE, MI-GEE(Aug) and MI-WGEE, both the dropout and imputation models need to be specified.

To assess the distinctive and relative merits of the methods of interest, we consider their performance, first in the case without any misspecification, then under various scenarios of misspecification in some or all of the models for the DR-GEEs versions. Next, we spell out the data-generating models, followed by a description of the simulation study's design, after which the results are presented.

### 4.2.1 Data-generating Models

We generated an outcome at three time points from the Bahadur model, formally introduced in Section 3.5.4. The measurement model incorporated a binary treatment indicator. For the dropout model, an MAR mechanism was considered. Assuming that dropout can occur only after the first time point, there are three possible dropout patterns: (1) dropout at the second time point, (2) dropout at the third time point, and (3) no dropout. Note that we restrict the simulation setting to short sequences, because the higher-order Bahadur models would become prohibitive to generate from.

Denote by  $t_j$  the time point at which measurement  $j$  is taken and by  $x_i$  the treatment indicator. Consider a Bahadur model, which follows general formulation (3.22), with

$$\text{logit}(\pi_{ij}) = \text{logit}[P(Y_{ij} = 1|x_i, t_j)] = \beta_0 + \beta_x x_i + \beta_t t_j + \beta_{xt} x_i t_j, \quad (4.10)$$

where we choose  $\beta_0 = -0.25$ ,  $\beta_x = 0.5$ ,  $\beta_t = 0.2$  and  $\beta_{xt} = -0.8$ , with two- and three-way correlation coefficients equal to  $\rho_{ij_1j_2} = 0.2$  and  $\rho_{ij_1j_2j_3} = 0$ , respectively. The latter define an exchangeable correlation structure. Our choice for linear time evolutions, at the scale of the linear predictor and within each of the treatment arms, allows us to distinguish between misspecification effects on cross-sectional parameters ( $\beta_0$  and  $\beta_x$ ), longitudinal parameters ( $\beta_t$ ), and parameters combining aspects of both ( $\beta_{xt}$ ). In practice, for example, in a clinical trial, it might be advisable to allow for an unstructured, saturated treatment-by-time model, reducing the risk of model misspecification and in line with recommendations made by Molenberghs *et al.* (2004) and several references listed therein.

The missingness process is assumed to be MAR and the probability of dropout at time point  $j$  given  $x_i$  and the measurement at the previous time point, is modeled by a logistic regression of the form

$$\text{logit}[P(D_i = j|x_i, y_{i,j-1}, D_i \geq j)] = \psi_0 + \psi_x x_i + \psi_{prev} y_{i,j-1}, \quad (4.11)$$

where  $j = 2, 3, 4$ ,  $\psi_0 = -0.5$ ,  $\psi_x = -0.6$ , and  $\psi_{prev} = -3.5$ . The combination of this MAR logistic dropout model with the Bahadur measurement model (4.10) defines our data generating model, which will hereinafter be referred to as GM I. GM I yields the following proportions of subjects within each of the response patterns: 68% completers (33% for  $x = 0$  and 35% for  $x = 1$ ), 15% with the last outcome missing (7% for  $x = 0$  and 8% for  $x = 1$ ), and 18% with only the first outcome observed (10% for  $x = 0$  and 8% for  $x = 1$ ).

We further consider a second data generating model, GM II, in which the outcomes are generated based on (4.10) and non-random missingness is induced via the following MNAR logistic regression model:

$$\text{logit}[P(D_i = j|x_i, y_{i,j-1}, D_i \geq j)] = \psi_0 + \psi_x x_i + \psi_{prev} y_{i,j-1} + \psi_{curr} y_{ij}, \quad (4.12)$$

where  $j = 2, 3, 4$ ,  $\psi_0 = -0.5$ ,  $\psi_x = -0.6$ ,  $\psi_{prev} = -1.0$  and  $\psi_{curr} = -2.0$ . For GM II, there are around 71% completers (36% for  $x = 0$  and 35% for  $x = 1$ ), 13% with only the last outcome missing (6% for  $x = 0$  and 8% for  $x = 1$ ), and 16% with only one non-missing outcome (8% each for  $x = 0$  and  $x = 1$ ).

It can be recalled that all the methods under investigation in this chapter assume an MAR missing data mechanism. As such, these methods are inherently “incorrect” or misspecified when applied to data with non-random missingness. Thus, comparisons under the MAR case are no longer meaningful for the MNAR case, unless, of course, modeling approaches for non-random missingness are entertained. Without recourse to the latter, we nevertheless proceed to investigate the performance of the



different approaches under certain misspecification settings to further evaluate their robustness or lack thereof.

### 4.2.2 Design of the Simulation Study

We assume a sample of size  $N = 500$  subjects, equally divided between the two treatment groups. Such a choice is practically relevant, given that many biopharmaceutical trials employ about 250 to 300 patients per treatment arm. Based on the underlying probabilities from GM I, 250 observations were generated randomly for each treatment group. A total of  $S = 500$  such samples were then generated. For the various extensions of GEE considered, the same working correlation structure as assumed during data generation, i.e., an exchangeable type, is employed in the analysis.

Several measures are computed to gauge the relative performance of the various methods. First, we define bias as the difference between the estimate and the true value of the parameter, i.e.,  $\text{Bias}(\hat{\beta}) = \bar{\beta} - \beta$ . Further, the average ( $\bar{\beta}$ ) of the estimators over all  $S = 500$  samples, its variance for a sample of size  $N$  ( $\text{Var}(\bar{\beta})$ ), and the Monte Carlo MSE ( $MSE_{MC}$ ) are computed as:

$$\bar{\beta} = \sum_{i=1}^S \frac{\hat{\beta}_i}{S}, \quad \text{Var}(\bar{\beta}) = \sum_{i=1}^S \frac{(\hat{\beta}_i - \bar{\beta})^2}{S-1},$$

$$\text{MSE} \equiv MSE_{MC}(\bar{\beta}) = \text{Bias}^2(\bar{\beta}) + \text{Var}(\bar{\beta}).$$

In addition, the empirically-corrected MSE ( $MSE_{\text{emp}}$ ) and model-based MSE ( $MSE_{\text{mod}}$ ) are computed.

### 4.2.3 Results

In the simulation study, the behavior of the methods is studied in terms of bias, variance and mean squared error (MSE) of the estimators, under correctly specified and misspecified models. In this way, robustness of the methods under misspecification of either the dropout model, the imputation model, or both, can be explored. We consider, in turn, various types of misspecification, first for the MAR-based GM I, then for the MNAR-based GM II.

#### MAR: Everything Correctly Specified

We first investigate the individual merits of each method when every one of its aspects is correctly specified. Recall that GM I is based on a Bahadur measurement model and

a logistic model for dropout that is reflective of an MAR mechanism, i.e., depending on the previous measurement as well as the treatment indicator. An appropriate analysis model would consist of a measurement model and either a dropout or imputation model that match those of this GM I. Because GEE methods are moment-based versions of the Bahadur model (Section 3.5.4), a GEE-based version, with the same structure as that of the underlying measurement model would be suitable. To address the MAR nature of missingness, the GEE-based approach is supplemented with either a weighting scheme and/or imputations obtained from a model of the same form as that of the underlying dropout model, that is, a weight model or/and an imputation model containing the treatment indicator and the previous measurement as predictor. WGEE, MI-GEE, as well as the doubly robust GEE versions (MI-WGEE and MI-GEE(Aug)) proposed in Section 4.1.3, were so fitted for GM I and the results are shown in Table 4.1. From Table 4.1, it can be observed that the doubly robust MI-WGEE approach consistently yields the least bias, and MI-GEE leads to slightly larger bias than the former. The MSEs of MI-WGEE are worse than MSEs of MI-GEE and MI-GEE(Aug) (specially for  $\beta_x$  and  $\beta_{xt}$ ). Compared to MI-GEE and the DR-GEEs, WGEE has the worst Monte Carlo MSE (except for  $\beta_0$ ). Furthermore, MI-GEE(Aug) gives empirically corrected MSEs that are closer to the asymptotic MSEs when the variability in the weights is acknowledged.

Table 4.1: *Simulation study for MAR (GM I): Everything correctly specified.*

| WGEE                    |         |                    |            |                    |             |                 |             |
|-------------------------|---------|--------------------|------------|--------------------|-------------|-----------------|-------------|
| Par.                    | Bias    | Var <sub>EST</sub> | $MSE_{MC}$ | No var. in weights |             | Var. in weights |             |
|                         |         |                    |            | $MSE_{emp}$        | $MSE_{mod}$ | $MSE_{emp}$     | $MSE_{mod}$ |
| $\beta_0$               | -0.0225 | 0.3031             | 0.3036     | 0.5096             | 0.1240      | 0.1968          | 0.0363      |
| $\beta_t$               | 0.0002  | 0.0662             | 0.0662     | 0.2137             | 0.0670      | 0.0936          | 0.0160      |
| $\beta_x$               | 0.0407  | 0.8947             | 0.8964     | 0.7639             | 0.1858      | 0.2998          | 0.0625      |
| $\beta_{xt}$            | -0.0557 | 0.1785             | 0.1815     | 0.3371             | 0.1095      | 0.1599          | 0.0317      |
| MI-GEE                  |         |                    |            |                    |             |                 |             |
| $\beta_0$               | 0.0124  | 0.0427             | 0.0428     | 0.1988             | 0.1986      |                 |             |
| $\beta_t$               | -0.0093 | 0.0106             | 0.0107     | 0.0973             | 0.0981      |                 |             |
| $\beta_x$               | -0.0348 | 0.0747             | 0.0759     | 0.2858             | 0.2852      |                 |             |
| $\beta_{xt}$            | 0.0236  | 0.0191             | 0.0197     | 0.1437             | 0.1439      |                 |             |
| MI-WGEE (Non-Augmented) |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.0102 | 0.2289             | 0.2290     | 0.5668             | 0.2936      | 0.3352          | 0.2714      |
| $\beta_t$               | -0.0054 | 0.0407             | 0.0407     | 0.2899             | 0.1999      | 0.2126          | 0.1933      |
| $\beta_x$               | -0.0033 | 0.8108             | 0.8108     | 0.8927             | 0.4637      | 0.5265          | 0.4344      |
| $\beta_{xt}$            | -0.0183 | 0.1273             | 0.1276     | 0.4816             | 0.3367      | 0.3588          | 0.3273      |
| MI-GEE (Augmented)      |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.8060 | 0.0532             | 0.7029     | 0.8591             | 0.8056      | 0.7967          | 0.7033      |
| $\beta_t$               | 0.2230  | 0.0130             | 0.0627     | 0.1494             | 0.1239      | 0.1166          | 0.0860      |
| $\beta_x$               | 0.2388  | 0.0863             | 0.1433     | 0.3491             | 0.2887      | 0.2673          | 0.1301      |
| $\beta_{xt}$            | -0.0695 | 0.0211             | 0.0259     | 0.1487             | 0.1194      | 0.1056          | 0.0551      |

### MAR: Dropout and Measurement Models Correct, Imputation Model Incorrect

We compare, under this scenario, MI-GEE with the DR-GEEs, all having a correctly specified measurement and dropout model, but with an incorrectly specified imputation model in the sense that the previous outcome is omitted from the imputation model. Since the underlying missingness model (4.11) does include  $y_{i,j-1}$ , omission of such in the imputation model is a clear misspecification. The results are given in

Table 4.2.

Table 4.2: *Simulation study for MAR (GM I): Misspecified imputation model. (Misspecification in the form of omission of the previous outcome,  $y_{i,j-1}$ , from the imputation model).*

| MI-GEE                  |         |                    |            |                    |             |                 |             |
|-------------------------|---------|--------------------|------------|--------------------|-------------|-----------------|-------------|
| Par.                    | Bias    | Var <sub>EST</sub> | $MSE_{MC}$ | No var. in weights |             | Var. in weights |             |
|                         |         |                    |            | $MSE_{emp}$        | $MSE_{mod}$ | $MSE_{emp}$     | $MSE_{mod}$ |
| $\beta_0$               | -0.0047 | 0.0424             | 0.0424     | 0.1971             | 0.1999      |                 |             |
| $\beta_t$               | 0.0271  | 0.0106             | 0.0114     | 0.0975             | 0.0992      |                 |             |
| $\beta_x$               | -0.0211 | 0.0737             | 0.0742     | 0.2823             | 0.2855      |                 |             |
| $\beta_{xt}$            | 0.0030  | 0.0191             | 0.0191     | 0.1415             | 0.1437      |                 |             |
| MI-WGEE (Non-Augmented) |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.0401 | 0.1999             | 0.2015     | 0.5539             | 0.2886      | 0.3299          | 0.2653      |
| $\beta_t$               | 0.0239  | 0.0331             | 0.0337     | 0.2844             | 0.1947      | 0.2075          | 0.1874      |
| $\beta_x$               | 0.0198  | 0.7399             | 0.7403     | 0.8731             | 0.4523      | 0.5131          | 0.4219      |
| $\beta_{xt}$            | -0.0324 | 0.1071             | 0.1082     | 0.4655             | 0.3201      | 0.3414          | 0.3099      |
| MI-GEE (Augmented)      |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.8197 | 0.0522             | 0.7240     | 0.8801             | 0.8276      | 0.8159          | 0.7242      |
| $\beta_t$               | 0.2473  | 0.0128             | 0.0739     | 0.1604             | 0.1354      | 0.1271          | 0.0967      |
| $\beta_x$               | 0.2496  | 0.0851             | 0.1474     | 0.3527             | 0.2945      | 0.2695          | 0.1357      |
| $\beta_{xt}$            | -0.0813 | 0.0206             | 0.0272     | 0.1497             | 0.1219      | 0.1067          | 0.0576      |

MI-GEE gives the smallest or at least comparable bias with MI-WGEE, while MI-GEE(Aug) shows somewhat larger bias. Moreover, in comparison with the weighted versions, MI-GEE leads to the smallest empirically corrected MSEs, as can be expected, since no variability in the weights is added to the conventional MI variability. Between the DR-GEEs, smaller empirically corrected MSEs are obtained under MI-GEE(Aug) when the weight variability is taken into account.

**MAR: Imputation and Measurement Models Correct, Dropout Model Incorrect**

Keeping the measurement and imputation models correctly specified, we now investigate the effects of misspecification in the dropout model. Here, as in the previous subsection, the misspecification in the dropout model is again in the form of omission of the previous outcome from the dropout model, from which weights are obtained. The results are given in Table 4.3.

In this setting, the doubly robust MI-GEE(Aug) consistently gives the least biased and least variable estimates, and thus, the smallest empirically corrected MSEs when the variability in the weights is accounted for. Moreover, these corrected MSEs are closer to the Monte Carlo MSE for MI-GEE(Aug) than they are under the other two methods. For the three approaches compared, WGEE, MI-WGEE and MI-GEE(Aug), improved variability estimates are obtained when one acknowledges the variability in the weights.

MI-WGEE seems to yield quite imprecise estimates when the weight variability is considered. This does not come as a surprise because a misspecification in the dropout model comes into play twice for this method. First, imprecise weights that arise from a misspecified dropout model can impact the WGEE part of the method, in which the cases are weighted (incorrectly). In addition, a misspecification in the dropout model can also affect the precision of the estimates via the adjustment done to account for the variability in the weights. Hence, it is not totally surprising that the impact of a misspecified dropout model for MI-WGEE can be quite substantial.

Table 4.3: *Simulation study for MAR (GM I): Misspecified dropout model. (Misspecification in the form of omission of the previous outcome,  $y_{i,j-1}$ , from the dropout model).*

| WGEE                    |         |                    |            |                    |             |                 |             |
|-------------------------|---------|--------------------|------------|--------------------|-------------|-----------------|-------------|
| Par.                    | Bias    | Var <sub>EST</sub> | $MSE_{MC}$ | No var. in weights |             | Var. in weights |             |
|                         |         |                    |            | $MSE_{emp}$        | $MSE_{mod}$ | $MSE_{emp}$     | $MSE_{mod}$ |
| $\beta_0$               | -0.7283 | 0.0534             | 0.5839     | 0.7714             | 0.6408      | 0.7442          | 0.5551      |
| $\beta_t$               | 0.0680  | 0.0140             | 0.0186     | 0.1212             | 0.0645      | 0.0990          | 0.0152      |
| $\beta_x$               | 0.2373  | 0.1482             | 0.2045     | 0.4469             | 0.2262      | 0.3929          | 0.0998      |
| $\beta_{xt}$            | -0.1141 | 0.0393             | 0.0523     | 0.2206             | 0.1149      | 0.1947          | 0.0335      |
| MI-WGEE (Non-Augmented) |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.8220 | 0.0579             | 0.7335     | 0.9540             | 0.8615      | 0.9398          | 0.8335      |
| $\beta_t$               | 0.1864  | 0.0165             | 0.0513     | 0.1869             | 0.1513      | 0.1743          | 0.1425      |
| $\beta_x$               | 0.1207  | 0.1265             | 0.1411     | 0.4365             | 0.2791      | 0.4023          | 0.2351      |
| $\beta_{xt}$            | -0.0276 | 0.0325             | 0.0333     | 0.2421             | 0.1774      | 0.2192          | 0.1627      |
| MI-GEE (Augmented)      |         |                    |            |                    |             |                 |             |
| $\beta_0$               | 0.2757  | 0.0400             | 0.1160     | 0.2704             | 0.2322      | 0.2049          | 0.1292      |
| $\beta_t$               | -0.0717 | 0.0100             | 0.0152     | 0.0965             | 0.0809      | 0.0695          | 0.0426      |
| $\beta_x$               | -0.1518 | 0.0706             | 0.0937     | 0.3025             | 0.2489      | 0.2060          | 0.0927      |
| $\beta_{xt}$            | 0.0505  | 0.0181             | 0.0207     | 0.1382             | 0.1149      | 0.0957          | 0.0518      |

### MAR: Measurement Model Correct, Imputation and Dropout Models Incorrect

A final comparison for the MAR case evaluates the relative performance of the augmented and non-augmented versions of DR-GEE with ordinary WGEE and singly robust MI-GEE, under misspecification in both the dropout and imputation models. For both cases, misspecification is again in terms of omission of the previous outcome. The results are given in Table 4.4; for ease of comparison, the first panels of Tables 4.2 and 4.3 are replicated here.

Table 4.4: *Simulation study for MAR (GM I): Both models misspecified. (Misspecification in the form of omission of the previous outcome,  $y_{i,j-1}$ , from both the dropout and imputation models).*

| WGEE                    |         |                    |                   |                    |                    |                    |                    |
|-------------------------|---------|--------------------|-------------------|--------------------|--------------------|--------------------|--------------------|
| Par.                    | Bias    | Var <sub>EST</sub> | MSE <sub>MC</sub> | No var. in weights |                    | Var. in weights    |                    |
|                         |         |                    |                   | MSE <sub>emp</sub> | MSE <sub>mod</sub> | MSE <sub>emp</sub> | MSE <sub>mod</sub> |
| $\beta_0$               | -0.7283 | 0.0534             | 0.5839            | 0.7714             | 0.6408             | 0.7442             | 0.5551             |
| $\beta_t$               | 0.0680  | 0.0140             | 0.0186            | 0.1212             | 0.0645             | 0.0990             | 0.0152             |
| $\beta_x$               | 0.2373  | 0.1482             | 0.2045            | 0.4469             | 0.2262             | 0.3929             | 0.0998             |
| $\beta_{xt}$            | -0.1141 | 0.0393             | 0.0523            | 0.2206             | 0.1149             | 0.1947             | 0.0335             |
| MI-GEE                  |         |                    |                   |                    |                    |                    |                    |
| $\beta_0$               | -0.0047 | 0.0424             | 0.0424            | 0.1971             | 0.1999             |                    |                    |
| $\beta_t$               | 0.0271  | 0.0106             | 0.0114            | 0.0975             | 0.0992             |                    |                    |
| $\beta_x$               | -0.0211 | 0.0737             | 0.0742            | 0.2823             | 0.2855             |                    |                    |
| $\beta_{xt}$            | 0.0030  | 0.0191             | 0.0191            | 0.1415             | 0.1437             |                    |                    |
| MI-WGEE (Non-Augmented) |         |                    |                   |                    |                    |                    |                    |
| $\beta_0$               | -0.8527 | 0.0478             | 0.7749            | 0.9977             | 0.9119             | 0.9877             | 0.8828             |
| $\beta_t$               | 0.2605  | 0.0135             | 0.0814            | 0.2175             | 0.1839             | 0.2069             | 0.1746             |
| $\beta_x$               | 0.1352  | 0.1087             | 0.1269            | 0.4240             | 0.2801             | 0.3974             | 0.2364             |
| $\beta_{xt}$            | -0.0529 | 0.0273             | 0.0301            | 0.2331             | 0.1737             | 0.2129             | 0.1588             |
| MI-GEE (Augmented)      |         |                    |                   |                    |                    |                    |                    |
| $\beta_0$               | 0.2598  | 0.0396             | 0.1071            | 0.2615             | 0.2240             | 0.1942             | 0.1192             |
| $\beta_t$               | -0.0451 | 0.0099             | 0.0120            | 0.0934             | 0.0780             | 0.0657             | 0.0386             |
| $\beta_x$               | -0.1398 | 0.0690             | 0.0886            | 0.2982             | 0.2467             | 0.2003             | 0.0894             |
| $\beta_{xt}$            | 0.0362  | 0.0176             | 0.0189            | 0.1367             | 0.1148             | 0.0941             | 0.0512             |

The observed bias is smallest under singly-robust MI-GEE and is larger under the ordinary WGEE and MI-WGEE. The variability obtained from MI-GEE(Aug) is better than that of WGEE and comparable to MI-GEE. The variability under the “hybrid” MI-WGEE is better than that of WGEE (especially for  $\beta_x$  and  $\beta_{xt}$ ). With respect to empirically corrected MSEs that are computed by taking the variability in weights into account, the doubly robust approach MI-GEE(Aug) consistently yields

the smallest values, showing substantial improvement over WGEE and improves even further the already precise estimates obtained under singly-robust MI-GEE.

### **MNAR: Results**

In Section 4.2.1, we defined a second data generating MNAR mechanism (GM II). We consider two particular settings. First, the current outcome,  $y_{ij}$ , in (4.12) is omitted as a predictor from both the dropout and the imputation model. This implies that MAR type models are thus fitted for data with an underlying mechanism that is MNAR. Moreover, whereas WGEE and MI-GEE are singly misspecified, since only one component is incorrect, the DR-GEEs are actually “doubly” misspecified, in the sense that two of its components – dropout and imputation model – are incorrect. A second setting, with more extreme misspecification, is further examined. In addition to the current outcome,  $y_{ij}$ , the previous outcome,  $y_{i,j-1}$ , in (4.12) is also omitted from both the dropout and imputation models. This represents a more grossly misspecified scenario, since, essentially, MCAR models are employed for MNAR type data. The results are now discussed in turn. The results for the first setting, in which only the current outcome is omitted as a predictor from both the dropout and the imputation model, are summarized in Table 4.5.

It can be observed that the bias for WGEE is substantially larger compared to MI-GEE, with slightly smaller empirically corrected MSEs for the latter. This underscores, once again, the inefficiency of WGEE under misspecification in the dropout model. While MI-WGEE shows the worst results, the doubly robust MI-GEE(Aug) generally shows improvement over, or is at least comparable with, the singly robust MI-GEE. Moreover, for the doubly robust MI-GEE(Aug), the empirically corrected MSEs that adjust for the variability in the weights are closer to the Monte Carlo MSEs.



Table 4.5: *Simulation study for MNAR (GM II): Singly misspecified dropout and imputation models. (Misspecification in the form of omission of the current outcome,  $y_{ij}$ , from both the dropout and imputation models).*

| WGEE                    |         |                    |            |                    |             |                 |             |
|-------------------------|---------|--------------------|------------|--------------------|-------------|-----------------|-------------|
| Par.                    | Bias    | Var <sub>EST</sub> | $MSE_{MC}$ | No var. in weights |             | Var. in weights |             |
|                         |         |                    |            | $MSE_{emp}$        | $MSE_{mod}$ | $MSE_{emp}$     | $MSE_{mod}$ |
| $\beta_0$               | -0.2588 | 0.0905             | 0.1574     | 0.3587             | 0.1921      | 0.2610          | 0.1024      |
| $\beta_t$               | 0.2209  | 0.0212             | 0.0700     | 0.1860             | 0.1185      | 0.1430          | 0.0649      |
| $\beta_x$               | 0.1029  | 0.1816             | 0.1921     | 0.4255             | 0.1909      | 0.2901          | 0.0686      |
| $\beta_{xt}$            | -0.0816 | 0.0435             | 0.0502     | 0.2073             | 0.1092      | 0.1459          | 0.0336      |
| MI-GEE                  |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.0952 | 0.0417             | 0.0508     | 0.2044             | 0.2063      |                 |             |
| $\beta_t$               | 0.1424  | 0.0103             | 0.0306     | 0.1159             | 0.1169      |                 |             |
| $\beta_x$               | -0.0069 | 0.0735             | 0.0735     | 0.2803             | 0.2832      |                 |             |
| $\beta_{xt}$            | -0.0276 | 0.0188             | 0.0195     | 0.1404             | 0.1423      |                 |             |
| MI-WGEE (Non-Augmented) |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.1777 | 0.0759             | 0.1075     | 0.3502             | 0.2211      | 0.2752          | 0.1879      |
| $\beta_t$               | 0.1684  | 0.0163             | 0.0447     | 0.1984             | 0.1510      | 0.1699          | 0.1403      |
| $\beta_x$               | 0.0348  | 0.1581             | 0.1593     | 0.4624             | 0.2756      | 0.3536          | 0.2299      |
| $\beta_{xt}$            | -0.0363 | 0.0335             | 0.0348     | 0.2511             | 0.1818      | 0.2095          | 0.1666      |
| MI-GEE (Augmented)      |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.4306 | 0.0472             | 0.2326     | 0.3800             | 0.3451      | 0.3127          | 0.2327      |
| $\beta_t$               | 0.2390  | 0.0111             | 0.0682     | 0.1502             | 0.1340      | 0.1195          | 0.0899      |
| $\beta_x$               | 0.0214  | 0.0754             | 0.0759     | 0.2781             | 0.2316      | 0.1879          | 0.0692      |
| $\beta_{xt}$            | -0.0454 | 0.0191             | 0.0211     | 0.1373             | 0.1160      | 0.0956          | 0.0499      |

Under the second setting for the MNAR case, more severe misspecification in both the dropout and imputation models (i.e., omission of both the current *and* the previous outcome) was considered, the results of which are presented in Table 4.6.

Table 4.6: *Simulation study for MNAR (GM II): Doubly misspecified dropout and imputation models. (Misspecification in the form of omission of the current,  $y_{ij}$ , and previous,  $y_{i,j-1}$ , outcomes from both the dropout and imputation models.)*

| WGEE                    |         |                    |            |                    |             |                 |             |
|-------------------------|---------|--------------------|------------|--------------------|-------------|-----------------|-------------|
| Par.                    | Bias    | Var <sub>EST</sub> | $MSE_{MC}$ | No var. in weights |             | Var. in weights |             |
|                         |         |                    |            | $MSE_{emp}$        | $MSE_{mod}$ | $MSE_{emp}$     | $MSE_{mod}$ |
| $\beta_0$               | -0.6604 | 0.0691             | 0.5052     | 0.6950             | 0.5491      | 0.6366          | 0.4650      |
| $\beta_t$               | 0.2743  | 0.0162             | 0.0914     | 0.1947             | 0.1379      | 0.1674          | 0.0881      |
| $\beta_x$               | 0.1853  | 0.1457             | 0.1801     | 0.4150             | 0.2001      | 0.3333          | 0.0834      |
| $\beta_{xt}$            | -0.1195 | 0.0350             | 0.0493     | 0.2002             | 0.1103      | 0.1626          | 0.0369      |
| MI-GEE                  |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.1042 | 0.0409             | 0.0518     | 0.2057             | 0.2088      |                 |             |
| $\beta_t$               | 0.1561  | 0.0101             | 0.0344     | 0.1199             | 0.1214      |                 |             |
| $\beta_x$               | 0.0036  | 0.0717             | 0.0717     | 0.2799             | 0.2849      |                 |             |
| $\beta_{xt}$            | -0.0359 | 0.0186             | 0.0199     | 0.1410             | 0.1441      |                 |             |
| MI-WGEE (Non-Augmented) |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.6370 | 0.0640             | 0.4698     | 0.6965             | 0.5878      | 0.6531          | 0.5587      |
| $\beta_t$               | 0.3078  | 0.0154             | 0.1101     | 0.2526             | 0.2118      | 0.2336          | 0.2024      |
| $\beta_x$               | 0.1151  | 0.1208             | 0.1340     | 0.4331             | 0.2745      | 0.3729          | 0.2341      |
| $\beta_{xt}$            | -0.0769 | 0.0285             | 0.0344     | 0.2387             | 0.1786      | 0.2125          | 0.1651      |
| MI-GEE (Augmented)      |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.0015 | 0.0408             | 0.0408     | 0.1927             | 0.1614      | 0.1236          | 0.0467      |
| $\beta_t$               | 0.1283  | 0.0099             | 0.0264     | 0.1082             | 0.0947      | 0.0785          | 0.0491      |
| $\beta_x$               | -0.0101 | 0.0717             | 0.0718     | 0.2763             | 0.2313      | 0.1782          | 0.0680      |
| $\beta_{xt}$            | -0.0327 | 0.0183             | 0.0194     | 0.1348             | 0.1154      | 0.0918          | 0.0487      |

In the first two panels of the table, the results for WGEE and MI-GEE are also

shown. Note that for these two methods, misspecification occurs in only one component – either in the dropout model for WGEE or in the imputation model for MI-GEE, while for the DR-GEEs, both components are incorrectly specified. MI-GEE yields much less biased estimates than WGEE, and these improve even further under MI-GEE(Aug), despite the doubly, grossly misspecified nature of the latter. Though the doubly robust MI-WGEE shows slight improvement over WGEE in terms of bias, its empirically corrected MSEs tend to be worse. It can also be observed that the empirically corrected MSEs for MI-GEE are much better than those for WGEE, and even further reduction can be seen under MI-GEE(Aug). The latter observation underscores the doubly robust nature of MI-GEE(Aug) in the sense that, despite being misspecified in its two components, it nevertheless does better than the (singly-robust) MI-GEE. For all the methods, accounting for the variability in the weights brings the empirically corrected MSEs closer to the Monte Carlo MSEs.

Two additional settings for the MNAR case are further examined here. In either case, one model is misspecified (i.e., by omission of the current outcome as predictor) and the other model is grossly misspecified (i.e., by omission of both the current and previous outcomes as predictors). The first setting considers a double misspecification in the imputation model, by omission of both the current and previous outcomes as predictors, but only a single misspecification in the dropout model, by omission only of the current outcome as predictor. Thus, though both the imputation and dropout models are incorrectly specified, a more severe misspecification is actually made in the imputation model. The results for this comparison are shown in Table 4.7.

MI-GEE yields less biased estimates than the DR-GEE. Despite an additional misspecification in the dropout model, the doubly robust GEE (MI-GEE(Aug)), provide MSEs that better than or comparable to that of MI-GEE (except for  $\beta_0$ ).

Table 4.7: *Simulation study for MNAR (GM II): Singly misspecified dropout model and doubly misspecified imputation model. (Misspecification in the dropout model in the form of omission of the current outcome,  $y_{ij}$ ; misspecification in the imputation model in the form of omission of the current,  $y_{ij}$ , and previous,  $y_{i,j-1}$ , outcomes).*

| MI-GEE                  |         |                    |            |                    |             |                 |             |
|-------------------------|---------|--------------------|------------|--------------------|-------------|-----------------|-------------|
| Par.                    | Bias    | Var <sub>EST</sub> | $MSE_{MC}$ | No var. in weights |             | Var. in weights |             |
|                         |         |                    |            | $MSE_{emp}$        | $MSE_{mod}$ | $MSE_{emp}$     | $MSE_{mod}$ |
| $\beta_0$               | -0.1042 | 0.0409             | 0.0518     | 0.2057             | 0.2088      |                 |             |
| $\beta_t$               | 0.1561  | 0.0101             | 0.0344     | 0.1199             | 0.1214      |                 |             |
| $\beta_x$               | 0.0036  | 0.0717             | 0.0717     | 0.2799             | 0.2849      |                 |             |
| $\beta_{xt}$            | -0.0359 | 0.0186             | 0.0199     | 0.1410             | 0.1441      |                 |             |
| MI-WGEE (Non-Augmented) |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.1968 | 0.0754             | 0.1141     | 0.3538             | 0.2277      | 0.2811          | 0.1936      |
| $\beta_t$               | 0.1845  | 0.0165             | 0.0505     | 0.2027             | 0.1561      | 0.1747          | 0.1450      |
| $\beta_x$               | 0.0707  | 0.1501             | 0.1551     | 0.4633             | 0.2813      | 0.3574          | 0.2351      |
| $\beta_{xt}$            | -0.0612 | 0.0317             | 0.0355     | 0.2526             | 0.1855      | 0.2126          | 0.1697      |
| MI-GEE (Augmented)      |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.4401 | 0.0464             | 0.2401     | 0.3879             | 0.3536      | 0.3195          | 0.2401      |
| $\beta_t$               | 0.2510  | 0.0107             | 0.0737     | 0.1560             | 0.1401      | 0.1248          | 0.0951      |
| $\beta_x$               | 0.0281  | 0.0755             | 0.0763     | 0.2779             | 0.2325      | 0.1866          | 0.0695      |
| $\beta_{xt}$            | -0.0514 | 0.0189             | 0.0216     | 0.1376             | 0.1170      | 0.0956          | 0.0505      |

The reverse misspecification was also considered. We omit the current outcome from the imputation model, and misspecify, more grossly, the dropout model by omitting both the current and previous outcomes as predictors. Table 4.8, which presents the results for this scenario, highlights the extreme inefficiency of WGEE under misspecification of the dropout model, as can be seen from the relative large empirically corrected MSEs. Not surprisingly, the doubly robust MI-GEE(Aug) not only shows significant reduction in bias as compared to WGEE, but also substantial gains in precision. Again, the hybrid MI-WGEE approach seems slightly worse than ordinary WGEE, since the former involves an additional misspecification in comparison with the latter.

Table 4.8: *Simulation study for MNAR (GM II): Singly misspecified imputation model and doubly misspecified dropout model. (Misspecification in the imputation model in the form of omission of the current outcome,  $y_{ij}$ ; misspecification in the dropout model in the form of omission of the current,  $y_{ij}$ , and previous,  $y_{i,j-1}$ , outcomes).*

| WGEE                    |         |                    |            |                    |             |                 |             |
|-------------------------|---------|--------------------|------------|--------------------|-------------|-----------------|-------------|
| Par.                    | Bias    | Var <sub>EST</sub> | $MSE_{MC}$ | No var. in weights |             | Var. in weights |             |
|                         |         |                    |            | $MSE_{emp}$        | $MSE_{mod}$ | $MSE_{emp}$     | $MSE_{mod}$ |
| $\beta_0$               | -0.6604 | 0.0691             | 0.5052     | 0.6950             | 0.5491      | 0.6366          | 0.4650      |
| $\beta_t$               | 0.2743  | 0.0162             | 0.0914     | 0.1947             | 0.1379      | 0.1674          | 0.0881      |
| $\beta_x$               | 0.1853  | 0.1457             | 0.1801     | 0.4150             | 0.2001      | 0.3333          | 0.0834      |
| $\beta_{xt}$            | -0.1195 | 0.0350             | 0.0493     | 0.2002             | 0.1103      | 0.1626          | 0.0369      |
| MI-WGEE (Non-Augmented) |         |                    |            |                    |             |                 |             |
| $\beta_0$               | -0.6091 | 0.0655             | 0.4365     | 0.6652             | 0.5537      | 0.6197          | 0.5252      |
| $\beta_t$               | 0.2698  | 0.0150             | 0.0878     | 0.2304             | 0.1892      | 0.2109          | 0.1800      |
| $\beta_x$               | 0.0811  | 0.1306             | 0.1371     | 0.4318             | 0.2673      | 0.3684          | 0.2268      |
| $\beta_{xt}$            | -0.0514 | 0.0306             | 0.0332     | 0.2359             | 0.1731      | 0.2084          | 0.1594      |
| MI-GEE (Augmented)      |         |                    |            |                    |             |                 |             |
| $\beta_0$               | 0.0224  | 0.0420             | 0.0425     | 0.1945             | 0.1623      | 0.1300          | 0.0499      |
| $\beta_t$               | 0.1052  | 0.0104             | 0.0215     | 0.1037             | 0.0898      | 0.0767          | 0.0457      |
| $\beta_x$               | -0.0209 | 0.0725             | 0.0729     | 0.2778             | 0.2313      | 0.1844          | 0.0688      |
| $\beta_{xt}$            | -0.0234 | 0.0187             | 0.0193     | 0.1350             | 0.1148      | 0.0940          | 0.0486      |

### 4.3 Analysis of the Toenail Data

The data set was introduced in Section 2.1. Here the response of interest is the severity of the infection, coded as 0 (not severe) or 1 (severe). The question of interest was whether the percentage of severe infections decreased over time, and whether that evolution was different for the two treatment groups. A graphical representation of the number of patients in the study at each time-point, and the number of patients with severe infections is given in Figure 4.1.

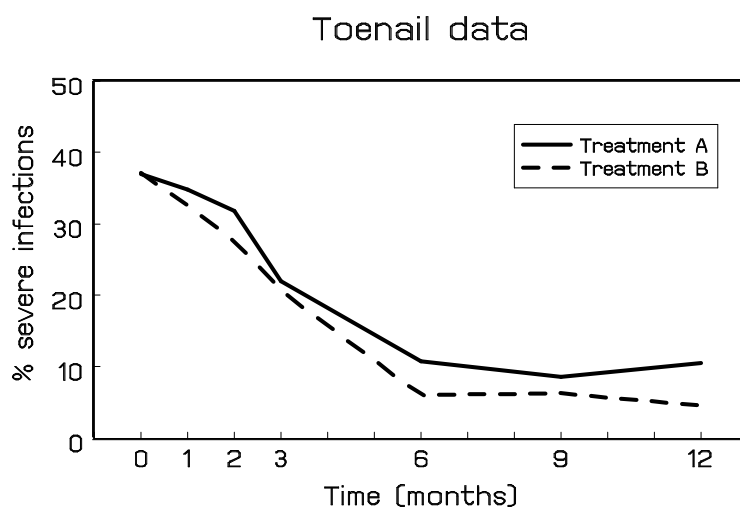


Figure 4.1: *Toenail Data. Evolution of the percentage of severe toenail infections in the two treatment groups separately. (Source: Molenberghs and Verbeke, 2005, p.10).*

Our case study, geared towards assessing the difference in improvement rate between both treatments for onychomycosis, allows us to further empirically assess the behavior of the various GEE versions that we consider. Results are summarized in Table 4.9. For the main covariate of interest, i.e., treatment, all the considered models result in insignificant difference between the treatment groups.

Table 4.9: *Toenail Data. Severity of toenail infection. Parameter estimates (empirically corrected standard errors ignoring variability in the weights; empirically corrected standard errors acknowledging variability in the weights).*

| Effect    | WGEE                | MI-GEE        | MI-WGEE              | MI-WGEE(Aug)        |
|-----------|---------------------|---------------|----------------------|---------------------|
| Intercept | -0.322(0.309;0.220) | -0.059(0.224) | -0.679(0.395; 0.249) | -0.936(0.448;0.430) |
| Time      | 0.820(0.641;0.305)  | 0.017(0.322)  | 0.874(0.708; 0.327)  | 0.028(0.145;0.143)  |
| Trt       | -0.219(0.073;0.045) | -0.313(0.047) | -0.070(0.087; 0.059) | 0.236(0.323;0.271)  |
| Trt*time  | -0.124(0.183;0.059) | -0.051(0.068) | -0.134(0.140; 0.065) | -0.105(0.085;0.077) |

That said, a few additional observations can be made. First, taking the variability in the weights into account does not necessarily lead to smaller standard errors, even though the difference with the standard errors not accommodating this variability is evident. Second, the point estimates do vary considerably between the different methods. This, combined with insight from the simulation study, may lead us to conclude that it is prudent to consider a doubly robust method, perhaps in conjunction with multiple imputation, for ease of use and good performance.

## 4.4 Discussion

In this chapter, we have presented a variety of versions of generalized estimating equations for use when data are incomplete and of an MAR nature. These are based on the principles of inverse probability weighting, doubly robust extensions thereof, and multiple imputation. We have paid particular attention to some of the combinations, such as multiple imputation combined with inverse probability weighting, and its doubly robust counterpart. The latter is due to Daniel (2008). We consider both a principled version, owing to this author, and an approximation that facilitates computation.

Oftentimes, when a weighting scheme is considered, one fits the prescribed model using *estimated* weights, rather than the true weights. Because such a substitution can inherently affect the estimated precision, we have proposed a modification in the variance estimation that accounts for the variability in the (estimated) weights. While acknowledging variability in the weights seems entirely reasonable to do, the user should be aware that this also renders the method more computationally demanding, in the sense that these adjusted variance estimates are not readily available in standard software implementations of WGEE. Nevertheless, taking the variability into account is definitely an advantage, particularly since the weight model used for analysis is frequently just an estimate of the true underlying weight model.

An additional point regarding inverse probability weighting methods is that their validity requires the so-called *positivity assumption* in the sense that probabilities for being observed should be bounded away from zero, Rotnitzky (2009). Cases where this assumption is at stake would be of interest in their own right; however, they are outside of the scope of the current (simulation) study.

From an extensive small-sample simulation study it emerges that WGEE is rather inefficient, especially for small sample sizes, and was observed to be sensitive to misspecification in the dropout model, which is not surprising. In contrast, though, MI-GEE is relatively robust to misspecification in the imputation model, even though such model is at the heart of multiple imputation. Often, it was also observed that the singly robust MI-GEE yielded fairly precise estimates, with slight improvement gained under the doubly robust MI-GEE(Aug).

In the MAR case, when the model is fully correctly specified, MI-GEE and MI-GEE(Aug) show better performance than the approaches that employ WGEE (e.g., ordinary WGEE and the “hybrid” MI-WGEE), demonstrating the inefficiency of inverse probability weighting despite a correct specification of the weight model. Moreover, the inefficiency tends to be even grosser when small sample sizes are in play.

When the dropout and measurement models are correctly specified and the imputation model is not, either MI-GEE or MI-GEE(Aug) show better results, depending on the parameter under consideration. In contrast, when the dropout model is incorrectly specified but the rest is correct, then the doubly robust MI-GEE(Aug) yields substantive improvement over WGEE – a clear manifestation of the double robustness of the former. Finally, in the situation wherein both the imputation and dropout models are incorrect, MI-GEE(Aug) clearly outperforms all the other methods, providing considerable improvement over WGEE and slightly increased precision over MI-GEE.

For the MNAR case, inefficiency of WGEE was again observed, especially under severe misspecification of the dropout model. MI-GEE exhibited a reasonable amount of robustness with respect to misspecification – severe or otherwise – in the imputation model. MI-GEE(Aug) was seen to be considerably robust, doing better than all the other methods, despite severe or not-so-severe misspecification in both of its components.

Overall, the multiple imputation based methods are recommendable for practice and it is advisable to take variability in the weights into account.



## Chapter 5

# Pseudo-likelihood Estimation for Incomplete Data

Pseudo-likelihood approaches have become a practical alternative to full likelihood methods, particularly for applications involving complex likelihood forms. In the case of incomplete data, Pseudo-likelihood (PL) is valid under the assumption of an MCAR mechanism operating, but this does not generally extend to MAR mechanisms, except in a limited number of special cases, such as full exchangeability, as will be shown in Section 5.2. The reason for this is twofold. First, in line with Kenward and Molenberghs (2009), even likelihood methods commonly have frequentist elements, such as the expected information matrix. Second, because pseudo-likelihood is not a genuine likelihood but rather a modification of it, it no longer enjoys the results derived for the likelihood by Rubin (1976). This second issue is shared with GEE.

Unlike for GEE, little work has been done for PL estimation with incomplete data. A noteworthy exception is Parzen *et al.* (2006) who apply PL ideas not just to the vector of outcomes, but to the entire vector of outcomes, covariates, and missing data indicators. In what follows, a different route, using inverse probability weighting and double robustness ideas (Scharfstein, Rotnitzky and Robins, 1999; Van der Laan and Robins, 2003; Bang and Robins, 2005; Rotnitzky, 2009) is followed.

In Section 5.1, general expressions are presented and their validity established. These expressions are then applied to pseudo-likelihood (Section 5.2). This implies that they hold, beyond pseudo-likelihood, to a large class of estimating equations, in line with the work of Robins, Rotnitzky, and colleagues. Thereafter, we pay particular attention to two special PL families: (1) marginal and (2) full conditional. In the first case, marginal pseudo-likelihood for Gaussian (Section 5.2.2) is considered in more

detail. In the second case, an exponential family model for binary clustered data is scrutinized further (Section 5.2.3). The proposed methods are applied to two case studies in Section 5.3. Marginal models for binary data will be the main focus of Chapter 6.

## 5.1 General Forms of Estimating Equations for Incomplete Data

Assume that we have a set of estimating equations, whether resulting from full likelihood or pseudo-likelihood, of a conventional generalized estimating equations type, or beyond:

$$\mathbf{U} = \sum_{i=1}^N \mathbf{U}_i(\boldsymbol{\theta} | \mathbf{Y}_i, \mathbf{x}_i) \stackrel{\text{notation}}{=} \sum_{i=1}^N \mathbf{U}_i(\mathbf{Y}_i). \quad (5.1)$$

Assume that  $E(\mathbf{U}) = \mathbf{0}$ .

Let us first consider two obvious ‘naive’ estimating equations, originating from (5.1):

$$\mathbf{U}_{\text{naive, CC}} = \sum_{i=1}^N \tilde{R}_i \mathbf{U}_i(\mathbf{Y}_i), \quad (5.2)$$

$$\mathbf{U}_{\text{naive, AC}} = \sum_{i=1}^N \mathbf{U}_i(\mathbf{Y}_i^o). \quad (5.3)$$

Here,  $\tilde{R}_i = 1$  if subject  $i$  is fully observed and 0 otherwise, and  $\mathbf{U}_i(\mathbf{Y}_i^o)$  is the score pertaining the observed outcomes on subject  $i$ . Further, ‘naive’ refers to the fact that these estimating equations would generally be biased under MAR; ‘CC’ denotes complete cases, i.e., subjects with all measurements taken; and ‘AC’ stands for available cases. For the latter, it is necessary to derive the score contribution of the sub-vector of observed components of  $\mathbf{Y}_i$ . Because this involves integration over the incomplete data, it is trivial in the marginal case, but less so, for example, for conditionally specified PL functions.

Singly robust versions of (5.2) and (5.3) would take the form:

$$\mathbf{U}_{\text{IPWCC}} = \sum_{i=1}^N \frac{\tilde{R}_i}{\pi_i} \mathbf{U}_i(\mathbf{Y}_i), \quad (5.4)$$

$$\mathbf{U}_{\text{IPWAC}} = \sum_{i=1}^N \frac{1}{\pi'_i} \cdot E_{Y^m|y^o} \mathbf{U}(Y_i), \quad (5.5)$$

$$\mathbf{U}_{\text{IPWAC,seq}} = \sum_{i=1}^N \sum_{j=1}^{n_i} \frac{R_{ij}}{\pi_{ij}} \cdot \mathbf{U}(Y_{ij}|\mathbf{Y}_{i\bar{j}}). \quad (5.6)$$

Here  $R_{ij}$  is the indicator for a subject to be observed at occasion  $j$  and  $\pi_{ij}$  is the probability of being observed up until and including occasion  $j$ , i.e.,  $\pi_{ij} = \prod_{\ell=1}^j (1 - p_{i\ell})$ . Further,  $\mathbf{Y}_{i\bar{j}}$  is shorthand for the history  $(Y_{i1}, \dots, Y_{i,j-1})$ , and the corresponding function  $\mathbf{U}(Y_{ij}|\mathbf{Y}_{i\bar{j}})$  is the score for the outcome at occasion  $j$  given the history. Recall that  $\pi_i$  and  $\pi'_i$  have been defined in (4.1) and (4.2) respectively.

Doubly robust versions are:

$$\mathbf{U}_{\text{IPWCC,dr}} = \sum_{i=1}^N \left[ \frac{\tilde{R}_i}{\pi_i} \mathbf{U}_i(\mathbf{Y}_i) + \left(1 - \frac{\tilde{R}_i}{\pi_i}\right) E_{Y_i^m|y_i^o} \mathbf{U}_i(\mathbf{Y}_i) \right], \quad (5.7)$$

$$\begin{aligned} \mathbf{U}_{\text{IPWAC,dr}} = & \sum_{i=1}^N \left\{ \sum_{j=1}^{n_i} \left[ \frac{R_{ij}}{\pi_{ij}} \cdot \mathbf{U}(Y_{ij}|\mathbf{Y}_{i\bar{j}}) \right. \right. \\ & \left. \left. + \left(1 - \frac{R_{ij}}{\pi_{ij}}\right) \cdot E_{Y_{ij}^m|y_{ij}^o} \mathbf{U}(Y_{ij}|\mathbf{Y}_{i\bar{j}}) \right] \right\}. \quad (5.8) \end{aligned}$$

Here  $\mathbf{U}(Y_{ij}|\mathbf{Y}_{i\bar{j}})$  is the score pertaining to outcome  $Y_{ij}$  given the history, denoted by  $\mathbf{Y}_{i\bar{j}}$ . We are now in a position to establish the single and double robustness of the above definitions.

**Theorem 2 (Single robustness of  $\mathbf{U}_{\text{IPWCC}}$ ,  $\mathbf{U}_{\text{IPWAC}}$ , and  $\mathbf{U}_{\text{IPWAC,seq}}$ .)**

*Under MAR, and if  $p_{i\ell}$  in (4.1)–(4.2) is non-parametrically or correctly parametrically specified as  $p_{i\ell}(\boldsymbol{\psi})$ , then  $\mathbf{U}_{\text{IPWCC}}$ ,  $\mathbf{U}_{\text{IPWAC}}$ , and  $\mathbf{U}_{\text{IPWAC,seq}}$  are consistent.*

**Proof.** This follows from their expectation being 0, as follows:

$$\begin{aligned}
E(\mathbf{U}_{\text{IPWCC}}) &= E_Y \left\{ \sum_{i=1}^N E_{R|Y} \left[ \frac{\tilde{R}_i}{\pi_i} \mathbf{U}(\mathbf{Y}_i) \right] \right\} \\
&= E_Y \left\{ \sum_{i=1}^N \left[ \frac{E_{R|Y}(\tilde{R}_i)}{\pi_i} \mathbf{U}(\mathbf{Y}_i) \right] \right\} \\
&= E_Y \left[ \sum_{i=1}^N \mathbf{U}_i(\mathbf{Y}_i) \right] = \mathbf{0}. \tag{5.9}
\end{aligned}$$

$$\begin{aligned}
E(\mathbf{U}_{\text{IPWAC}}) &= E_Y \left\{ \sum_{i=1}^N E_{R|Y} \left[ \frac{R'_i}{\pi'_i} E_{Y^m|y^o} \mathbf{U}(\mathbf{Y}_i) \right] \right\} \\
&= E_Y \left\{ \sum_{i=1}^N \left[ \frac{E_{R|Y}(R'_i)}{\pi'_i} E_{Y^m|y^o} \mathbf{U}(\mathbf{Y}_i) \right] \right\} \\
&= \sum_{i=1}^N E_Y E_{Y^m|y^o} \mathbf{U}(\mathbf{Y}_i) = E_Y \left[ \sum_{i=1}^N \mathbf{U}_i(\mathbf{Y}_i) \right] = \mathbf{0}. \tag{5.10}
\end{aligned}$$

$$\begin{aligned}
E(\mathbf{U}_{\text{IPWAC,seq}}) &= E_Y \left\{ \sum_{i=1}^N E_{R|Y} \left[ \sum_{j=1}^{n_i} \frac{R_{ij}}{\pi_{ij}} E_{Y^m|y^o} \mathbf{U}(Y_{ij} | \mathbf{Y}_{i\bar{j}}) \right] \right\} \\
&= E_Y \left\{ \sum_{i=1}^N \left[ \sum_{j=1}^{n_i} E_{R_j|R_{\bar{j}}Y} \frac{R_{ij}}{\pi_{ij}} E_{Y^m|y^o} \mathbf{U}(Y_{ij} | \mathbf{Y}_{i\bar{j}}) \right] \right\} \\
&= E_Y \left[ \sum_{i=1}^N \sum_{j=1}^{n_i} \mathbf{U}_i(\mathbf{Y}_i) \right] = \mathbf{0}. \tag{5.11}
\end{aligned}$$

Here,  $R'_i = 1$  if a subject drops at occasion  $d_i$  and 0 otherwise. Note that, in the CC case, we used  $E_{R|Y}(R_i) = E_{R|Y^o}(R_i) = \pi_i$ , owing to MAR. A similar statement holds in the AC case. This completes the proof.

In the above, and also in what follows, the same regularity conditions apply as in Rotnitzky (2009). In particular, it is important that the probability of being observed for a measurement be bounded away from zero.

**Theorem 3 (Double robustness of  $U_{\text{IPWCC,dr}}$  and  $U_{\text{IPWAC,dr}}$ .)** *Under MAR, and (a) if  $p_{i\ell}$  in (4.1)–(4.2) is non-parametrically or correctly parametrically specified as  $p_{i\ell}(\psi)$  and/or (b) if the predictive models in (5.7) and (5.8) are correctly specified, then  $U_{\text{IPWCC,dr}}$  and  $U_{\text{IPWAC,dr}}$  are consistent.*

**Proof.** If condition (a) holds, then the result trivially follows from Theorem 2 and the observation that the expectation of the first factors of the second terms on the

right hand sides equal zero. Under condition (b), write  $E_{R|Y}(R_i) = E_{R|Y^o}(R_i) = \lambda_i$ . Then,

$$\begin{aligned}
E(\mathbf{U}_{\text{IPWCC,dr}}) &= E_Y \left\{ \sum_{i=1}^N \left[ \frac{\lambda_i}{\pi_i} \mathbf{U}_i(\mathbf{Y}_i) + \left(1 - \frac{\lambda_i}{\pi_i}\right) E_{Y_i^m|y_i^o} \mathbf{U}_i(\mathbf{Y}_i) \right] \right\} \\
&= \sum_{i=1}^N \left\{ \frac{\lambda_i}{\pi_i} E_{Y^o} E_{Y^m|Y^o} [\mathbf{U}_i(\mathbf{Y}_i)] \right. \\
&\quad \left. + \left(1 - \frac{\lambda_i}{\pi_i}\right) E_{Y^o} E_{Y^m|Y^o} [E_{Y_i^m|y_i^o} \mathbf{U}_i(\mathbf{Y}_i)] \right\} \\
&= \sum_{i=1}^N E_{Y^o} E_{Y^m|Y^o} [\mathbf{U}_i(\mathbf{Y}_i)] = \sum_{i=1}^N E_Y [\mathbf{U}_i(\mathbf{Y}_i)] = \mathbf{0}. \quad (5.12)
\end{aligned}$$

The AC case starts with similar logic for the case condition (a) holds. When (b) holds, but not necessarily (a):

$$\begin{aligned}
E(\mathbf{U}_{\text{IPWAC,dr}}) &= E_Y \left\{ \sum_{i=1}^N \left[ \sum_{j=1}^{n_i} \frac{\lambda_{ij}}{\pi_{ij}} \mathbf{U}_i(Y_i | \mathbf{Y}_{i\bar{j}}) \right. \right. \\
&\quad \left. \left. + \left(1 - \frac{\lambda_{ij}}{\pi_{ij}}\right) E_{Y_i^m|y_i^o} \mathbf{U}_i(Y_i | \mathbf{Y}_{i\bar{j}}) \right] \right\} \\
&= \sum_{i=1}^N \sum_{j=1}^{n_i} \left\{ \frac{\lambda_{ij}}{\pi_{ij}} E_{Y^o} E_{Y^m|Y^o} [\mathbf{U}_i(Y_i | \mathbf{Y}_{i\bar{j}})] \right. \\
&\quad \left. + \left(1 - \frac{\lambda_{ij}}{\pi_{ij}}\right) E_{Y^o} E_{Y^m|Y^o} [E_{Y_i^m|y_i^o} \mathbf{U}_i(Y_i | \mathbf{Y}_{i\bar{j}})] \right\} \\
&= \sum_{i=1}^N \sum_{j=1}^{n_i} E_{Y^o} E_{Y^m|Y^o} [\mathbf{U}_i(Y_i | \mathbf{Y}_{i\bar{j}})] \\
&= \sum_{i=1}^N E_Y [\mathbf{U}_i(\mathbf{Y}_i)] = \mathbf{0}. \quad (5.13)
\end{aligned}$$

This completes the proof.

As mentioned in Section 4.1.4, the predictive model may show varying degrees of complexity, depending on the type of PL function considered. For example, marginal models for continuous data, marginal models for binary data, and conditional models for binary data, may all pose specific challenges. This means that, in some settings, the predictive model might be of higher dimension than the components of the actual PL function and/or contain components of the full likelihood that are not needed for it. While this may seem to defeat the purpose of using PL methodology, there are several practically useful strategies to handle this. To see this, it helps to distinguish

between two uses of the likelihood within the framework: (1) estimation; and (2) prediction. It is predominantly for estimation that PL leads to important economies by not having to manipulate the full likelihood. For prediction, several alternative strategies are available.

First, even though using the entire joint distribution is oftentimes prohibitive for estimation, it may be tractable for prediction purposes, provided all the necessary parameters are obtained from the likelihood. An example is provided by a full conditional PL, with a counterexample being a purely marginal PL for binary data, consisting of lower-order margins only. Second, sufficiently rich predictive models and alternative approaches, as discussed in Section 4.1.4, could be used. We return to this in Section 5.2.

### Precision Estimation

A general expression for the precision of the estimates obtained using pseudo-likelihood is given in Theorem 1 of Chapter 3. If estimation of the dropout model parameters is not involved, then  $I_0$  and  $I_1$  are as in (3.18) and (3.19) respectively. When the singly robust and doubly robust versions of the pseudo-likelihood estimation are used, the variability induced by estimation of the  $\boldsymbol{\psi}$  parameters needs to be accommodated. The asymptotic variance-covariance matrix is estimated by  $\widehat{I}_0^{-1} \widehat{I}_1 \widehat{I}_0^{-1}$  where  $I_0$  and  $I_1$  are as in (4.8) and (4.9) respectively. While the  $\boldsymbol{U}$  in  $\boldsymbol{U} = \sum_{i=1}^N \boldsymbol{V}_i(\boldsymbol{\beta})$  (Section 4.1.4) comes from the score equations, here the  $\boldsymbol{U}$  are from the (pairwise) pseudo-likelihood.

## 5.2 Pseudo-likelihood Estimating Equations

In the previous section, we focused on estimating equations in the broadest sense. When we turn to pseudo-likelihood, the generic forms can be made more specific and expanded further:

$$\boldsymbol{U}_{\text{naive, CC}} = \sum_{i=1}^N R_i \sum_{s \in S} \delta_s \boldsymbol{U}_s(\boldsymbol{y}_i^{(s)}), \quad (5.14)$$

$$\boldsymbol{U}_{\text{naive, CS}} = \sum_{i=1}^N \sum_{s \in S} R_{i,s} \delta_s \boldsymbol{U}_s(\boldsymbol{y}_i^{(s)}), \quad (5.15)$$

$$\mathbf{U}_{\text{naive, AC}} = \sum_{i=1}^N \sum_{s \in S} \delta_s E_{\mathbf{Y}^m | \mathbf{y}^o} \mathbf{U}_s(\mathbf{y}_i^{(s)}), \quad (5.16)$$

$$\mathbf{U}_{\text{IPWCC}} = \sum_{i=1}^N \frac{R_i}{\pi_i} \cdot \sum_{s \in S} \delta_s \mathbf{U}_s(\mathbf{y}_i^{(s)}), \quad (5.17)$$

$$\mathbf{U}_{\text{IPWCS}} = \sum_{i=1}^N \sum_{s \in S} \frac{R_{i,s}}{\pi_{i,s}} \cdot \delta_s \mathbf{U}_s(\mathbf{y}_i^{(s)}), \quad (5.18)$$

$$\mathbf{U}_{\text{IPWAC}} = \sum_{i=1}^N \sum_{s \in S} \delta_s \sum_{j=1}^{n_i} I(j \in s) \cdot \frac{R_{ij}}{\pi_{ij}} \cdot \mathbf{U}_s(y_{ij} | \mathbf{y}_{i\bar{j}}^{(s)}), \quad (5.19)$$

$$\begin{aligned} \mathbf{U}_{\text{IPWCC, dr}} = & \sum_{i=1}^N \left\{ \frac{R_i}{\pi_i} \left[ \sum_{s \in S} \delta_s \mathbf{U}_s(\mathbf{y}_i^{(s)}) \right] \right. \\ & \left. + \left( 1 - \frac{R_i}{\pi_i} \right) \cdot E_{\mathbf{Y}_i^m | \mathbf{y}_i^o} \left[ \sum_{s \in S} \delta_s \mathbf{U}_s(\mathbf{y}_i^{(s)}) \right] \right\}, \end{aligned} \quad (5.20)$$

$$\begin{aligned} \mathbf{U}_{\text{IPWCS, dr}} = & \sum_{i=1}^N \sum_{s \in S} \left\{ \frac{R_{i,s}}{\pi_{i,s}} \cdot \delta_s \mathbf{U}_s(\mathbf{y}_i^{(s)o}) \right. \\ & \left. + \left( 1 - \frac{R_{i,s}}{\pi_{i,s}} \right) \cdot \delta_s E_{\mathbf{Y}_i^m | \mathbf{y}_i^o} \mathbf{U}_s(\mathbf{y}_i^{(s)}) \right\}. \end{aligned} \quad (5.21)$$

$$\begin{aligned} \mathbf{U}_{\text{IPWAC, dr}} = & \sum_{i=1}^N \sum_{s \in S} \delta_s \sum_{j=1}^{n_i} I(j \in s) \left[ \frac{R_{ij}}{\pi_{ij}} \cdot \mathbf{U}_s(y_{ij} | \mathbf{y}_{i\bar{j}}^{(s)}) \right. \\ & \left. + \left( 1 - \frac{R_{ij}}{\pi_{ij}} \right) \cdot E_{\mathbf{Y}_i^m | \mathbf{y}_i^o} \mathbf{U}_s(y_{ij} | \mathbf{Y}_{i\bar{j}}^{(s)}) \right]. \end{aligned} \quad (5.22)$$

where  $R_i$ ,  $\pi_i$ ,  $R_{ij}$ , and  $\pi_{ij}$  retain their former meaning. Similarly,  $R_{i,s}$  and  $\pi_{i,s}$  are the indicator and probability for the sub-vector  $\mathbf{y}_i^{(s)}$  of  $\mathbf{y}_i$  to be observed, respectively. Further, ‘CS’ stands for ‘complete sets’.

When the outcome sequence is fully exchangeable, in the sense that the distribution of any sub-vector of  $\mathbf{Y}_i$  equals that of any other sub-vector of equal length or a permutation thereof, then  $\mathbf{U}_{\text{IPWCS, dr}}$  simplifies considerably. Indeed,

$$E_{\mathbf{Y}_i^m | \mathbf{y}_i^o} \mathbf{U}_s(\mathbf{y}_i^{(s)}) = E_{\mathbf{Y}_i^m | \mathbf{y}_i^o} \left[ \mathbf{U}_s(\mathbf{y}_i^{(s)o}) + \mathbf{U}_s(\mathbf{y}_i^{(s)m} | \mathbf{y}_i^{(s)o}) \right].$$

Now, the expectation over the second term on the right hand side can be replaced by  $E_{\mathbf{Y}_i^{(s)m} | \mathbf{y}_i^{(s)o}} \mathbf{U}_s(\mathbf{y}_i^{(s)m} | \mathbf{y}_i^{(s)o})$ , thanks to full exchangeability and the fact that the score contributions stem from derivatives of sub-vectors of  $\mathbf{y}_i$ . Upon this replacement, the conditional expectation vanishes. As a consequence, under exchangeability there is no need to explicitly model the missing data mechanism. Hence, (5.21) reduces to

$$\mathbf{U}_{\text{IPW, exch}} = \sum_{i=1}^N \sum_{s \in S} \delta_s \mathbf{U}_s(\mathbf{y}_i^{(s)o}), \quad (5.23)$$

Thus, in this special but important case, neither the weights nor the conditional expectations are necessary to obtain valid inferences.

We now focus on special cases of pseudo-likelihood. To begin with, we will consider the case of pairwise pseudo-likelihood, and then apply it to normally distributed data. Thereafter, a conditional pseudo-likelihood for binary outcomes will be entertained.

### 5.2.1 Pairwise (Pseudo-)likelihood

While in principle general missingness could be considered, we focus on the important special case of dropout, to streamline mathematical development. The forms (5.14)–(5.22) take the following form for the specific case of pairwise likelihood:

$$\mathbf{U}_{\text{naive, CC}} = \sum_{i=1}^N \tilde{R}_i \sum_{j < k} \mathbf{U}_i(y_{ij}, y_{ik}), \quad (5.24)$$

$$\mathbf{U}_{\text{naive, CP}} = \sum_{i=1}^N \sum_{j < k < d_i} \mathbf{U}_i(y_{ij}, y_{ik}), \quad (5.25)$$

$$\mathbf{U}_{\text{naive, AC}} = \sum_{i=1}^N \left[ \sum_{j < k < d_i} \mathbf{U}_i(y_{ij}, y_{ik}) + \sum_{j=1}^{d_i-1} (n_i - d_i + 1) \mathbf{U}_i(y_{ij}) \right], \quad (5.26)$$

$$\mathbf{U}_{\text{IPWCC}} = \sum_{i=1}^N \frac{\tilde{R}_i}{\pi_i} \left[ \sum_{j < k} \mathbf{U}_i(y_{ij}, y_{ik}) \right], \quad (5.27)$$

$$\mathbf{U}_{\text{IPWCP}} = \sum_{i=1}^N \sum_{j < k < d_i} \frac{R_{ijk}}{\pi_{ijk}} \cdot \mathbf{U}_i(y_{ij}, y_{ik}), \quad (5.28)$$

$$\mathbf{U}_{\text{IPWAC}} = \sum_{i=1}^N \sum_{j < k} \left[ \frac{R_{ij}}{\pi_{ij}} \cdot \mathbf{U}_i(y_{ij}) + \frac{R_{ik}}{\pi_{ik}} \cdot \mathbf{U}_i(y_{ik} | y_{ij}) \right], \quad (5.29)$$



$$\begin{aligned} \mathbf{U}_{\text{IPWCC,dr}} &= \sum_{i=1}^N \left\{ \frac{\tilde{R}_i}{\pi_i} \left[ \sum_{j < k} \mathbf{U}_i(y_{ij}, y_{ik}) \right] \right. \\ &\quad \left. + \left( 1 - \frac{\tilde{R}_i}{\pi_i} \right) E_{\mathbf{Y}_i^m | \mathbf{y}_i^o} \left[ \sum_{j < k} \mathbf{U}_i(y_{ij}, y_{ik}) \right] \right\}, \end{aligned} \quad (5.30)$$

$$\begin{aligned} \mathbf{U}_{\text{IPWCP,dr}} &= \sum_{i=1}^N \sum_{j < k < n_i} \left[ \frac{R_{ijk}}{\pi_{ijk}} \cdot \mathbf{U}_i(y_{ij}, y_{ik}) \right. \\ &\quad \left. + \left( 1 - \frac{R_{ijk}}{\pi_{ijk}} \right) \cdot E_{\mathbf{Y}_i^m | \mathbf{y}_i^o} \mathbf{U}_i(y_{ij}, y_{ik}) \right], \end{aligned} \quad (5.31)$$

$$\begin{aligned} \mathbf{U}_{\text{IPWAC,dr}} &= \sum_{i=1}^N \sum_{j < k} \left[ \frac{R_{ij}}{\pi_{ij}} \cdot \mathbf{U}_i(y_{ij}) + \frac{R_{ik}}{\pi_{ik}} \cdot \mathbf{U}_i(y_{ik} | y_{ij}) \right. \\ &\quad \left. + \left( 1 - \frac{R_{ij}}{\pi_{ij}} \right) \cdot E_{\mathbf{Y}_i^m | \mathbf{y}_i^o} \mathbf{U}_i(y_{ij}) \right. \\ &\quad \left. + \left( 1 - \frac{R_{ik}}{\pi_{ik}} \right) \cdot E_{\mathbf{Y}_i^m | \mathbf{y}_i^o} \mathbf{U}_i(y_{ik} | y_{ij}) \right]. \end{aligned} \quad (5.32)$$

where  $R_{ijk}$  and  $\pi_{ijk}$  are the indicator and probability, respectively, for observing both  $y_{ij}$  and  $y_{ik}$ . We can now write  $\pi_i = \prod_{\ell=2}^{n_i} (1 - p_{i\ell})$ , where still  $p_{i\ell} = P(D_i = \ell | D_i \geq \ell, Y_{i\bar{\ell}}, X_{i\bar{\ell}})$ . The second term in (5.26) results from all pairs with the first component observed and the second one unobserved.

It is interesting, and easy to show, that all three the doubly robust versions coincide in this case, which adds to their attraction:

$$\begin{aligned} \mathbf{U}_{\text{IPWCC,dr}} &= \mathbf{U}_{\text{IPWCP,dr}} = \mathbf{U}_{\text{IPWAC,dr}} \\ &= \sum_{i=1}^N \left\{ \sum_{j < k < d_i} \mathbf{U}_i(y_{ij}, y_{ik}) + \sum_{j=1}^{d_i-1} (n_i - d_i + 1) \cdot \mathbf{U}_i(y_{ij}) \right. \\ &\quad \left. + \sum_{j < d_i \leq k} E[\mathbf{U}_i(y_{ik} | y_{ij})] + \sum_{d_i \leq j < k} E[\mathbf{U}_i(y_{ij}, y_{ik})] \right\}. \end{aligned} \quad (5.33)$$

A key feature in (5.33) is that the need to model the missing data mechanism is avoided. Note that this expression is related to (5.26) in the sense that both terms of the latter expression occur here as well, with in addition the predictive terms. There are two types of predictive terms, corresponding to: (a) a pair with the first component observed and the second one missing; (b) a pair with both components missing. All predictive models involve two types of contributions: for  $E[\mathbf{U}_i(y_{ik} | y_{ij})]$  where  $y_{ij}$  is observed but  $y_{ik}$  is not, and for  $E[\mathbf{U}_i(y_{ij}, y_{ik})]$  with both unobserved. These will be considered for the special but important cases that follow next.

It is very easy to derive an exchangeable form, starting from (5.32), because then, in this expression, the expectations vanish. Hence, clearly, the exchangeable form is equal to (5.26), making the naive available case version not only valid, but actually doubly robust. Of course, this is the case only under exchangeability.

A very important observation is that in the doubly robust versions (5.33), the need to specify the missing data model is avoided, even though the predictive model for the unobserved outcomes is needed.

### 5.2.2 Marginal (Pairwise) Pseudo-likelihood for Gaussian Data

Assume  $\mathbf{Y}_i \sim N(\boldsymbol{\mu}, \Sigma)$ . Then first, suppressing the index  $i$  from notation, and writing down the expressions for observed values, we find:

$$\begin{aligned} U(y_k|y_j) &= \frac{\partial(\mu_{k|j}, \sigma_{kk|j})}{\partial(\mu_j, \mu_k, \sigma_{jj}, \sigma_{jk}, \sigma_{kk})} \cdot \frac{\partial \ln \phi(y_k|y_j; \mu_{k|j}, \sigma_{kk|j})}{\partial(\mu_{k|j}, \sigma_{kk|j})} \\ &= \begin{pmatrix} -\frac{\sigma_{jk}}{\sigma_{jj}} & 0 \\ 1 & 0 \\ -\frac{\sigma_{jk}}{\sigma_{jj}^2}(y_j - \mu_j) & \frac{\sigma_{jk}^2}{\sigma_{jj}^2} \\ \frac{y_j - \mu_j}{\sigma_{jj}} & -\frac{2\sigma_{jk}}{\sigma_{jj}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{y_k - \mu_{k|j}}{\sigma_{kk|j}} \\ -\frac{1}{2\sigma_{kk|j}} + \frac{1}{2} \frac{(y_k - \mu_{k|j})^2}{\sigma_{kk|j}^2} \end{pmatrix}, \end{aligned} \quad (5.34)$$

where  $\phi(\cdot)$  is the normal density with mean and variance given by:

$$\mu_{k|j} = \mu_k + \frac{\sigma_{jk}}{\sigma_{jj}}(y_j - \mu_j) \quad \text{and} \quad \sigma_{kk|j} = \frac{\sigma_{jj}\sigma_{kk} - \sigma_{jk}^2}{\sigma_{jj}}.$$

The only stochastic elements in (5.34) are the conditional residual and its square. We need to take their expectation conditional upon the observed outcomes, producing for the second factor in (5.34):

$$\begin{pmatrix} \frac{\sigma_{jj}\Sigma_k \bar{\mathbf{d}} \Sigma_{\bar{\mathbf{d}}}^{-1}(\mathbf{y}_{\bar{\mathbf{d}}} - \boldsymbol{\mu}_{\bar{\mathbf{d}}}) - \sigma_{jk}(y_j - \mu_j)}{\sigma_{jj}\sigma_{kk} - \sigma_{jk}^2} \\ \frac{\sigma_{jj}(\sigma_{jk}^2 - \sigma_{jj}\Sigma_k \bar{\mathbf{d}} \Sigma_{\bar{\mathbf{d}}}^{-1} \Sigma_{\bar{\mathbf{d}}} + [\sigma_{jj}\Sigma_k \bar{\mathbf{d}} \Sigma_{\bar{\mathbf{d}}}^{-1}(\mathbf{y}_{\bar{\mathbf{d}}} - \boldsymbol{\mu}_{\bar{\mathbf{d}}}) - \sigma_{jk}(y_j - \mu_j)]^2)}{2(\sigma_{jj}\sigma_{kk} - \sigma_{jk}^2)^2} \end{pmatrix}. \quad (5.35)$$

Here,  $\bar{\mathbf{d}}$  refers to the set of indices  $(1, 2, \dots, d-1)$ , corresponding to the observed portion of  $\mathbf{y}$ .

Turning to the other expectation, we find:

$$\begin{aligned} \mathbf{U}(y_j, y_k) &= \frac{\partial \ln \phi(y_j, y_k; \mu_j, \mu_k, \sigma_{jj}, \sigma_{jk}, \sigma_{kk})}{\partial(\mu_j, \mu_k, \sigma_{jj}, \sigma_{jk}, \sigma_{kk})} \\ &= \begin{pmatrix} \Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu}) \\ h_{jj} + Q_{jj} \\ h_{jk} + Q_{jk} \\ h_{kk} + Q_{kk} \end{pmatrix}, \end{aligned} \quad (5.36)$$

where

$$h_{jj} = -\frac{1}{2} \frac{\sigma_{kk}}{\varphi}, \quad h_{jk} = \frac{\sigma_{jk}}{\varphi}, \quad h_{kk} = -\frac{1}{2} \frac{\sigma_{jj}}{\varphi},$$

$$\varphi = \sigma_{jj}\sigma_{kk} - \sigma_{jk}^2,$$

$$Q_\sigma = \frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \Sigma^{-1} S_\sigma \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu}),$$

$$S_{jj} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad S_{jk} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad S_{kk} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Here,  $S_\sigma$  is generic notation for either one of the three pairs  $(j, j)$ ,  $(j, k)$ , and  $(k, k)$ .

To calculate the expectation of (5.36), we need:

$$E(\mathbf{Y} | \mathbf{y}_{\bar{d}}) = \boldsymbol{\mu}_{jk}^c = \boldsymbol{\mu} + \Sigma_{jk, \bar{d}} \Sigma_{\bar{d}, \bar{d}}^{-1} (\mathbf{y}_{\bar{d}} - \boldsymbol{\mu}_{\bar{d}}), \quad (5.37)$$

$$\text{var}(\mathbf{Y} | \mathbf{y}_{\bar{d}}) = \Sigma_{jk, jk} - \Sigma_{jk, \bar{d}} \Sigma_{\bar{d}, \bar{d}}^{-1} \Sigma_{\bar{d}, jk}. \quad (5.38)$$

It now follows that

$$E[\mathbf{U}(y_j, y_k) | \mathbf{y}_{\bar{d}}] = \begin{pmatrix} \Sigma_{jk, jk}^{-1} \Sigma_{jk, \bar{d}} \Sigma_{\bar{d}, \bar{d}}^{-1} (\mathbf{y}_{\bar{d}} - \boldsymbol{\mu}_{\bar{d}}) \\ h_{jj} + E[Q_{jj} | \mathbf{y}_{\bar{d}}] \\ h_{jk} + E[Q_{jk} | \mathbf{y}_{\bar{d}}] \\ h_{kk} + E[Q_{kk} | \mathbf{y}_{\bar{d}}] \end{pmatrix}, \quad (5.39)$$

where some straightforward algebra produces:

$$\begin{aligned} E[Q_\sigma | \mathbf{y}_{\bar{d}}] &= \frac{1}{2} \text{tr} \left\{ \Sigma_{jk, jk}^{-1} S_\sigma \Sigma_{jk, jk}^{-1} \left[ \Sigma_{jk, jk} + \Sigma_{jk, \bar{d}} \Sigma_{\bar{d}, \bar{d}}^{-1} \right. \right. \\ &\quad \left. \left. \times \left( (\mathbf{y}_{\bar{d}} - \boldsymbol{\mu}_{\bar{d}})(\mathbf{y}_{\bar{d}} - \boldsymbol{\mu}_{\bar{d}})' - \Sigma_{\bar{d}, \bar{d}} \right) \Sigma_{\bar{d}, \bar{d}}^{-1} \Sigma_{\bar{d}, jk} \right] \right\}. \end{aligned} \quad (5.40)$$

In the special case of two measurements, the first of which always observed,  $\bar{d} = 1$  in (5.35), i.e., it refers to the first measurement. Hence, both expectations in (5.35) reduce to 0, implying in turn that then  $E_{y_m | y_o} \mathbf{U}(y_2 | y_1) = E_{y_2 | y_1} \mathbf{U}(y_2 | y_1) = \mathbf{0}$ , as it should because in this simple case pseudo-likelihood coincides with full likelihood.

For each of the estimators, the sandwich estimator can be computed. Next, we will provide generic expressions of the sandwich estimator for the case of IPWCC and its doubly robust version.

### Sandwich Estimator for $U_{IPWCC}$ and $U_{IPWCC,dr}$ With Normal Data

Write a subject's contribution to (5.27) as

$$\mathbf{V}_i = \frac{R_i}{\pi_i} \sum_{j < k} \mathbf{U}(y_{ij}, y_{ik}) = \frac{R_i}{\pi_i} \sum_{j < k} \frac{\partial \ell_{ijk}}{\partial \boldsymbol{\theta}} = \frac{R_i}{\pi_i} \mathbf{U}_i. \quad (5.41)$$

The model for missingness can be written in logistic form as:

$$\pi_i = \prod_{j=2}^{n_i} \left( 1 + e^{\mathbf{z}'_{ij} \boldsymbol{\psi}} \right)^{-1},$$

where  $\mathbf{z}_{ij}$  is a vector containing relevant covariates and outcomes from the history prior to occasion  $j$ . Then,

$$\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}} = \frac{R_i}{\pi_i} \cdot K' \frac{\partial^2 \ell_{ijk}}{\partial (\boldsymbol{\mu}, \boldsymbol{\sigma}) \partial (\boldsymbol{\mu}, \boldsymbol{\sigma})'} K, \quad (5.42)$$

$$\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\psi}} = \frac{R_i}{\pi_i} \cdot \mathbf{U}_i \sum_{k=2}^{n_i} \mathbf{z}_{ik} p_{ik}, \quad (5.43)$$

with

$$K = \begin{pmatrix} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{\sigma}}{\partial \boldsymbol{\alpha}} \end{pmatrix}, \quad p_{ik} = \frac{e^{\mathbf{z}'_{ik} \boldsymbol{\psi}}}{1 + e^{\mathbf{z}'_{ik} \boldsymbol{\psi}}}.$$

Next, the estimating equation  $W_i$  for the  $\boldsymbol{\psi}$  parameters follows from its logistic structure, with data of the form  $(R_{ij}, \mathbf{z}_{ij})$ , for  $i = 1, \dots, N$  and  $j = 1, \dots, d_i$ , and  $R_{ij} = 0$  if  $j < d_i$ , and 1 otherwise. Following standard generalized linear models theory, we have that

$$\mathbf{W}_i = \sum_{j=2}^{d_i} \mathbf{z}'_{ij} (R_{ij} - p_{ij}). \quad (5.44)$$

Hence,

$$\frac{\partial \mathbf{W}_i}{\partial \boldsymbol{\psi}} = - \sum_{j=2}^{d_i} (\mathbf{z}_{ij} \cdot \mathbf{z}'_{ij}) p_{ij} (1 - p_{ij}). \quad (5.45)$$

The sandwich estimator then follows from plugging the expressions (5.41) and (5.44) for the scores, and (5.42), (5.43), and (5.45) for the second derivatives, into (4.8) and (4.9). We still need an expression for

$$\frac{\partial^2 \ell_{ijk}}{\partial (\boldsymbol{\beta}, \boldsymbol{\alpha}) \partial (\boldsymbol{\beta}, \boldsymbol{\alpha})'}.$$

Define

$$H^{(2)} = \frac{\partial \mathbf{h}}{\partial \boldsymbol{\sigma}}, \quad Q^{(2)} = \frac{\partial \mathbf{Q}}{\partial \boldsymbol{\sigma}},$$

with  $\mathbf{h} = (h_{jj}, h_{jk}, h_{kk})'$  and  $\mathbf{Q} = (Q_{jj}, Q_{jk}, Q_{kk})'$ . Then,

$$H^{(2)} = \frac{1}{\varphi^2} \begin{pmatrix} -\frac{1}{2}\sigma_{kk}^2 & \sigma_{jj}\sigma_{kk} & \frac{1}{2}\sigma_{jk}^2 \\ -\sigma_{kk}\sigma_{jk} & \sigma_{jj}\sigma_{kk} + \sigma_{jk}^2 & -\sigma_{jj}\sigma_{jk} \\ \frac{1}{2}\sigma_{jk}^2 & \sigma_{jj}\sigma_{kk} & -\frac{1}{2}\sigma_{jj}^2 \end{pmatrix}.$$

The generic element of  $Q^{(2)}$  is

$$Q_{\sigma,\tau} = -\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_i)' \Sigma^{-1} (S_\sigma \Sigma^{-1} S_\tau + S_\tau \Sigma^{-1} S_\sigma) \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i).$$

Finally,

$$\frac{\partial^2 \ell_{ijk}}{\partial(\beta, \alpha) \partial(\beta, \alpha)'} = \left( \begin{array}{c|c} -\Sigma^{-1} & T^{(2)} \\ \hline T^{(2)'} & H^{(2)} + Q^{(2)} \end{array} \right),$$

where  $T^{(2)}$  is a  $2 \times 3$  matrix with columns  $-\Sigma^{-1} S_\sigma \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)$ .

Let us turn to the doubly robust version (5.30). Evidently,  $\mathbf{W}_i$  and  $\partial \mathbf{W}_i / \partial \boldsymbol{\psi}$  remain as before, with the same holding true for the form of  $\mathbf{S}_i$  and  $A_i$ . However, the contribution  $\mathbf{V}_i$  of subject  $i$  changes and can also be written as

$$\begin{aligned} \mathbf{V}_i &= \mathbf{V}_i^{(1)} + \left(1 - \frac{R_i}{\pi_i}\right) \mathbf{V}_i^{(2)}, \\ \mathbf{V}_i^{(1)} &= \sum_{j < k < d_i} \mathbf{U}(y_{ij}, y_{ik}), \\ \mathbf{V}_i^{(2)} &= \sum_{j=1}^{d_i-1} (n_i - d_i + 1) \mathbf{U}(y_{ij}) + \sum_{j < d_i \leq k} E[\mathbf{U}(y_{ik} | y_{ij})] + \sum_{d_i \leq j < k} E[\mathbf{U}(y_{ij}, y_{ik})]. \end{aligned}$$

We merely need derivatives with respect to  $\boldsymbol{\theta}$  and  $\boldsymbol{\psi}$ . Regarding the latter, we obtain:

$$\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\psi}} = -\frac{R_i}{\pi_i} \mathbf{V}_i^{(2)} \sum_{k=2}^{n_i} \mathbf{z}_{ik} p_{ik},$$

while for the former, the general form is

$$\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}} = \frac{\partial \mathbf{V}_i^{(1)}}{\partial \boldsymbol{\theta}} + \left(1 - \frac{R_i}{\pi_i}\right) \frac{\partial \mathbf{V}_i^{(2)}}{\partial \boldsymbol{\theta}}.$$

Now, denote by  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{n_i})'$ , the entire mean vector and by  $\boldsymbol{\sigma} = \text{vech}(\Sigma)$ , the

vector of unique variance-covariance matrix elements. It then easily follows that

$$\frac{\partial \mathbf{V}_i^{(1)}}{\partial \boldsymbol{\theta}} = K' \left( \sum_{j < k < d_i} \frac{\partial^2 \ell_{ijk}}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma}) \partial(\boldsymbol{\mu}, \boldsymbol{\sigma})'} \right) K, \quad (5.46)$$

$$\begin{aligned} \frac{\partial \mathbf{V}_i^{(2)}}{\partial \boldsymbol{\theta}} = & K' \left[ \sum_{j < d_i} (n_i - d_i + 1) \frac{\partial^2 \ell_{ij}}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma}) \partial(\boldsymbol{\mu}, \boldsymbol{\sigma})'} + \sum_{j < d_i \leq k} \frac{\partial}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})} E \left( \frac{\partial \ell_{ik|j}}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})'} \right) \right. \\ & \left. + \sum_{d_i \leq j < k} \frac{\partial}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})} E \left( \frac{\partial \ell_{ijk}}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})'} \right) \right] K. \end{aligned} \quad (5.47)$$

The derivatives in (5.46)–(5.47) follow in the same fashion as in the single robust case, starting from explicit expressions (5.37)–(5.40).

### 5.2.3 Conditional Pseudo-likelihood for Binary Data

Consider a single clustered outcome, such as in the National Toxicology Program Data (Section 2.4) and assume the model (Molenberghs and Ryan, 1999; Aerts *et al.*, 2002; Molenberghs and Verbeke, 2005):

$$\begin{aligned} f_i(\mathbf{y}_i; \boldsymbol{\Theta}_i) = & \quad (5.48) \\ \exp \left\{ \sum_{j=1}^{n_i} \theta_{ij} y_{ij} + \sum_{j < j'} \delta_{ijj'}^* y_{ij} y_{ij'} + \cdots + \omega_{i1 \dots n_i} y_{i1} \dots y_{in_i} - A(\boldsymbol{\Theta}_i^*) \right\}. \end{aligned}$$

or its quadratic simplification (Zhao and Prentice, 1990; Molenberghs and Ryan, 1999):

$$f_i(\mathbf{y}_i; \boldsymbol{\Theta}_i^*, n_i) = \exp \left\{ \sum_{j=1}^{n_i} \theta_i^* y_{ij} + \sum_{j < j'} \delta_i^* y_{ij} y_{ij'} - A(\boldsymbol{\Theta}_i^*) \right\}, \quad (5.49)$$

with  $\delta_i^*$  describing the association between pairs of measurements within the  $i^{\text{th}}$  unit. It is useful to code the outcomes as 1 and  $-1$ , rather than 1 and 0, whenever the number of measurements per unit is variable, to ensure coding invariance. Focusing on an exchangeable situation, define the number of measurements from unit  $i$  with positive response to be  $Z_i$ . Model (5.49) then becomes, upon absorbing constant terms into the normalizing constant and using the re-parameterization  $\theta_i = 2\theta_i^*$  and  $\delta_i = 2\delta_i^*$ :

$$f_i(\mathbf{y}_i; \boldsymbol{\Theta}_i, n_i) = \exp \left\{ \theta_i z_i^{(1)} + \delta_i z_i^{(2)} - A(\boldsymbol{\Theta}_i) \right\}, \quad (5.50)$$

with  $z_i^{(1)} = z_i$  and  $z_i^{(2)} = -z_i(n_i - z_i)$ . The normalizing constant takes the form:

$$A(\boldsymbol{\Theta}_i) = \ln \left[ \sum_{k=0}^{n_i} \binom{n_i}{k} \exp \left\{ \theta_i k^{(1)} + \delta_i k^{(2)} \right\} \right],$$

where  $k^{(1)} = k$  and  $k^{(2)} = -k(n_i - k)$ . For model (5.50), independence corresponds to  $\delta_i = 0$ . A positive  $\delta_i$  corresponds to classical clustering or overdispersion, whereas a negative parameter value occurs in the under-dispersed case. As such, estimation of the association parameter can be of interest.

Fitting the model is awkward for long sequences, owing to the presence of the normalizing constant. Therefore, it is convenient to replace the corresponding likelihood function by a pseudo-likelihood alternative, found by replacing the joint density  $f_i(\mathbf{y}_i; \Theta_i)$  by the product of univariate full conditional densities  $f(y_{ij} | \{y_{ij'}\}, j' \neq j; \Theta_i)$  for  $j = 1, \dots, n_i$ . This idea can be put into the framework (3.16) by choosing  $\delta_{\mathbf{1}_{n_i}} = n_i$  and  $\delta_{\mathbf{s}_j} = -1$  for  $j = 1, \dots, n_i$  where  $\mathbf{1}_{n_i}$  is a vector of ones and  $\mathbf{s}_j$  consists of ones everywhere, except for the  $j^{\text{th}}$  entry. For all other vectors  $\mathbf{s}$ ,  $\delta_{\mathbf{s}}$  equals zero. This pseudo-likelihood has the effect of replacing a joint mass function with a complicated normalizing constant by  $n_i$  univariate functions of logistic type.

If we can assume that outcomes within a unit are exchangeable, then there are merely two types of contribution: (1) the conditional probability of an additional success, given there are  $z_i - 1$  successes and  $n_i - z_i$  failures (this contribution occurs with multiplicity  $z_i$ ):

$$p_{is} = \frac{\exp[\theta_i - \delta_i(n_i - 2z_i + 1)]}{1 + \exp[\theta_i - \delta_i(n_i - 2z_i + 1)]},$$

and (2) the conditional probability of an additional failure, given there are  $z_i$  successes and  $n_i - z_i - 1$  failures (with multiplicity  $n_i - z_i$ ):

$$p_{if} = \frac{\exp[-\theta_i + \delta_i(n_i - 2z_i - 1)]}{1 + \exp[-\theta_i + \delta_i(n_i - 2z_i - 1)]}.$$

The log PL contribution for unit  $i$  can then be expressed as

$$p\ell_i = z_i \ln p_{is} + (n_i - z_i) \ln p_{if}. \quad (5.51)$$

The contribution of unit  $i$  to the pseudo-likelihood score vector takes the form

$$\begin{bmatrix} z_i(1 - p_{is}) - (n_i - z_i)(1 - p_{if}) \\ -z_i(n_i - 2z_i + 1)(1 - p_{is}) + (n_i - z_i)(n_i - 2z_i - 1)(1 - p_{if}) \end{bmatrix}.$$

Note that, if  $\delta_i \equiv 0$ , then  $p_{is} \equiv 1 - p_{if}$  and the first component of the score vector is a sum of terms  $z_i - n_i p_{is}$ , i.e., standard logistic regression follows.

Data can be incomplete, for example, because some litter mates die or get resorbed into the uterus line. Let there be  $m_i$  litter mates,  $n_i$  of which are viable and assessed for success/failure. This then means that (5.51) would pertain to the observed data only, whereas there are an additional  $m_i - n_i$  missing outcomes.

The general expressions (5.14)–(5.22) now take the form:

$$\mathbf{U}_{\text{naive, CC}} = \sum_{i=1}^N R_i \mathbf{U}_i(z_i, n_i - z_i) = \sum_{i=1}^N R_i \mathbf{U}_i(z_i, m_i - z_i), \quad (5.52)$$

$$\mathbf{U}_{\text{naive, AC}} = \sum_{i=1}^N \mathbf{U}_i(z_i, n_i - z_i), \quad (5.53)$$

$$\mathbf{U}_{\text{IPWCC}} = \sum_{i=1}^N \frac{R_i}{\pi_i(m_i|m_i)} \mathbf{U}_i(z_i, n_i - z_i), \quad (5.54)$$

$$\mathbf{U}_{\text{IPWAC}} = \sum_{i=1}^N \frac{I(n_i|m_i)}{\pi_i(n_i|m_i)} \mathbf{U}_i^o(z_i, n_i, m_i), \quad (5.55)$$

$$\begin{aligned} \mathbf{U}_{\text{IPWCC, dr}} = & \sum_{i=1}^N \left\{ \frac{R_i}{\pi_i(m_i|m_i)} \mathbf{U}_i(z_i, n_i - z_i) \right. \\ & \left. + \left[ 1 - \frac{R_i}{\pi_i(m_i|m_i)} \right] E_{k|z_i, n_i} [\mathbf{U}_i(z_i + k, m_i - z_i - k)] \right\}, \end{aligned} \quad (5.56)$$

$$\begin{aligned} \mathbf{U}_{\text{IPWAC, dr}} = & \sum_{i=1}^N \left\{ \frac{I(n_i|m_i)}{\pi_i(n_i|m_i)} \mathbf{U}_i^o(z_i, n_i, m_i) \right. \\ & \left. + \left[ 1 - \frac{I(n_i|m_i)}{\pi_i(n_i|m_i)} \right] E_{k|z_i, n_i} [\mathbf{U}_i(z_i + k, m_i - z_i - k)] \right\}. \end{aligned} \quad (5.57)$$

Here,  $R_i$  is the usual indicator for a complete cluster, and  $I(n_i|m_i)$  is an indicator for observing  $n_i$  out of  $m_i$  litter mates. Furthermore,  $\pi_i(n_i|m_i)$  is the probability of observing  $n_i$  out of  $m_i$  litter mates. Evidently,  $\pi_i(m_i|m_i)$  is the special case of observing a complete cluster. Result (5.57) follows from observing that the observed version of the score and the expectation over the incomplete data follow, in this case, in exactly the same way.

The quantity  $\mathbf{U}_i^o(z_i, n_i, m_i)$  in (5.55) and (5.57) follows from

$$p\ell_i^o = \ln \left\{ \sum_{k=0}^{m_i - n_i} \binom{m_i - n_i}{k} p_{is}(z_i, k)^{z_i + k} [1 - p_{if}(z_i, k)]^{m_i - z_i - k} \right\}, \quad (5.58)$$

and then constructing

$$\mathbf{U}_i^o = \frac{\partial p\ell_i^o}{\partial(\theta_i, \delta_i)}, \quad (5.59)$$

where

$$\begin{aligned} \text{logit}[p_{is}(z_i, k)] &= \theta_i - \delta_i[m_i - 2(z_i + k) + 1], \\ \text{logit}[p_{if}(z_i, k)] &= -\theta_i + \delta_i[m_i - 2(z_i + k) - 1]. \end{aligned}$$



In the NTP data (Section 2.4), especially for the higher dose groups, complete clusters may be rare, thence the AC versions become not only attractive, but actually necessary to make progress.

Overall, the AC forms are slightly more cumbersome, owing to somewhat less tractable expressions, such as (5.58). Consider full exchangeability, whence form (5.23) can be used, we obtain:

$$\mathbf{U}_{\text{IPWAC,exch}} = \sum_{i=1}^N \mathbf{U}_i^o(z_i, n_i, m_i). \quad (5.60)$$

Even though the missing data mechanism is removed, as follows from (5.23) in general, construction (5.58)–(5.59) needs to be used. This is different from the pairwise likelihood case, thanks to the marginal specification of the latter. Of course, (5.60) can be used with a numerical optimizer or equation solver, thanks to the explicit expression (5.58).

Now, using (5.50), the expectations can be written as:

$$E_{k|z_i, n_i} [\mathbf{U}_i(z_i + k, m_i - z_i - k)] = \frac{\sum_{k=0}^{m_i - n_i} e^{\theta_i k - \delta_i k(m_i - 2z_i - k)} \mathbf{U}_i(z_i + k, m_i - z_i - k)}{\sum_{k=0}^{m_i - n_i} e^{\theta_i k - \delta_i k(m_i - 2z_i - k)}}.$$

To formulate a sensible missingness model in this case, write the individual responses as  $(y_{i1}, \dots, y_{in_i}, y_{i, n_i+1}, \dots, y_{im_i})$ , with the first  $n_i$  observed and the later  $m_i - n_i$  missing. Likewise, the missingness indicators are  $(r_{i1}, \dots, r_{in_i}, r_{i, n_i+1}, \dots, r_{im_i})$ , the first set being 1 and the second part 0. Let  $x_i$  indicate the dose administered to litter  $i$ . Now, the joint distribution of  $\mathbf{Y}_i$  and  $\mathbf{R}_i$  factors as

$$f(y_{i1}, \dots, y_{in_i}, y_{i, n_i+1}, \dots, y_{im_i} | x_i) \times \\ \times f(r_{i1}, \dots, r_{in_i}, r_{i, n_i+1}, \dots, r_{im_i} | y_{i1}, \dots, y_{in_i}, y_{i, n_i+1}, \dots, y_{im_i}, x_i).$$

Here, the first factor is the one for which pseudo-likelihood is considered, whereas the second one can be written in summary-statistics form, thanks to exchangeability:  $f(n_i, m_i - n_i | z_i, n_i - z_i, x_i)$ . To explicitly acknowledge within-cluster correlation, a beta-binomial model (Skellam, 1948; Kleinman, 1973; Molenberghs and Verbeke, 2005), for example, would be a reasonable choice:

$$p_i = \frac{B[n_i + \nu_i(\rho^{-1} - 1), m_i - n_i + (1 - \nu_i)(\rho^{-1} - 1)]}{B[\nu_i(\rho^{-1} - 1), (1 - \nu_i)(\rho^{-1} - 1)]}, \quad (5.61)$$

in terms of the mean parameter  $\nu_i$  and correlation  $\rho$ , and then

$$f_i(n_i, m_i - n_i | \nu_i, \rho) = \binom{m_i}{n_i} p_i^{m_i - n_i} (1 - p_i)^{n_i}. \quad (5.62)$$

Here,  $B(\cdot, \cdot)$  is the beta function. One might write, for example:

$$\text{logit}(\nu_i) = \psi_0 + \psi_1 d_i + \psi_2 (z_i/n_i). \quad (5.63)$$

Fitting the model and other manipulations is straightforward (Molenberghs and Verbeke, 2005), even though it is not commonly implemented in standard statistical software. Alternatively, one might choose to simplify matters and simply replace (5.61) by logistic regression, in which case (5.62) and (5.63) would be retained.

### Sandwich Estimator for $U_{IPWCC}$ and $U_{IPWCC,dr}$ With Conditional PL for Binary Data

For the sandwich estimator, take for example  $IPWCC$ , which can be written in shorthand as

$$U_{IPWCC} = \sum_{i=1}^N V_i = \sum_{i=1}^N \frac{R_i}{\pi_i} U_i.$$

Then,

$$\frac{\partial V_i}{\partial(\theta, \delta)} = \frac{R_i}{\pi_i} Q_i, \quad \frac{\partial V_i}{\partial \psi} = -\frac{R_i}{\pi_i^2} \frac{\partial \pi_i}{\partial \psi} U_i.$$

Here,  $Q_i$  has elements:

$$\begin{aligned} q_{i,11} &= -z_i p_{is}(1 - p_{is}) - (n_i - z_i) p_{if}(1 - p_{if}), \\ q_{i,12} = q_{i,21} &= z_i(n_i - 2z_i + 1) p_{is}(1 - p_{is}) + (n_i - z_i)(n_i - 2z_i - 1) p_{if}(1 - p_{if}), \\ q_{i,22} &= -z_i(n_i - 2z_i + 1)^2 p_{is}(1 - p_{is}) \\ &\quad - (n_i - z_i)(n_i - 2z_i - 1)^2 p_{if}(1 - p_{if}). \end{aligned}$$

The derivative w.r.t.  $\psi$  evidently depends on whether the beta-binomial model, or rather simpler logistic regression is chosen. Finally, let  $\mathbf{W}_i$  be the beta-binomial score equation contribution of litter  $i$ . From this, the derivative  $\partial \mathbf{W}_i / \partial \psi$  follows immediately. For the other forms, similar calculations apply.

## 5.3 Analysis of Case Studies

In this section, similarities and differences between the various types of marginal and conditional pseudo-likelihood are illustrated with two of the cases studies introduced in Chapter 2.

### 5.3.1 The Onychomycosis Trial

The response that will be investigated is the unaffected nail length, measured from the nail bed to the infected part of the nail, which is always at the free end of the nail, expressed in *mm*. This outcome has been studied extensively in Verbeke and Molenberghs (2000). Figure 5.1 shows the observed profiles of 30 randomly selected subjects from treatment group *A* and treatment group *B*, respectively.

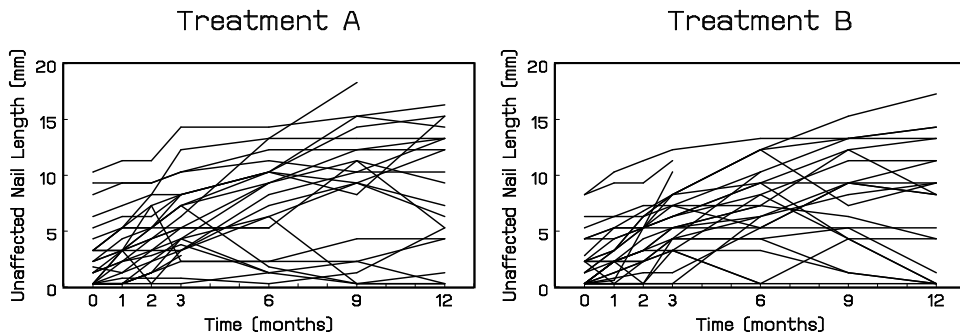


Figure 5.1: *Toenail Data*. Individual profiles of 30 randomly selected subjects in each of the treatment groups in the toenail experiment. (Source: Verbeke and Molenberghs, 2000, p.10).

The design and data type of this study is sufficiently simple to allow for full likelihood, providing a basis for comparison with which to compare the proposed pseudo-likelihood methods. Next to this, we will use several forms of pairwise marginal likelihood, as described in Section 5.2.1, in particular with the multivariate normal versions as in Section 5.2.2.

For the unaffected nail length  $Y_{ij}$ , measured at time occasion  $j$  for patient  $i$ , we specified a linear mixed-effects model:

$$Y_{ij}|b_i \sim N[b_i + \beta_0 \cdot I(T_i = 0) + \beta_1 \cdot I(T_i = 1) + \beta_2 t_j \cdot I(T_i = 0) + \beta_3 t_j \cdot I(T_i = 1), \sigma^2], \quad (5.64)$$

$$b_i \sim N(0, \tau^2),$$

where  $T_i = 0$  if patient  $i$  received standard treatment and 1 for experimental therapy ( $i = 1, \dots, 298$ ). Further,  $t_j$  is the time at which the  $j^{th}$  measurement is taken ( $j = 1, \dots, 7$ ). Finally,  $I(\cdot)$  is an indicator function. Parameter estimates and standard errors, obtained through maximum likelihood and pairwise likelihood, are presented in Table 5.1.

Table 5.1: *Toenail Data. (Unaffected nail length outcome). Parameter estimates (purely model-based standard errors; empirically corrected standard errors) for full likelihood, and naive, singly robust, and doubly robust pairwise likelihood.*

| Effect | Par.       | $\mathbf{U}_{\text{full.lik.}}$  | $\mathbf{U}_{\text{naive, CC}}$ | $\mathbf{U}_{\text{naive, CP}}$ | $\mathbf{U}_{\text{naive, AC}}$ |
|--------|------------|--|---------------------------------|---------------------------------|---------------------------------|
| Int.A  | $\beta_0$  | 2.52(0.247;0.228)  | 2.77(0.086;0.272)               | 2.70(0.081;0.248)               | 2.56(0.075;0.231)               |
| Int.B  | $\beta_1$  | 2.77(0.243;0.249)  | 2.82(0.083;0.271)               | 2.81(0.078;0.254)               | 2.77(0.073;0.250)               |
| Sl.A   | $\beta_2$  | 0.56(0.023;0.045)  | 0.55(0.011;0.046)               | 0.56(0.011;0.045)               | 0.57(0.011;0.045)               |
| Sl.B   | $\beta_3$  | 0.61(0.022;0.043)  | 0.60(0.011;0.044)               | 0.61(0.011;0.043)               | 0.61(0.010;0.043)               |
| R.I.v. | $\tau^2$   | 6.49(0.628;0.633)  | 6.71(0.226;0.731)               | 6.67(0.213;0.680)               | 6.41(0.200;0.645)               |
| Res.v. | $\sigma^2$ | 6.94(0.248;0.466)  | 7.31(0.150;0.520)               | 7.13(0.140;0.483)               | 7.05(0.137;0.472)               |
| Effect | Par.       | $\mathbf{U}_{\text{wt.lik.}}$  | $\mathbf{U}_{\text{IPWCC}}$     | $\mathbf{U}_{\text{IPWCP}}$     | $\mathbf{U}_{\text{IPWAC}}$     |
| Int.A  | $\beta_0$  | 1.85(0.092;0.303)  | 2.71(0.074;0.266)               | 2.77(0.079;0.270)               | 2.59(0.069;0.237)               |
| Int.B  | $\beta_1$  | 2.65(0.089;0.517)  | 2.78(0.073;0.265)               | 2.82(0.077;0.269)               | 2.77(0.069;0.249)               |
| Sl.A   | $\beta_2$  | 0.68(0.014;0.068)  | 0.54(0.010;0.046)               | 0.53(0.010;0.044)               | 0.55(0.010;0.045)               |
| Sl.B   | $\beta_3$  | 0.73(0.013;0.101)  | 0.60(0.010;0.044)               | 0.59(0.010;0.044)               | 0.60(0.010;0.043)               |
| R.I.v. | $\tau^2$   | 6.21(0.235;1.032)  | 6.66(0.195;0.717)               | 6.72(0.209;0.753)               | 6.44(0.187;0.669)               |
| Res.v. | $\sigma^2$ | 5.05(0.088;0.603)  | 7.29(0.130;0.513)               | 7.59(0.142;0.562)               | 7.35(0.130;0.514)               |
| Effect | Par.       | $\mathbf{U}_{\text{IPW,exch}} = \mathbf{U}_{\text{IPWCP,dr}} = \mathbf{U}_{\text{IPWAC,dr}}$ |                                 |                                 |                                 |
| Int.A  | $\beta_0$  | 2.52(0.074;0.226)  |                                 |                                 |                                 |
| Int.B  | $\beta_1$  | 2.77(0.072;0.247)  |                                 |                                 |                                 |
| Sl.A   | $\beta_2$  | 0.56(0.011;0.046)  |                                 |                                 |                                 |
| Sl.B   | $\beta_3$  | 0.61(0.011;0.044)  |                                 |                                 |                                 |
| R.I.v. | $\tau^2$   | 6.23(0.197;0.636)  |                                 |                                 |                                 |
| Res.v. | $\sigma^2$ | 7.09(0.139;0.483)  |                                 |                                 |                                 |

For the comparison purpose, weighted likelihood analysis ( $\mathbf{U}_{\text{wt.lik.}}$ ) was also considered. Observe that all point estimates are relatively close to each other, except for some deviation in the weighted likelihood analysis. Note that, with likelihood, there is little rationale to weigh here, and actually happens to result in a poorer fit.

The purely model-based standard errors are meaningful only in the standard likelihood case, where they are reasonable close to the empirically corrected ones. They are not meaningful in the weighted analyses, because they are based on the incorrect assumption that the weights represent replication at the subject (or pair) level. Fur-

thermore, naive standard errors in the pseudo-likelihood case are based on the entirely incorrect assumption that every pair results from independent replication, whereas, for example in a completely observed sequence, every measurement is used in six different pairs. This is no problem, as long as one resorts to the proper empirically corrected standard errors for inferential purposes.

It is clear that using complete cases only results in a small loss of efficiency, in the naive and IPW cases, whereas the available-case approach makes optimal use of the data. Turning to the doubly robust versions, not only is it confirmed that all three coincide, they are also very close to full likelihood, both in terms of point estimates and precision.

In a relatively large data set with continuous outcomes, like this one, treating the weights in the weighted analysis as either fixed or random does not have a noticeable impact on the standard errors. In the next study, though, there is more of a difference. The weights are based on the following logistic model (standard errors enclosed in parenthesis below coefficients):

$$\begin{aligned} & \text{logit}[P(D_i = j | D_i \geq j, T_i, t_j, Y_{i,j-1})] \\ &= \underset{(0.24)}{-3.17} - \underset{(0.24)}{0.28} T_i + \underset{(0.036)}{0.072} t_j - \underset{(0.036)}{0.035} Y_{i,j-1} \end{aligned} \quad (5.65)$$

Note that, while the effect of the previous measure is not significant, only borderline so, the weighted analyses are different from the unweighted ones. In this sense, it is a strong asset that the doubly robust versions obviate the need for using the weights, as long as the expectations are included. This is not always the case, as it is a consequence of the pairwise marginal nature of the likelihood contributions.

### 5.3.2 The National Toxicology Program Data

The NTP data, introduced in Section 2.4 will be analyzed in this section. The response of interest is the collapsed binary malformation outcome coded as one if at least one of the three malformations (external, visceral, and skeletal) occur and 0 otherwise.

We fit the models described in Section 5.2.3 with further specification:  $\theta_i = \beta_0 + \beta_1 x_i$  and  $\delta_i = \beta_d$ . Here  $x_i$  is rescaled dose, in the sense that the DEHP consumption doses of 0, 44, 91, 191, and 292 mg/kg/day are replaced by unit-interval standardized values 0.0000, 0.1507, 0.3116, 0.6541, and 1.0000, respectively.

We considered, apart from full likelihood, naive CC, naive AC, IPWCC, IPWAC, and exchangeable IPWAC. Given the equivalence of the latter to double robustness in the case of exchangeability, there is no need to further consider the other doubly robust versions.

While in this case it is obviously possible to specify full likelihood, there may be reasons to select one of the singly or doubly robust available-case versions. Indeed, in the case of likelihood, the model parameters are interpreted conditional on the number of viable fetuses, and this itself is driven by the dose assignment, an experimental variable. The available-case versions take into account the number of implants,  $m_i$ . Of course, an available-case likelihood version is in principle possible as well, which has been done and labeled  $\mathbf{U}_{\text{full.lik., AC}}$ , and based on the following modification of (5.50):

$$f_i(\mathbf{y}_i; \boldsymbol{\Theta}_i, n_i) = \sum_{k=0}^{m_i - n_i} \binom{m_i}{z_i + k} \times \exp \{ \theta_i(z_i + k) - \delta_i(z_i + k)(m_i - z_i + k) - A(\boldsymbol{\Theta}_i) \}. \quad (5.66)$$

Again, this expression has the advantage of properly acknowledging the discrepancy between the number of implants and the number of viable fetuses.

Estimated parameter and standard errors are presented in Table 5.2. For IPWAC and IPWCC, where explicit models for the weights are needed, we consider (5.63), with parameter estimates (standard errors),  $\hat{\psi}_0 = 1.960(0.110)$ ,  $\hat{\psi}_1 = 0.018(0.419)$ , and  $\hat{\psi}_2 = -2.558(0.391)$ .

Table 5.2: *Developmental Toxicity Study (DEHP). Parameter estimates (standard errors) for full likelihood, and naive, singly robust, and doubly robust pseudo-likelihood.*

| Effect | Par.      | $\mathbf{U}_{\text{full.lik.}}$ | $\mathbf{U}_{\text{full.lik., AC}}$ | $\mathbf{U}_{\text{naive, CC}}$  | $\mathbf{U}_{\text{naive, AC}}$ |
|--------|-----------|---------------------------------|-------------------------------------|----------------------------------|---------------------------------|
| Int.   | $\beta_0$ | -1.992(0.340)                   | -2.460(0.535)                       | -1.772(2.005)                    | -1.749(0.344)                   |
| Dose   | $\beta_d$ | 2.955(0.510)                    | 3.207(0.674)                        | 2.363(2.644)                     | 2.925(0.552)                    |
| Assoc. | $\beta_a$ | 0.164(0.027)                    | 0.053(0.041)                        | 0.163(0.155)                     | 0.200(0.029)                    |
| Effect | Par.      | $\mathbf{U}_{\text{IPWCC}}$     | $\mathbf{U}_{\text{IPWAC}}$         | $\mathbf{U}_{\text{IPWAC,exch}}$ |                                 |
| Int.   | $\beta_0$ | -2.888(3.825)                   | -1.335(0.831)                       | -1.470(0.164)                    |                                 |
| Dose   | $\beta_d$ | 2.145(5.969)                    | 4.588(1.021)                        | 2.225(0.293)                     |                                 |
| Assoc. | $\beta_a$ | 0.130(0.275)                    | 0.314(0.055)                        | 0.184(0.022)                     |                                 |

There are 23 complete litters, where the number of implants equals the number of viable fetuses, out of 108 litters with at least one viable fetus. This dramatic reduction of sample size shows through greatly inflated standard errors for  $\mathbf{U}_{\text{naive, CC}}$  and  $\mathbf{U}_{\text{IPWCC}}$ , up to the point where an otherwise highly significant dose effect is wiped

out. Also, the weighted version  $U_{IPWAC}$  shows a decreased efficiency. In contrast,  $U_{IPWAC, \text{exch}}$  is efficient and, while doubly robust, does not need an explicit model for the missingness probabilities; hence, it may be preferable.

## 5.4 Discussion

In this chapter, we have laid out a general framework for handling incomplete data predominantly within the pseudo-likelihood setting. Our methodology, applicable under MAR, employs ideas from inverse probability weighting and double robustness. After general development, we have focused on the pseudo-likelihood setting, elucidating in detail specific marginal and conditional instances.

Having shown that, under MAR, naive complete-case and available-case estimating equations are biased, we have formulated several alternative versions that overcome this problem, including both singly and doubly robust forms. The second of these requires evaluation of conditional expectations of the unobserved outcomes given the observed ones, which in turn may require joint distributions of a higher order than those used in the singly robust version. While at first sight this seems to undermine the appeal of pseudo-likelihood, the role of such joint distributions is solely to construct expectations which invokes considerably less computational burden. Sometimes, this might still be impractical, but then the model-based expectation can be replaced by a simpler but sufficiently rich model, in line with Bang and Robins (2005) and Meng (1994).

While in general doubly robust versions require the specification of both a weight and a predictive model, considerable simplification applies to the important special case of marginal pairwise (or, more generally,  $n$ -way) likelihood. This case is also known as composite likelihood. In this case, the doubly robust versions merely require the formulation of a predictive model. In many models these are relatively easy to compute or approximate, as was illustrated for the normal case. This is a strong asset of the combined use of doubly-robust and composite likelihood ideas. In some cases, though, the formulation of the margins (pairs) may be challenging in its own right. For example, when the conditionally specified model of Section 5.2.3 is used, formulating the full conditional pseudo-likelihood is much easier than the pairs. Thus, there is a tradeoff between simplicity in terms of weights and predictive terms on the one hand, and the pseudo-likelihood contributions themselves on the other.

For the estimation of precision we have indicated how a conventional sandwich-type estimator can be used. Should the derivation of explicit forms be deemed cumbersome, one could resort to such sampling-based methods as stochastic EM, multiple

imputation, the bootstrap, and MCMC machinery.

We have provided examples of the method, using continuous data from a clinical trial in onychomycosis and binary outcomes from a developmental toxicity study. The advantage of a variety of proposals is that the user has freedom of selection. While single robustness requires the correct specification of the weights, this requirement is less critical in the doubly robust version, because it is also possible to attain unbiasedness through the predictive term. That said, this result is in need of further qualification.



## Chapter 6

# Efficient Doubly Robust Pseudo-likelihood for Hierarchical Categorical Data

As already discussed in previous chapters, Pseudo-likelihood (PL) is closely related to but different from full likelihood. Therefore it is not guaranteed to be valid under MAR, even though in some specific cases it might, because Rubin (1976) provided conditions for *ignorability* that are sufficient but not always necessary. Yi, Zeng and Cook (2011) address this issue via a pairwise (pseudo-)likelihood method for incomplete longitudinal binary clustered data which does not require modeling the missing data process, thereby circumventing the need for the assumption of MAR. Molenberghs *et al.* (2011), on the other hand, work with the idea of supplementing PL with weighting to extend its validity under MAR. As discussed in Chapter 5, they proposed a suite of corrections to pseudo-likelihood in its standard form, also to ensure its validity under MAR. These corrections hold for pseudo-likelihood in general and follow both single and double robustness ideas, making use of inverse probability weighting (IPW), possibly supplemented with a predictive model for the unobserved outcomes given the observed ones, wherever appropriate. They have provided examples of the proposed methods using practical data.

In this chapter, we investigate the performance of the corrections proposed in Molenberghs *et al.* (2011), relative to full likelihood. As mentioned in the previous chapters, a number of concerns arise in modeling, marginally, incomplete non-Gaussian longitudinal data. Thus, here we focus on the specific case of *marginal* pairwise pseudo-likelihood for binary outcomes where missingness takes the form of

dropout. We conduct a modest simulation study to assess the performance of the latter and supplement this with a case study.

The remainder of the chapter is organized as follows. Estimating equations for pairwise likelihood are reviewed in Section 6.1. Marginal pseudo-likelihood for binary data are discussed in Section 6.2. Simulation study and analysis of a case study are presented in Section 6.3 and Section 6.4 respectively.

SAS/IML was used for all analyses. In the models where no predictive terms are involved, i.e., for the naive and singly robust versions of our models, non-linear optimization methods were used. The naive and singly robust estimators were maximized using the pseudo-likelihood and weighted-pseudo-likelihood functions, respectively.

## 6.1 Estimating Equations for Pairwise Likelihood

The ‘naive’, singly robust and doubly robust pairwise (pseudo-)likelihood estimating equations were discussed in Section 5.2.

The three estimating equation for the ‘naive’ cases are given in (5.24)–(5.26). In the  $\mathbf{U}_{\text{naive, CC}}$  case, the pairwise pseudo-likelihood (score) contributions in (5.24) are from all pairs of outcomes for subjects with fully observed  $\mathbf{Y}_i$ . The contributions in (5.25), on the other hand, come from all pairs (‘CP’ denoting complete pairs) of outcomes for subjects having at least 2 outcomes, including of course, the completers. Note that while the contributions in (5.24) come only from completers, in (5.25), these are supplemented with contributions from dropouts having incomplete  $\mathbf{Y}_i$  but with at least 2 outcomes observed. Finally, the contributions in (5.25) are further supplemented with contributions from each observed outcome, i.e., from the so-called widows, and form  $\mathbf{U}_{\text{naive, AC}}$ .

Singly robust versions of (5.24) to (5.26) are based on IPW methods, in which each subject’s contribution is weighted by the inverse probability, either of being fully observed (IPWCC), or of being observed up to a certain time (IPWAC). The singly robust versions take the form (5.27) to (5.29). Single robustness is established in Section 5.1. Note that for the case of monotone missingness or dropout, whenever  $j < k$ ,

$$R_{ijk} \equiv R_{ik} \quad \text{and} \quad \pi_{ijk} \equiv \pi_{ik} = \prod_{\ell=2}^k (1 - p_{i\ell}),$$

in which case (5.28) can be re-expressed as:

$$\mathbf{U}_{\text{IPWCP}} = \sum_{i=1}^N \sum_{j < k < d_i} \frac{R_{ik}}{\pi_{ik}} \mathbf{U}_i(y_{ij}, y_{ik}).$$

Finally, doubly robust versions (5.30–5.32) incorporate predictive models, for the unobserved outcomes given the observed ones, into the singly robust expressions.

Double robustness is shown in Section 5.1. As shown in (5.33), the three doubly robust versions coincide and give:

$$\sum_{i=1}^N \left\{ \sum_{j < k < d_i} U_i(y_{ij}, y_{ik}) + \sum_{j=1}^{d_i-1} (n_i - d_i + 1) U_i(y_{ij}) + \sum_{j < d_i \leq k} E[U_i(y_{ik} | y_{ij})] \right. \\ \left. + \sum_{d_i \leq j < k} E[U_i(y_{ij}, y_{ik})] \right\}$$

As discussed in Section 5.2.1, a number of observations should be made. The first important implication of (5.33) is that the need to specify the missing data model is obviated, as the weights are no longer present in common expression (5.33). A predictive model, however, is still necessary for the unobserved outcomes. The last two terms of (5.33) relate to the predictive models, all of which involve two types of contributions: (a)  $E[U_i(y_{ik} | y_{ij})]$  for pairs with the first component  $y_{ij}$  observed and the second one  $y_{ik}$  missing, and (b)  $E[U_i(y_{ij}, y_{ik})]$  for pairs with both components  $y_{ij}$  and  $y_{ik}$  missing. It is also worthwhile to point out that the equivalence of the three doubly robust versions holds for pseudo-likelihood in general, not just for pairwise (pseudo-)likelihood as presented here. As shown in Section 5.2.1, the expectations in (5.33) vanish under exchangeability, rendering (5.33) essentially equivalent to (5.26), thereby making the naive available case version not only valid, but actually doubly robust, of course, only under exchangeability. This result is not unexpected, because under exchangeability the expectations of an unobserved measurement given the history and be replaced, consistently, by the expectation given the *observed* portion of the history, which then vanishes.

### Precision Estimation

For the singly and doubly robust versions (5.27) to (5.29), when one posits a parametric model for dropout, the uncertainty induced by the estimation of the  $\psi$  parameters of the latter needs to be accommodated. In line with Section 4.1.4 and Section 5.1, the asymptotic variance-covariance matrix is estimated by  $\widehat{I}_0^{-1} \widehat{I}_1 \widehat{I}_0^{-1}$  where  $I_0$  and  $I_1$  are as in (4.8) and (4.9) respectively.

Note, however, that for the doubly robust case, by virtue of (5.33), an explicit model for dropout is not needed, which would in turn imply that these modifications

to  $I_0$  and  $I_1$  may actually be unnecessary. Hence,

$$I_0 = \sum_{i=1}^N \frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}} \quad \text{and} \quad I_1 = \sum_{i=1}^N \mathbf{S}_i(\hat{\boldsymbol{\theta}}) \mathbf{S}'_i(\hat{\boldsymbol{\theta}}).$$

This is also the case for the naive versions, in which no weighting is used in the estimating equations.

## 6.2 Marginal Pseudo-likelihood for Binary Data

Let us assume that we have a model for multivariate and hence also for bivariate binary data. As discussed in Section 3.4, Bahadur (1961) proposed a marginal model that accounts for the association via marginal correlations. Using the notation  $\nu_{ij} = P(Y_{ij} = 1)$ ,  $\nu_{ijk} = P(Y_{ij} = 1, Y_{ik} = 1)$ , and  $\nu_{ik|j} = P(Y_{ik} = 1 | y_{ij} = \ell)$  ( $\ell = 0, 1$ ), pairwise Bahadur probabilities take the form

$$\nu_{ijk} = \nu_{ij} \nu_{ik} \left[ 1 + \rho_{ijk} \frac{1 - \nu_{ij}}{\sqrt{\nu_{ij}(1 - \nu_{ij})}} \frac{1 - \nu_{ik}}{\sqrt{\nu_{ik}(1 - \nu_{ik})}} \right]. \quad (6.1)$$

The expressions are implicit and fitting the model is challenging from a computation time standpoint. The multivariate Bahadur model can be written as  $f(\mathbf{y}_i) = f_1(\mathbf{y}_i)c(\mathbf{y}_i)$ , where  $f_1(\mathbf{y}_i)$  and  $c(\mathbf{y}_i)$  are as in (3.23) and (3.24) respectively. Here, the  $\rho$  parameters are pairwise and higher-order correlations. Even though the model admits a convenient and concise closed form, its fitting is less than trivial, owing to strong and intractable constraints on the parameter space, be it in fully general or second-order form (where the third- and higher-order correlations are set equal to zero). This makes pseudo-likelihood attractive.

A generic contribution to the pairwise log-likelihood takes the form:

$$\begin{aligned} p\ell_{ijk} &= y_{ij}y_{ik} \ln \nu_{ijk} + y_{ij}(1 - y_{ik}) \ln(\nu_{ij} - \nu_{ijk}) + (1 - y_{ij})y_{ik} \ln(\nu_{ik} - \nu_{ijk}) \\ &\quad + (1 - y_{ij})(1 - y_{ik}) \ln(1 - \nu_{ij} - \nu_{ik} + \nu_{ijk}). \end{aligned}$$

As before, let  $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ , where  $\nu_{ij} = \nu_{ij}(\boldsymbol{\beta})$  and the association parameters are functions of  $\boldsymbol{\alpha}$ . Hence,  $\nu_{ijk} = \nu_{ijk}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ . Pairwise and conditional contributions to the score take the form:

$$\begin{aligned} U_{ijk} &= \frac{y_{ij}y_{ik}}{\nu_{ijk}} \frac{\partial}{\partial \boldsymbol{\theta}} \nu_{ijk} + \frac{y_{ij}(1 - y_{ik})}{\nu_{ij} - \nu_{ijk}} \frac{\partial}{\partial \boldsymbol{\theta}} (\nu_{ij} - \nu_{ijk}) + \frac{(1 - y_{ij})y_{ik}}{\nu_{ik} - \nu_{ijk}} \frac{\partial}{\partial \boldsymbol{\theta}} (\nu_{ik} - \nu_{ijk}) \\ &\quad + \frac{(1 - y_{ij})(1 - y_{ik})}{1 - \nu_{ij} - \nu_{ik} + \nu_{ijk}} \frac{\partial}{\partial \boldsymbol{\theta}} (1 - \nu_{ij} - \nu_{ik} + \nu_{ijk}) \end{aligned} \quad (6.2)$$

and

$$\begin{aligned}
U_{ik|j} &= \frac{y_{ij}y_{ik}\nu_{ij}}{\nu_{ijk}} \frac{\partial}{\partial \boldsymbol{\theta}} \left( \frac{\nu_{ijk}}{\nu_{ij}} \right) + \frac{y_{ij}(1-y_{ik})\nu_{ij}}{\nu_{ij}-\nu_{ijk}} \frac{\partial}{\partial \boldsymbol{\theta}} \left( \frac{\nu_{ij}-\nu_{ijk}}{\nu_{ij}} \right) \\
&\quad + \frac{(1-y_{ij})y_{ik}(1-\nu_{ij})}{\nu_{ik}-\nu_{ijk}} \frac{\partial}{\partial \boldsymbol{\theta}} \left( \frac{\nu_{ik}-\nu_{ijk}}{1-\nu_{ij}} \right) \\
&\quad + \frac{(1-y_{ij})(1-y_{ik})(1-\nu_{ij})}{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}} \frac{\partial}{\partial \boldsymbol{\theta}} \left( \frac{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}{1-\nu_{ij}} \right). \quad (6.3)
\end{aligned}$$

In addition, we need expectations of these over the conditional distribution of the unobserved outcomes given the observed ones. Evidently, because (6.2)–(6.3) are linear in the triplet  $y_{ij}, y_{ik}$  and  $y_{ij}y_{ik}$ , it suffices to calculate the expectations over these. Their corresponding probabilities are

$$\nu_{ij|\bar{d}} = \frac{\nu_{i\bar{d}j}}{\nu_{i\bar{d}}} \quad \text{and} \quad \nu_{ijk|\bar{d}} = \frac{\nu_{i\bar{d}jk}}{\nu_{i\bar{d}}}, \quad (6.4)$$

where  $\bar{d}$  is as defined in Section 5.2.2.

Combining (6.2) and (6.3) with (6.4) leads to:

$$\begin{aligned}
E(U_{ijk}) &= \frac{\nu_{i\bar{d}jk}}{\nu_{i\bar{d}}\nu_{ijk}} \frac{\partial}{\partial \boldsymbol{\theta}} \nu_{ijk} + \frac{\nu_{i\bar{d}j}-\nu_{i\bar{d}jk}}{\nu_{i\bar{d}}(\nu_{ij}-\nu_{ijk})} \frac{\partial}{\partial \boldsymbol{\theta}} (\nu_{ij}-\nu_{ijk}) \\
&\quad + \frac{\nu_{i\bar{d}k}-\nu_{i\bar{d}jk}}{\nu_{i\bar{d}}(\nu_{ik}-\nu_{ijk})} \frac{\partial}{\partial \boldsymbol{\theta}} (\nu_{ik}-\nu_{ijk}) \\
&\quad + \frac{\nu_{i\bar{d}}-\nu_{i\bar{d}j}-\nu_{i\bar{d}k}+\nu_{i\bar{d}jk}}{\nu_{i\bar{d}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk})} \frac{\partial}{\partial \boldsymbol{\theta}} (1-\nu_{ij}-\nu_{ik}+\nu_{ijk}) \quad (6.5)
\end{aligned}$$

and

$$\begin{aligned}
E(U_{ik|j}) &= \frac{y_{ij}\nu_{i\bar{d}k}\nu_{ij}}{\nu_{i\bar{d}}\nu_{ijk}} \frac{\partial}{\partial \boldsymbol{\theta}} \left( \frac{\nu_{ijk}}{\nu_{ij}} \right) + \frac{y_{ij}(\nu_{i\bar{d}}-\nu_{i\bar{d}k})\nu_{ij}}{\nu_{i\bar{d}}(\nu_{ij}-\nu_{ijk})} \frac{\partial}{\partial \boldsymbol{\theta}} \left( \frac{\nu_{ij}-\nu_{ijk}}{\nu_{ij}} \right) \\
&\quad + \frac{(1-y_{ij})\nu_{i\bar{d}k}(1-\nu_{ij})}{\nu_{i\bar{d}}(\nu_{ik}-\nu_{ijk})} \frac{\partial}{\partial \boldsymbol{\theta}} \left( \frac{\nu_{ik}-\nu_{ijk}}{1-\nu_{ij}} \right) \\
&\quad + \frac{(1-y_{ij})(\nu_{i\bar{d}}-\nu_{i\bar{d}k})(1-\nu_{ij})}{\nu_{i\bar{d}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk})} \frac{\partial}{\partial \boldsymbol{\theta}} \left( \frac{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}{1-\nu_{ij}} \right). \quad (6.6)
\end{aligned}$$

All probabilities involving  $\bar{d}$  are potentially high-dimensional; they would follow from the multivariate Bahadur model. We have seen, however, that several alternative routes are open. For example, here, one could simply resort to the singly robust version. Alternatively, the expectations could be replaced by simple, e.g., logistic, models. Precisely,  $E\mathbf{Y}_i^m | \mathbf{y}_i^o(y_{ij})$  could be written as a standard logistic model, where the

already present covariates are supplemented with  $\mathbf{y}_{i\bar{d}}$ , whereas for  $E_{\mathbf{Y}_i^m | \mathbf{y}_i^o}(y_{ij}y_{ik})$  the pairwise model under consideration can be used, again supplementing the covariate information with  $\mathbf{y}_{i\bar{d}}$ .

Apart from the already-mentioned approaches, expressions (6.5) and (6.6) can also be alternatively evaluated by using the classical definition of an expectation for discrete distributions. That is,

$$E(\mathbf{U}_{ijk}) \equiv E[\mathbf{U}_i(y_{ij}, y_{ik})] = \sum_{y_{ij}=0}^1 \sum_{y_{ik}=0}^1 \mathbf{U}_i(y_{ij}, y_{ik})P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik}), \quad (6.7)$$

$$E(\mathbf{U}_{ik|j}) \equiv E[\mathbf{U}_i(y_{ik}|y_{ij})] = \sum_{y_{ik}=0}^1 \mathbf{U}_i(y_{ik}|y_{ij})P(Y_{ik} = y_{ik}|Y_{ij} = y_{ij}), \quad (6.8)$$

where  $\mathbf{U}_i(y_{ij}, y_{ik})$  and  $\mathbf{U}_i(y_{ik}|y_{ij})$  are as defined in (6.2) and (6.3), while  $P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik})$  and  $P(Y_{ik} = y_{ik}|Y_{ij} = y_{ij})$  are the pairwise and conditional probabilities for the Bahadur model, respectively.

It is also worthwhile to point out that expressions (6.2)–(6.8) require derivatives with respect to the univariate and pairwise probabilities. For most pairwise models, such as the Bahadur model, they are reasonably straightforward and have been derived by various authors. The detail can be seen from Molenberghs and Verbeke (2005). The derivation of the sandwich estimator follows from logic similar to the one laid in Section 5.2.2.

### 6.3 Simulation Study

In this section, we investigate the performance of the estimating equations summarized in Section 6.1 by means of a simulation study. Simulation study results will be presented in Section 6.3.1.

For the simulation, we first generated an outcome at four time points from a Bahadur model, in which the measurement model incorporates a binary treatment indicator and evolution over time. Denoting  $x_i$  as the treatment indicator and  $t_j$  the time point at which measurement  $j$  is taken, we specify the following logistic formulation:

$$\text{logit } \nu_{ij} = \text{logit } P(Y_{ij} = 1|x_i, t_j) = \beta_0 + \beta_x x_i + \beta_t t_j + \beta_{xt} x_i t_j, \quad (6.9)$$

with  $\beta_0 = -2.5$ ,  $\beta_x = 0.1$ ,  $\beta_t = 1.0$ , and  $\beta_{xt} = -0.5$ . The correlation among the

outcomes is assumed to follow a Toeplitz structure of the form:

$$\begin{pmatrix} 1 & \rho^{(1)} & \rho^{(2)} & \rho^{(3)} \\ \rho^{(1)} & 1 & \rho^{(1)} & \rho^{(2)} \\ \rho^{(2)} & \rho^{(1)} & 1 & \rho^{(2)} \\ \rho^{(3)} & \rho^{(2)} & \rho^{(1)} & 1 \end{pmatrix}, \quad (6.10)$$

where  $\rho^{(k)}$ ,  $k = 1, 2, 3$  denotes the correlation between outcomes that are  $k$  time points apart. Hence, the Bahadur density is  $f(\mathbf{y}_i) = f_1(\mathbf{y}_i)c(\mathbf{y}_i)$ , with  $f_1(\mathbf{y}_i)$  as in (3.23) with  $n_i = 4$  and (3.24) taking the specific form:

$$\begin{aligned} c(\mathbf{y}_i) &= 1 + \sum_{\substack{j_1 < j_2 \\ j_2 - j_1 = k}} \rho_{ij_1 j_2}^{(k)} e_{ij_1} e_{ij_2}, \\ &= 1 + \rho^{(1)} (e_{i1}e_{i2} + e_{i2}e_{i3} + e_{i3}e_{i4}) + \rho^{(2)} (e_{i1}e_{i3} + e_{i2}e_{i4}) + \rho^{(3)} e_{i1}e_{i4}, \end{aligned} \quad (6.11)$$

where  $\rho^{(1)} = 0.30$ ,  $\rho^{(2)} = 0.15$  and  $\rho^{(3)} = 0.05$ .

For the missingness, we assume an MAR mechanism, with dropout possible only after the first time point. This results in a total of 4 possible dropout patterns: (1) dropout at the second time point, (2) dropout at the third time point, (3) dropout at the fourth time point, and (4) no dropout. The probability of dropout at time point  $j$ , given  $x_i$  and the measurement at the previous time point, is modeled by a logistic regression of the form

$$\text{logit } P(D_i = j | D_i \geq j, x_i, y_{i,j-1}) = \psi_0 + \psi_x x_i + \psi_{prev} y_{i,j-1}, \quad (6.12)$$

( $j = 2, 3, 4, 5$ ). To explore the effect of the amount of missingness, we consider two sets of values for the  $\psi$  parameters. First,  $\psi_0 = -2.2$ ,  $\psi_x = 0.5$  and  $\psi_{prev} = 2.0$ , for a scenario with a 26% dropout rate (Setting 1), and  $\psi_0 = -1.5$ ,  $\psi_x = 0.5$ , and  $\psi_{prev} = 2.0$ , for a scenario with a 43% dropout rate (Setting 2). For both settings, we assume a sample of size  $N = 600$  subjects, equally divided among the two treatment groups. A total of  $S = 500$  simulated data sets were generated for each scenario.

### 6.3.1 Results

For the simulated data sets, under each setting, we fit the Bahadur model as specified by (6.9)–(6.11) and the dropout model (6.12), using full likelihood, the naive estimating equations (5.24)–(5.26), the singly robust versions (5.27)–(5.29) and the doubly robust version (5.33). For the latter, the predictive terms were obtained in two ways: using the potentially high-dimensional expressions given in (6.5)–(6.6) and using the

alternative formulations given in (6.7)–(6.8). In our tabulation of results, the former case will be labeled as ‘A’ and the latter as ‘B’.

The simulation studies are summarized in terms bias, square root of the mean squared error ( $RMSE$ ) and square root of relative efficiency ( $Reff$ ), expressed as the ratio of the square root of the mean squared errors of the estimates obtained with the full likelihood and the pseudo-likelihood estimating equation, where

$$\text{Bias}(\hat{\beta}) = \bar{\beta} - \beta, \quad \bar{\beta} = \sum_{i=1}^S \frac{\hat{\beta}_i}{S}.$$

$$\text{MSE} = \text{Bias}^2(\bar{\beta}) + \text{Var}(\bar{\beta}),$$

and

$$\text{Reff} = \frac{RMSE_{full}}{RMSE_{method}}.$$

A value of  $Reff$  lower than one indicates loss of efficiency.

The results obtained for the proposed estimating equations for Settings 1 and 2, are presented respectively in Tables 6.1 and 6.2. We discuss these in turn.

For Setting 1 (Table 6.1), we observe generally unbiased estimates for most of the parameters for all methods, except for a couple of measurement model parameters (e.g.,  $\beta_0$  and  $\beta_x$ ), particularly for the naive CC and naive CP methods. Bias is also consistently smaller under the doubly robust IPW version A compared to any of the other methods. Among the naive approaches, the naive AC case leads to the lowest MSE for the  $\beta$  parameters, but all three naive cases have comparable MSE for the association parameters. The doubly robust approach yields the smallest MSEs. Finally, some loss of efficiency is observed for the naive CC, and naive CP cases, while almost equivalent efficiency as full likelihood is observed under naive AC, IPWCC, IPWCP, and IPWAC. Small to moderate gains in efficiency, as much as 330%, has been seen under the doubly robust IPW approach.



Table 6.1: *Simulation Study (Setting 1). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for naive, singly and doubly robust pairwise likelihood and full likelihood.*

| Par.                             | Naive  |        |        | IPW    |        |        | IPW, dr |        | Full Lik. |
|----------------------------------|--------|--------|--------|--------|--------|--------|---------|--------|-----------|
|                                  | CC     | CP     | AC     | CC     | CP     | AC     | A       | B      |           |
| Bias                             |        |        |        |        |        |        |         |        |           |
| $\beta_0$                        | 0.166  | 0.137  | -0.009 | -0.013 | 0.064  | 0.114  | -0.000  | -0.004 | -0.031    |
| $\beta_x$                        | 0.142  | 0.102  | -0.018 | -0.003 | 0.044  | 0.071  | 0.000   | 0.000  | 0.033     |
| $\beta_t$                        | -0.022 | -0.025 | 0.001  | 0.002  | -0.008 | -0.027 | 0.000   | 0.006  | 0.010     |
| $\beta_{xt}$                     | -0.031 | -0.022 | 0.006  | 0.002  | -0.008 | -0.016 | -0.000  | 0.001  | -0.010    |
| $\rho^{(1)}$                     | 0.003  | 0.004  | -0.000 | 0.000  | 0.002  | 0.005  | 0.000   | 0.002  | 0.002     |
| $\rho^{(2)}$                     | 0.003  | 0.003  | -0.001 | 0.001  | 0.002  | 0.006  | 0.000   | -0.002 | 0.000     |
| $\rho^{(3)}$                     | 0.002  | 0.001  | 0.001  | 0.001  | 0.001  | 0.003  | 0.000   | -0.001 | 0.013     |
| RMSE                             |        |        |        |        |        |        |         |        |           |
| $\beta_0$                        | 0.258  | 0.237  | 0.178  | 0.170  | 0.200  | 0.214  | 0.153   | 0.174  | 0.187     |
| $\beta_x$                        | 0.339  | 0.318  | 0.258  | 0.243  | 0.287  | 0.277  | 0.228   | 0.241  | 0.273     |
| $\beta_t$                        | 0.075  | 0.073  | 0.067  | 0.059  | 0.069  | 0.067  | 0.041   | 0.056  | 0.069     |
| $\beta_{xt}$                     | 0.108  | 0.103  | 0.093  | 0.080  | 0.096  | 0.087  | 0.064   | 0.074  | 0.096     |
| $\rho^{(1)}$                     | 0.027  | 0.026  | 0.025  | 0.021  | 0.024  | 0.022  | 0.007   | 0.019  | 0.022     |
| $\rho^{(2)}$                     | 0.032  | 0.031  | 0.030  | 0.025  | 0.030  | 0.026  | 0.017   | 0.027  | 0.026     |
| $\rho^{(3)}$                     | 0.045  | 0.048  | 0.043  | 0.035  | 0.044  | 0.037  | 0.018   | 0.043  | 0.044     |
| Relative efficiency <sup>a</sup> |        |        |        |        |        |        |         |        |           |
| $\beta_0$                        | 0.718  | 0.784  | 1.043  | 1.088  | 0.926  | 0.867  | 1.212   | 1.068  |           |
| $\beta_x$                        | 0.804  | 0.883  | 1.056  | 1.121  | 0.948  | 0.982  | 1.197   | 1.129  |           |
| $\beta_t$                        | 0.911  | 0.914  | 1.018  | 1.148  | 0.991  | 1.013  | 1.655   | 1.214  |           |
| $\beta_{xt}$                     | 0.884  | 0.933  | 1.025  | 1.193  | 0.990  | 1.101  | 1.491   | 1.283  |           |
| $\rho^{(1)}$                     | 0.825  | 0.849  | 0.880  | 1.069  | 0.919  | 0.997  | 3.305   | 1.169  |           |
| $\rho^{(2)}$                     | 0.814  | 0.833  | 0.862  | 1.045  | 0.874  | 0.980  | 1.509   | 0.941  |           |
| $\rho^{(3)}$                     | 0.911  | 0.915  | 0.953  | 1.169  | 0.939  | 1.123  | 2.328   | 0.959  |           |

<sup>a</sup>computed before rounding off

The corresponding results for the data generated under Setting 2 are provided in Table 6.2.

Table 6.2: *Simulation Study (Setting 2). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for naive, singly and doubly robust pairwise likelihood and full likelihood.*

| Par.                             | Naive  |        |        | IPW    |        |        | IPW, dr |        | Full   |
|----------------------------------|--------|--------|--------|--------|--------|--------|---------|--------|--------|
|                                  | CC     | CP     | AC     | CC     | CP     | AC     | A       | B      | Lik.   |
| Bias                             |        |        |        |        |        |        |         |        |        |
| $\beta_0$                        | 0.319  | 0.265  | -0.021 | -0.018 | 0.131  | 0.217  | -0.000  | -0.003 | -0.030 |
| $\beta_x$                        | 0.261  | 0.180  | -0.035 | 0.003  | 0.090  | 0.117  | 0.000   | 0.001  | 0.031  |
| $\beta_t$                        | -0.040 | -0.048 | 0.001  | 0.004  | -0.016 | -0.050 | 0.000   | 0.007  | 0.010  |
| $\beta_{xt}$                     | -0.061 | -0.043 | 0.014  | -0.002 | -0.020 | -0.029 | -0.000  | 0.000  | -0.012 |
| $\rho^{(1)}$                     | 0.005  | 0.008  | 0.001  | -0.000 | 0.004  | 0.010  | 0.000   | 0.003  | 0.003  |
| $\rho^{(2)}$                     | 0.003  | 0.003  | -0.001 | 0.001  | 0.001  | 0.010  | 0.000   | -0.003 | 0.000  |
| $\rho^{(3)}$                     | 0.003  | 0.002  | 0.001  | 0.002  | 0.001  | 0.006  | 0.000   | -0.002 | 0.016  |
| RMSE                             |        |        |        |        |        |        |         |        |        |
| $\beta_0$                        | 0.383  | 0.333  | 0.174  | 0.159  | 0.236  | 0.281  | 0.141   | 0.167  | 0.192  |
| $\beta_x$                        | 0.429  | 0.357  | 0.249  | 0.218  | 0.308  | 0.286  | 0.207   | 0.226  | 0.282  |
| $\beta_t$                        | 0.087  | 0.089  | 0.069  | 0.054  | 0.073  | 0.077  | 0.037   | 0.053  | 0.072  |
| $\beta_{xt}$                     | 0.131  | 0.117  | 0.097  | 0.071  | 0.103  | 0.085  | 0.057   | 0.069  | 0.103  |
| $\rho^{(1)}$                     | 0.031  | 0.029  | 0.027  | 0.019  | 0.025  | 0.022  | 0.005   | 0.017  | 0.024  |
| $\rho^{(2)}$                     | 0.036  | 0.034  | 0.032  | 0.022  | 0.032  | 0.026  | 0.014   | 0.028  | 0.029  |
| $\rho^{(3)}$                     | 0.051  | 0.051  | 0.047  | 0.032  | 0.049  | 0.034  | 0.013   | 0.047  | 0.050  |
| Relative efficiency <sup>a</sup> |        |        |        |        |        |        |         |        |        |
| $\beta_0$                        | 0.496  | 0.571  | 1.091  | 1.195  | 0.805  | 0.676  | 1.352   | 1.138  |        |
| $\beta_x$                        | 0.656  | 0.789  | 1.130  | 1.291  | 0.915  | 0.983  | 1.357   | 1.248  |        |
| $\beta_t$                        | 0.822  | 0.807  | 1.044  | 1.321  | 0.983  | 0.936  | 1.960   | 1.364  |        |
| $\beta_{xt}$                     | 0.781  | 0.879  | 1.054  | 1.440  | 0.997  | 1.203  | 1.785   | 1.495  |        |
| $\rho^{(1)}$                     | 0.792  | 0.828  | 0.899  | 1.297  | 0.959  | 1.088  | 4.723   | 1.458  |        |
| $\rho^{(2)}$                     | 0.795  | 0.828  | 0.887  | 1.273  | 0.900  | 1.103  | 2.040   | 1.031  |        |
| $\rho^{(3)}$                     | 0.913  | 0.915  | 1.002  | 1.459  | 0.957  | 1.360  | 3.491   | 1.007  |        |

<sup>a</sup>computed before rounding off

Similar observations can be made in this setting with more missingness. As before, some amount of bias can be observed for a few parameters, for some versions of the

naive and singly robust approaches. Also, the singly robust methods seem to lead to some improvement over the naive cases in terms of MSE. The IPWCC and doubly robust IPW methods show comparable MSE, with the latter generally yielding the lowest MSE compared to all the other methods. With respect to efficiency compared to full likelihood, the naive CC, and naive CP approaches again result in some loss of efficiency, while the naive AC and the singly robust IPW approaches seems to have fairly comparable efficiency with full likelihood (except for  $\beta_0$ ). Finally, the efficiency gains under the doubly robust IPW range from about 1% to 470%.

Comparison of the results in Tables 6.1 and 6.2 indicates that more bias is observed under the setting with more missingness for the naive and singly robust approaches, but for the doubly robust case, bias remains quite small under both settings. With respect to relative efficiency, methods that are less efficient than full likelihood become even more so with more missingness, but, for the doubly robust IPW, which already indicated increased efficiency for moderate amounts of missingness (Setting 1), the gain in efficiency is more pronounced when there is more missingness (Setting 2).

## 6.4 Analysis of the Analgesic Trial

The Analgesic trial has been introduced in Section 2.2. The dichotomized version of the ordinally scored ‘Global Satisfaction Assessment’, GSABIN, is the response of interest. For all ensuing analyses on the analgesic trial data, we consider only completers and dropouts, i.e., a subset of 328 patients from the original data set. We first build a logistic regression for the dropout indicator, in terms of the previous outcome and pain control assessment at baseline, i.e.,

$$\text{logit } P(D_i = j | D_i \geq j, x_i, y_{i,j-1}) = \psi_0 + \psi_x x_i + \psi_{prev} y_{i,j-1}. \quad (6.13)$$

The highly significant p-value ( $p < .0001$ ) for the parameter  $\psi_{prev}$  corresponding to the previous outcome provides evidence against MCAR in favor of MAR. Weights are then calculated based on predicted probabilities from this logistic model.

Preliminary analyses have indicated that, among a set of potential covariates, the linear and square effects of time  $t_{ij}$ , as well as the effect of baseline pain control assessment (PCA<sub>0</sub>, denoted  $x_i$ ) are of importance. The marginal regression model for the dichotomized GSA score, GSABIN, denoted as  $Y$ , is thus specified as

$$\text{logit } P(Y_{ij} = 1 | t_{ij}, x_i) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 x_i. \quad (6.14)$$

For the correlation across the within-subject outcomes, we posit a Toeplitz type correlation structure of the form (6.10).

Table 6.3: *Analgesic Trial. Parameter estimates (empirically-corrected standard errors) for naive, singly and doubly robust pairwise likelihood and for full likelihood.*

| Effect   | Par.         | $\mathbf{U}_{\text{naive, CC}}$   | $\mathbf{U}_{\text{naive, CP}}$   | $\mathbf{U}_{\text{naive, AC}}$ | $\mathbf{U}_{\text{full.lik.}}$ |
|--|--------------|-----------------------------------|-----------------------------------|---------------------------------|---------------------------------|
| Inter.   | $\beta_0$    | 3.131 (0.703)                     | 2.962 (0.562)                     | 2.691 (0.370)                   | 2.636 (0.523)                   |
| Time   | $\beta_1$    | -0.913 (0.504)                    | -0.908 (0.407)                    | -0.825 (0.304)                  | -0.763 (0.379)                  |
| Time <sup>2</sup>                                | $\beta_2$    | 0.170 (0.098)                     | 0.177 (0.081)                     | 0.183 (0.066)                   | 0.167 (0.078)                   |
| PCA <sub>0</sub>                                 | $\beta_3$    | -0.130 (0.136)                    | -0.125 (0.119)                    | -0.195 (0.069)                  | -0.187 (0.103)                  |
| corr <sub>1</sub>                                | $\rho^{(1)}$ | 0.217 (0.069)                     | 0.244 (0.056)                     | 0.210 (0.056)                   | 0.192 (0.474)                   |
| corr <sub>2</sub>                                | $\rho^{(2)}$ | 0.199 (0.075)                     | 0.234 (0.068)                     | 0.178 (0.068)                   | 0.160 (0.068)                   |
| corr <sub>3</sub>                                | $\rho^{(3)}$ | 0.224 (0.102)                     | 0.232 (0.103)                     | 0.116 (0.096)                   | 0.123 (0.102)                   |
| Considering Weights as Fixed                     |              |                                   |                                   |                                 |                                 |
| Effect   | Par.         | $\mathbf{U}_{\text{IPWCC}}$       | $\mathbf{U}_{\text{IPWCP}}$       | $\mathbf{U}_{\text{IPWAC}}$     |                                 |
| Inter.   | $\beta_0$    | 3.090 (0.297)                     | 2.717 (0.519)                     | 2.763 (0.381)                   |                                 |
| Time   | $\beta_1$    | -0.997 (0.200)                    | -0.774 (0.368)                    | -0.690 (0.253)                  |                                 |
| Time <sup>2</sup>                                | $\beta_2$    | 0.193 (0.039)                     | 0.154 (0.072)                     | 0.131 (0.047)                   |                                 |
| PCA <sub>0</sub>                                 | $\beta_3$    | -0.195 (0.061)                    | -0.141 (0.108)                    | -0.155 (0.074)                  |                                 |
| corr <sub>1</sub>                                | $\rho^{(1)}$ | 0.263 (0.028)                     | 0.275 (0.048)                     | 0.286 (0.031)                   |                                 |
| corr <sub>2</sub>                                | $\rho^{(2)}$ | 0.257 (0.031)                     | 0.255 (0.065)                     | 0.264 (0.033)                   |                                 |
| corr <sub>3</sub>                                | $\rho^{(3)}$ | 0.295 (0.041)                     | 0.267 (0.106)                     | 0.291 (0.042)                   |                                 |
| Incorporating Variability from Estimated Weights |              |                                   |                                   |                                 |                                 |
| Effect   | Par.         | $\mathbf{U}_{\text{IPWCC}}$       | $\mathbf{U}_{\text{IPWCP}}$       | $\mathbf{U}_{\text{IPWAC}}$     |                                 |
| Inter.   | $\beta_0$    | 3.079 (0.299)                     | 2.714 (0.521)                     | 2.767 (0.385)                   |                                 |
| Time   | $\beta_1$    | -0.999 (0.200)                    | -0.775 (0.368)                    | -0.701 (0.253)                  |                                 |
| Time <sup>2</sup>                                | $\beta_2$    | 0.194 (0.039)                     | 0.154 (0.072)                     | 0.134 (0.047)                   |                                 |
| PCA <sub>0</sub>                                 | $\beta_3$    | -0.193 (0.061)                    | -0.141 (0.108)                    | -0.155 (0.074)                  |                                 |
| corr <sub>1</sub>                                | $\rho^{(1)}$ | 0.258 (0.028)                     | 0.275 (0.049)                     | 0.284 (0.031)                   |                                 |
| corr <sub>2</sub>                                | $\rho^{(2)}$ | 0.252 (0.031)                     | 0.256 (0.065)                     | 0.260 (0.033)                   |                                 |
| corr <sub>3</sub>                                | $\rho^{(3)}$ | 0.284 (0.041)                     | 0.266 (0.107)                     | 0.276 (0.042)                   |                                 |
| Effect   | Par.         | $\mathbf{U}_{\text{IPW, dr (A)}}$ | $\mathbf{U}_{\text{IPW, dr (B)}}$ |                                 |                                 |
| Inter.   | $\beta_0$    | 2.637 (0.272)                     | 2.644 (0.301)                     |                                 |                                 |
| Time   | $\beta_1$    | -0.763 (0.182)                    | -0.761 (0.193)                    |                                 |                                 |
| Time <sup>2</sup>                                | $\beta_2$    | 0.167 (0.029)                     | 0.169 (0.033)                     |                                 |                                 |
| PCA <sub>0</sub>                                 | $\beta_3$    | -0.187 (0.017)                    | -0.188 (0.046)                    |                                 |                                 |
| corr <sub>1</sub>                                | $\rho^{(1)}$ | 0.192 (0.002)                     | 0.194 (0.053)                     |                                 |                                 |
| corr <sub>2</sub>                                | $\rho^{(2)}$ | 0.160 (0.003)                     | 0.161 (0.061)                     |                                 |                                 |
| corr <sub>3</sub>                                | $\rho^{(3)}$ | 0.123 (0.006)                     | 0.126 (0.092)                     |                                 |                                 |

The resulting parameter estimates, along with corresponding standard errors, for

model specification (6.14), with a Toeplitz correlation structure (6.10), using full likelihood and estimating equations (5.24) to (5.26) and (5.33) are presented in Table 6.3. There are two panels for the IPW cases: the first panel provides the results for IPW when the weights obtained from the dropout model are considered as fixed, while the second panel shows the corresponding results considering that the weights are estimated, in which case, the variability in the estimated weights is incorporated in the computation of the standard errors. The high degree of similarity in the results in these two panels indicates that the additional variability induced by estimation of the weight model does not seem to impact largely on either the estimates or their standard errors.

Fairly comparable results are also observed for the parameter estimates under full likelihood, naive AC and the doubly robust cases. Moreover, substantial efficiency over full likelihood seems to be gained under the naive AC and doubly robust approaches. Whereas these observations are not surprising for the doubly robust case, precisely because of their property, the relatively good performance of the naive AC case seems counterintuitive. However, under exchangeability, as shown in Chapter 5, the naive AC can be seen as a doubly robust estimator, given that then the expectation in (5.32) can be removed because observed and unobserved components from a subject's history are interchangeable. To this effect, we assessed the plausibility of the Toeplitz correlation structure of the analgesic trial data, using full likelihood, and determined that the three correlation parameters  $\rho^{(k)}$ ,  $k = 1, 2, 3$ , were not significantly different ( $p = 0.8091$ ), which implies that the underlying correlation structure might very well be exchangeable. This explains the excellent behavior of the naive AC estimator.

Next, we consider the CC and CP versions, both naive and singly robust (IPW). For the CC approaches, while the estimates for the parameters  $\beta_1, \beta_2$  and  $\beta_3$  are reasonably close to those under full likelihood, some disparity has been seen in the intercept  $\beta_0$  and in the correlation parameters, the latter particularly for the IPW cases. In addition, the standard errors under the naive CC approach are generally larger than those for the full likelihood. For IPWCC, in contrast, smaller standard errors are observed, a result that could be attributed to the single robustness of IPWCC. For the CP cases, in either the naive or the singly robust situations, the CP results seem to fall in between the CC and the AC results, implying somewhat of a compromise between the latter two. This can be inferred from the incremental nature of the contributions in expressions (5.24), (5.25), (5.26) and (5.27), (5.28), (5.29).

For the singly robust versions, throughout, all IPW versions yield correlation parameter estimates that are quite different to those obtained under the full likelihood.

This might be a result of the misspecified correlation structure, as mentioned earlier. There is seeming protection in the sense that the regression model parameters are generally reasonable, but the association parameters are not as well-protected.

## 6.5 Discussion

In this chapter, we assessed the performance of pseudo-likelihood approaches, supplemented with IPW-based corrections that take the form of singly and doubly robust estimators, as proposed in Molenberghs *et al.* (2011) and discussed in Chapter 5 of this thesis. In Chapter 5, a general framework for handling incomplete data, predominantly within the pseudo-likelihood setting, is laid out and IPW ideas are incorporated into estimating equations to ensure validity under MAR. In view of the various issues arising from marginally modeling incomplete non-Gaussian longitudinal data, we focused on *marginal pseudo-likelihood* and consider the specific case of incomplete longitudinal binary data. Our assessment was based on a simulation study, which was not meant to be extensive and exhaustive, but rather, was undertaken to demonstrate the workability and computational feasibility of the proposed methodology in Molenberghs *et al.* (2011). In addition, a number of features are also underscored in the analysis of the case study presented herein.

Simulation results indicated singly robust estimators to be at least as efficient as full likelihood, while doubly robust estimators were generally observed to be more efficient than full likelihood, with substantial gains in efficiency when there is more missingness within the data. Bias for some parameters was also observed to be larger for the naive and singly robust approaches for the setting with more missingness, whereas under the doubly robust IPW approach, bias remained fairly small under both settings considered. Similarly, from the analyses of the analgesic trial data, singly robust estimators with correctly specified dropout model and our doubly robust estimators without weights were found to be at least as efficient as direct likelihood methods. Moreover, under full or near exchangeability, the naive available case version is as efficient as the doubly robust estimators. This follows from the fact that, under the exchangeability, observed and missing components from the history can be traded for one another, implying consistency of the naive estimator. This is a very appealing property, because double robustness can be invoked without having to use weights or expectations. When the correlation structure differs from exchangeability, the weights still vanish in the AC case, but the expectations need to be calculated.

We further highlight a number of issues related to the methodology investigated in this chapter. Firstly, we underscore the difficulties encountered in fitting marginal

models for non-Gaussian longitudinal data. Random-effects models for the Gaussian case, i.e., linear mixed models, conveniently lead to marginal regression parameters within the same modeling framework, and as such, have naturally become the methodology of choice. This unified modeling framework, however, does not carry over to the non-Gaussian case, and hence, GLMMs, though popular, are ill-suited when marginal parameters are of scientific interest. While marginally specified models for non-Gaussian data exist, e.g., Bahadur (1961), these remain computationally challenging and, at times, restrictive. Further complications arise when data are incomplete. An alternative that has gained widespread acceptance makes use of so-called generalized estimating equations or GEE (Liang and Zeger, 1986), which though provides valid inferences only under MCAR, have been modified using inverse probability weighting to extend validity under MAR (Robins, Rotnitzky and Zhao, 1995). Parallel to the latter, Molenberghs *et al.* (2011) combine such weighting approaches with the advantages of pseudo-likelihood, to address numerical issues arising from complex, high-dimensional likelihood specifications, thereby providing yet another useful alternative to either full likelihood and/or weighted GEE. Here, we have demonstrated that under the most basic of situations, the methodology provides fairly efficient marginal parameters.

A second issue relates to the type of missingness. The corrections proposed in Molenberghs *et al.* (2011) and examined in this paper, consider the case of monotone missingness or dropout, which is the more common form of missingness that arises in clinical studies. Although non-monotone missing data are also encountered, these are usually in smaller numbers than dropout. In such cases, several routes are available. One might consider, multiple imputation (Rubin, 1987) to fill in all missing cases, both monotone and non-monotone, and proceed with the analysis method of choice, appropriately pooled according to multiple imputation principles. This of course does not circumvent the above-mentioned difficulties associated with marginally modeling non-Gaussian longitudinal data. Moreover, specification of imputation models for binary data, for instance, are perhaps just as problematic. Alternatively, one could impute non-monotone missing observations only to render the missingness monotone, and subsequently proceed as in the case of dropout. Yet a third consideration would be to combine multiple imputation with pseudo-likelihood (topic of Chapter 7), which would probably be a viable and advantageous approach since both the response model as well as the missingness model may be of high dimensions. And while multiple imputation approaches generally prescribe Gaussian type data, variations for non-Gaussian data can be utilized and seem reasonably stable even with model misspecifications, see, for instance, Beunckens, Sotito and Molenberghs (2008) and Birhanu *et al.* (2011).

In principle, our methodology can be used for non-monotone missingness as well; one then has to pay particular attention to the construction of both weights and predictions, and some non-trivial algebraic challenges will emerge.



## Chapter 7

# Comparison of Pseudo-likelihood and Generalized Estimating Equations for Incomplete Data

In correlated non-Gaussian data sets with moderate to large sequences, maximization of the full likelihood could be prohibitive due to excessive computational requirements. The problem gets even worse in the presence of incomplete data. As discussed in the previous chapters, two popular alternatives, GEE and PL, are only valid under the strongest missingness mechanism assumption, MCAR. Inverse probability weighting (IPW) (Robins, Rotnitzky and Zhao, 1995) and multiple imputation-based approaches (e.g., Paik, 1997) are widely used to take the missingness problem into account. Previous work by Carpenter, Kenward and Vansteelandt (2006), Beunckens, Sotto and Molenberghs (2008) and Birhanu *et al.* (2011) compared the relative performance of WGEE and MI-GEE under various misspecifications. The latter even compared WGEE and MI-GEE with Doubly Robust GEE (DR-GEE). In Chapters 5 and 6, IPW-based corrections that take the form of singly and doubly robust estimators of PL are discussed and their performance was assessed with simulation studies.

In this chapter, the efficiency and robustness of the various versions of GEE and PL in the presence of incomplete data are investigated and compared focusing on marginal

models for non-Gaussian longitudinal data with dropout. Comparisons will be made by means of a simulation study and a practical case study. Geys, Molenberghs, and Lipsitz (1998) compared GEE and PL for marginally specified odds ratio models for multivariate, clustered binary data, paying attention to exchangeable association structures. It has been reported that the efficiency of PL ranges from acceptably good to excellent (Geys, Molenberghs, and Lipsitz, 1998).

The outline of this chapter is as follows. In Section 7.1 we discuss the singly robust and doubly robust version of GEE and PL. A description and results of a simulation design is provided in Section 7.2. In Section 7.3, singly and doubly robust versions of PL and GEE will be illustrated using the analgesic trial data.

## 7.1 Inverse Probability Weighting and Multiple Imputation-based Methods

This section gives a brief review of the three approaches that are used in modeling non-Gaussian longitudinal data with dropout under an MAR missingness process.

### Inverse Probability Weighting Methods (IPW)

As discussed in previous chapters, the general idea behind the IPW method is to base estimation on the observed responses but to weight them to account for the probability of dropping out. Under MAR, the weights can be estimated as a function of the observed measurements and also as a function of the covariates and any additional variables that could help predict the dropout probability.

WGEE falls within the broad class of schemes that employ inverse probability weighting of complete cases. As described in Section 4.1.1, based on the completers only, the estimating equations for WGEE are given by

$$\text{WGEE} = \sum_{i=1}^N \frac{\tilde{R}_i}{\pi_i} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} V_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

Similarly, as discussed in Section 5.2.1, the pairwise pseudo-likelihood equivalent form to the above would be:

$$\text{WPL} = \mathbf{U}_{\text{IPWCC}} = \sum_{i=1}^N \frac{\tilde{R}_i}{\pi_i} \left[ \sum_{j < k} \mathbf{U}_i(y_{ij}, y_{ik}) \right],$$

with  $\pi_i$  as defined in (4.1) and  $\tilde{R}_i = 1$  if a subject is fully observed and 0 otherwise.

## Multiple Imputation

As was described in Section 3.6.2, the multiple imputation (MI) approach consists of multiply imputing the missing outcomes using a parametric model, followed by analyzing the resulting complete data sets using a standard method and finally combining the obtained inferences into a single one.

At the analysis stage of MI, any standard complete data method could be used. In case GEE is considered as the standard method, we refer to this combination of MI and GEE as “MI-GEE”. Similarly, one can combine MI with PL by handling the missingness using multiply imputed outcomes, followed by analyzing the resulting complete data sets using pseudo-likelihood. We refer to this combination of MI and PL as “MI-PL”.

## Doubly Robust Estimators

Doubly robust estimating equations (DR), where the weighting idea is supplemented with the use of a predictive model for the missing observations given the observed ones, is a recently developed approach to take the missingness problem into account. To yield consistent estimates, doubly robust (DR) methods need correct specification of either the weight or the predictive model, but not necessarily both.

The doubly robust GEE (DR-GEE) discussed in Section 4.1.3 takes the form:

$$\mathbf{U}_{\text{DR,GEE}} = \mathbf{U}_{\text{IPWCC,dr}} = \sum_{i=1}^N \left[ \frac{\tilde{R}_i}{\pi_i} \mathbf{U}_i(\mathbf{Y}_i) + \left( 1 - \frac{\tilde{R}_i}{\pi_i} \right) E_{\mathbf{Y}^m | \mathbf{y}^o} \mathbf{U}_i(\mathbf{Y}_i) \right],$$

As described in Section 5.1, a PL equivalence of the above double robust estimating equations would be:

$$\begin{aligned} \mathbf{U}_{\text{DR,PL}} = \mathbf{U}_{\text{IPWCC,dr}} = & \sum_{i=1}^N \left\{ \frac{\tilde{R}_i}{\pi_i} \left[ \sum_{j < k} \mathbf{U}_i(y_{ij}, y_{ik}) \right] \right. \\ & \left. + \left( 1 - \frac{\tilde{R}_i}{\pi_i} \right) E_{\mathbf{Y}_i^m | \mathbf{y}_i^o} \left[ \sum_{j < k} \mathbf{U}_i(y_{ij}, y_{ik}) \right] \right\}. \end{aligned}$$

where the second term in both expressions refer to the predictive terms of the unobserved outcomes given the observed ones.

As has been discussed in the previous chapters, for the predictive model several alternative routes are open, depending on the type of estimating equations. In Chapter 4, doubly robust estimating equations combined with multiple imputation ideas

as in Daniel (2008) were applied. In Chapter 5, we considered situations where a fully analytic approach is possible and an important special case of marginal pairwise under full or near exchangeability, where the formulation of a predictive model is merely required. In Chapter 6, we discussed two alternative routes: working with high-dimensional probabilities (equations 6.5 and 6.6) and using the classical definition of an expectation for discrete distributions (equations 6.7 and 6.8). In this chapter, we will follow Daniel's (2008) approach which combines doubly robust estimating equations with multiple imputation for both, GEE and PL, methods.

To motivate the approach for pseudo-likelihood, we refer to our data augmentation illustration of Section 4.1.3. For a longitudinal settings with two outcomes, the augmented data looks like:

| Augmented data  | Weight  |
|---|---|
| $\begin{pmatrix} Y_{1,i1} & Y_{1,i2}^{(m)} \\ Y_{2,i1} & Y_{2,i2}^{(m)} \\ Y_{2,i1} & Y_{2,i2} \end{pmatrix}$ | $\begin{matrix} 1 \\ 1 - \pi_i^{(2)-1} \\ \pi_i^{(2)-1} \end{matrix}$ |

The pairwise score contribution of the above augmented data is:

$$\sum_i \mathbf{U}(y_{1,i1}, y_{1,i2}^{(m)}) + \sum_i \left(1 - \frac{1}{\pi_i^{(2)}}\right) \cdot \mathbf{U}(y_{2,i1}, y_{2,i2}^{(m)}) + \sum_i \frac{1}{\pi_i^{(2)}} \cdot \mathbf{U}(y_{2,i1}, y_{2,i2}) \quad (7.1)$$

If we do impute the 'wiped out' outcomes in agreement with the prescription of the method, expression (7.1) take the form:

$$\sum_i \mathbf{U}(y_{1,i1}) + \sum_i \mathbf{U}(y_{1,i2}^{(m)} | y_{1,i1}) + \sum_i \mathbf{U}(y_{2,i1}) + \sum_i \left[ \frac{1}{\pi_i^{(2)}} \cdot \mathbf{U}(y_{2,i2} | y_{2,i1}) + \left(1 - \frac{1}{\pi_i^{(2)}}\right) \cdot \mathbf{U}(y_{2,i1}^{(m)}, y_{2,i2}) \right] \quad (7.2)$$

which is equivalent to (5.33). In this version the "weights" are necessary and the precision estimation is done by taking the variability in the "weights" into account as in Section 4.1.4.

## 7.2 Simulation Study

We investigate the performance of the PL and GEE in the presence of incomplete data using the simulation set up described in Section 6.3.

The data were generated from the marginally based Bahadur model which follows general formulation (6.9) with

$$\text{logit } \nu_{ij} = \text{logit } P(Y_{ij} = 1|x_i, t_j) = \beta_0 + \beta_x x_i + \beta_t t_j + \beta_{xt} x_i t_j,$$

where  $x_i$  and  $t_j$  represent a binary treatment indicator and time point at which measurements  $j$  was taken respectively. We choose  $\beta_0 = -2.5$ ,  $\beta_x = 0.1$ ,  $\beta_t = 1.0$ , and  $\beta_{xt} = -0.5$ . The correlation among the outcomes is assumed to follow a Toeplitz structure as in (6.10). The missingness is assumed to be MAR, and the probability of dropout at time point  $j$ , given  $x_i$  and the measurement at the previous time point ( $y_{i,j-1}$ ), is modeled by a logistic regression

$$\text{logit } P(D_i = j|D_i \geq j, x_i, y_{i,j-1}) = \psi_0 + \psi_x x_i + \psi_{prev} y_{i,j-1},$$

with  $\psi_0 = -2.2$ ,  $\psi_x = 0.5$  and  $\psi_{prev} = 2.0$ .

### 7.2.1 Results

In this section, we present simulation results of four scenarios. The scenarios are: correctly specified dropout and imputation model, either dropout or imputation models wrongly specified, and wrongly specified dropout and imputation models.

#### Everything Correctly Specified

We first investigate the relative performance of various versions of GEE and PL when both the dropout and the imputation model are correctly specified. The results are summarized in Table 7.1.

For comparison purposes, GEE and PL analyses on the complete data (i.e. before missingness is induced) are also considered. For this case, both GEE and PL yield unbiased and consistent estimators. Both are found to be as efficient as full likelihood. Since pseudo-likelihood allows for the estimation of both main effect parameters and association parameters, estimates of the correlation coefficient are reported for PL. GEE is restricted to main effect parameters.

When the dropout and imputation models are correctly specified, all methods consistently yield the minimum bias. WGEE produced slightly larger bias than the other methods. A notable loss of precision is observed under WGEE. All the other models except WGEE yield empirically corrected MSE that is comparable to that of the full likelihood and hence they are as efficient as the full likelihood. Since multiple imputation is rather a data filling approach, MI-GEE and MI-PL provided very comparable results for the mean parameters.

Table 7.1: *Simulation Study (correct imputation and dropout models). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for GEE and pairwise PL.*

| Par.                | GEE      |        |        |        | PL       |        |        |        | Full   |
|---------------------|----------|--------|--------|--------|----------|--------|--------|--------|--------|
|                     | Complete | WGEE   | MI-GEE | DR-GEE | Complete | WPL    | MI-PL  | DR-PL  | Lik.   |
| Bias                |          |        |        |        |          |        |        |        |        |
| $\beta_0$           | -0.014   | -0.115 | -0.019 | -0.120 | -0.017   | -0.013 | -0.019 | 0.008  | -0.031 |
| $\beta_x$           | 0.004    | 0.033  | 0.012  | -0.050 | 0.006    | -0.003 | 0.014  | -0.006 | 0.033  |
| $\beta_t$           | 0.002    | 0.087  | 0.005  | 0.026  | 0.004    | 0.002  | 0.005  | -0.004 | 0.010  |
| $\beta_{xt}$        | -0.001   | -0.035 | -0.005 | 0.014  | -0.002   | 0.002  | -0.006 | 0.004  | -0.010 |
| $\rho^{(1)}$        |          |        |        |        | -0.001   | 0.000  | -0.003 | -0.001 | 0.002  |
| $\rho^{(2)}$        |          |        |        |        | 0.001    | 0.001  | -0.003 | 0.000  | 0.000  |
| $\rho^{(3)}$        |          |        |        |        | 0.002    | 0.001  | -0.002 | 0.000  | 0.013  |
| RMSE                |          |        |        |        |          |        |        |        |        |
| $\beta_0$           | 0.185    | 0.406  | 0.190  | 0.158  | 0.184    | 0.170  | 0.189  | 0.119  | 0.185  |
| $\beta_x$           | 0.277    | 0.641  | 0.286  | 0.220  | 0.273    | 0.243  | 0.283  | 0.188  | 0.272  |
| $\beta_t$           | 0.066    | 0.055  | 0.070  | 0.058  | 0.065    | 0.059  | 0.069  | 0.035  | 0.068  |
| $\beta_{xt}$        | 0.092    | 0.081  | 0.099  | 0.081  | 0.091    | 0.080  | 0.098  | 0.056  | 0.095  |
| $\rho^{(1)}$        |          |        |        |        | 0.024    | 0.021  | 0.026  | 0.017  | 0.022  |
| $\rho^{(2)}$        |          |        |        |        | 0.028    | 0.025  | 0.030  | 0.015  | 0.026  |
| $\rho^{(3)}$        |          |        |        |        | 0.039    | 0.035  | 0.043  | 0.020  | 0.041  |
| Relative efficiency |          |        |        |        |          |        |        |        |        |
| $\beta_0$           | 1.002    | 0.457  | 0.975  | 1.170  | 1.007    | 1.088  | 0.979  | 1.555  |        |
| $\beta_x$           | 0.983    | 0.425  | 0.951  | 1.239  | 0.997    | 1.121  | 0.963  | 1.447  |        |
| $\beta_t$           | 1.031    | 1.241  | 0.975  | 1.176  | 1.047    | 1.148  | 0.984  | 1.943  |        |
| $\beta_{xt}$        | 1.037    | 1.184  | 0.962  | 1.180  | 1.048    | 1.193  | 0.978  | 1.712  |        |
| $\rho^{(1)}$        |          |        |        |        | 0.930    | 1.069  | 0.867  | 1.302  |        |
| $\rho^{(2)}$        |          |        |        |        | 0.923    | 1.045  | 0.854  | 1.683  |        |
| $\rho^{(3)}$        |          |        |        |        | 1.061    | 1.169  | 0.965  | 2.095  |        |

**Dropout Model Correct, Imputation Model Incorrect**

The focus of this scenario is to investigate how the MI-based and DR-based estimating equations perform under a misspecified imputation model. Here we compare MI-GEE, MI-PL, DR-GEE and DR-PL all having a correctly specified dropout model, but with an incorrectly specified imputation model in the sense that the previous outcome is omitted from the imputation model. Results are presented in Table 7.2.

All the considered four models perform well with a minor loss of efficiency compared their respective results of Table 7.1. This shows that MI-based methods are robust against misspecification in the imputation model and expected to perform well, as long as the imputation model is not grossly misspecified (Meng, 1994). Also misspecification occurring at the imputation step will only affect the unobserved (i.e., imputed) but not the observed part of the data.

Table 7.2: *Simulation Study (incorrect imputation model). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for singly and doubly robust GEE and pairwise PL.*

| Par.                | GEE    |        | PL     |        |
|---------------------|--------|--------|--------|--------|
|                     | MI-GEE | DR-GEE | MI-PL  | DR-PL  |
| Bias                |        |        |        |        |
| $\beta_0$           | -0.026 | -0.136 | -0.032 | 0.003  |
| $\beta_x$           | 0.008  | -0.050 | 0.013  | -0.002 |
| $\beta_t$           | 0.017  | 0.039  | 0.019  | -0.002 |
| $\beta_{xt}$        | 0.001  | 0.016  | 0.000  | 0.002  |
| $\rho^{(1)}$        |        |        | -0.046 | 0.000  |
| $\rho^{(2)}$        |        |        | -0.018 | 0.000  |
| $\rho^{(3)}$        |        |        | -0.001 | 0.000  |
| RMSE                |        |        |        |        |
| $\beta_0$           | 0.191  | 0.165  | 0.190  | 0.125  |
| $\beta_x$           | 0.287  | 0.224  | 0.284  | 0.189  |
| $\beta_t$           | 0.070  | 0.058  | 0.070  | 0.039  |
| $\beta_{xt}$        | 0.099  | 0.081  | 0.098  | 0.057  |
| $\rho^{(1)}$        |        |        | 0.029  | 0.020  |
| $\rho^{(2)}$        |        |        | 0.031  | 0.019  |
| $\rho^{(3)}$        |        |        | 0.043  | 0.022  |
| Relative efficiency |        |        |        |        |
| $\beta_0$           | 0.970  | 1.126  | 0.973  | 1.488  |
| $\beta_x$           | 0.948  | 1.214  | 0.960  | 1.438  |
| $\beta_t$           | 0.977  | 1.177  | 0.979  | 1.763  |
| $\beta_{xt}$        | 0.965  | 1.179  | 0.977  | 1.667  |
| $\rho^{(1)}$        |        |        | 0.781  | 1.109  |
| $\rho^{(2)}$        |        |        | 0.845  | 1.388  |
| $\rho^{(3)}$        |        |        | 0.973  | 1.866  |



**Imputation Model Correct, Dropout Model Incorrect**

The third simulation scenario is designed to investigate the effects of misspecification in the dropout model on inverse-probability weighting (IPW)-based methods: WGEE, DR-GEE, WPL, and DR-PL. Here, as in the previous subsection, the misspecification in the dropout model is again in the form of omission of the previous outcome ( $y_{i,j-1}$ ) from the dropout model, from which weights are obtained. The results are presented in Table 7.3.

A substantial amount of bias is observed in all methods, although this improves under DR-PL. Simulation results shows a general increase in empirically corrected MSEs and decrease in precision for WGEE and WPL. Up to 70% loss of efficiency is observed under WGEE, while the maximum loss of efficiency under WPL is 27%. Similar to the previous scenarios, the DR-GEE and DR-PL methods are the least biased and gains some efficiency compared to the full likelihood.

Table 7.3: *Simulation Study (incorrect dropout model). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for singly and doubly robust GEE and pairwise PL.*

| Par.                | GEE    |        | PL     |        |
|---------------------|--------|--------|--------|--------|
|                     | WGEE   | DR-GEE | WPL    | DR-PL  |
| Bias                |        |        |        |        |
| $\beta_0$           | -0.571 | 0.095  | 0.165  | 0.004  |
| $\beta_x$           | 0.155  | 0.096  | 0.131  | -0.004 |
| $\beta_t$           | 0.038  | -0.018 | -0.022 | -0.003 |
| $\beta_{xt}$        | 0.001  | -0.022 | -0.029 | 0.004  |
| $\rho^{(1)}$        |        |        | 0.004  | 0.000  |
| $\rho^{(2)}$        |        |        | 0.003  | 0.000  |
| $\rho^{(3)}$        |        |        | 0.001  | 0.000  |
| RMSE                |        |        |        |        |
| $\beta_0$           | 0.644  | 0.152  | 0.254  | 0.125  |
| $\beta_x$           | 0.440  | 0.226  | 0.333  | 0.189  |
| $\beta_t$           | 0.108  | 0.057  | 0.073  | 0.039  |
| $\beta_{xt}$        | 0.137  | 0.080  | 0.105  | 0.057  |
| $\rho^{(1)}$        |        |        | 0.026  | 0.020  |
| $\rho^{(2)}$        |        |        | 0.032  | 0.019  |
| $\rho^{(3)}$        |        |        | 0.048  | 0.022  |
| Relative efficiency |        |        |        |        |
| $\beta_0$           | 0.288  | 1.217  | 0.730  | 1.488  |
| $\beta_x$           | 0.619  | 1.206  | 0.818  | 1.437  |
| $\beta_t$           | 0.630  | 1.197  | 0.932  | 1.762  |
| $\beta_{xt}$        | 0.696  | 1.190  | 0.908  | 1.666  |
| $\rho^{(1)}$        |        |        | 0.858  | 1.109  |
| $\rho^{(2)}$        |        |        | 0.808  | 1.387  |
| $\rho^{(3)}$        |        |        | 0.862  | 1.866  |

**Imputation and Dropout Models Incorrect**

In this scenario, we explored the performance of DR-GEE and DR-PL under misspecification in both the dropout and imputation models. For both cases, misspecification is again in terms of omission of the previous outcome. The results are given in Table 7.4; for ease of comparison, the first panels of Tables 7.2 and 7.3 are replicated here.

Comparing the DR-GEE and DR-PL, both yield unbiased estimates with empirically corrected MSEs close to that of the full likelihood. Comparing the robustness of the singly and doubly robust methods, inverse-probability weighting (IPW)-based methods produce the highest bias and MSEs and hence, less efficient.

Table 7.4: *Simulation Study (incorrect imputation and dropout models). Bias, square root of empirically-corrected mean square error (RMSE) and relative efficiency (REff) with respect to full likelihood for singly and doubly robust GEE and pairwise PL.*

| Par.                | GEE    |        |        | PL     |        |        |
|---------------------|--------|--------|--------|--------|--------|--------|
|                     | WGEE   | MI-GEE | DR-GEE | WPL    | MI-PL  | DR-PL  |
| Bias                |        |        |        |        |        |        |
| $\beta_0$           | -0.571 | -0.026 | 0.085  | 0.165  | -0.032 | 0.001  |
| $\beta_x$           | 0.155  | 0.008  | 0.092  | 0.131  | 0.013  | 0.001  |
| $\beta_t$           | 0.038  | 0.017  | -0.009 | -0.022 | 0.019  | 0.000  |
| $\beta_{xt}$        | 0.001  | 0.001  | -0.019 | -0.029 | 0.000  | 0.000  |
| $\rho^{(1)}$        |        |        |        | 0.004  | -0.046 | -0.001 |
| $\rho^{(2)}$        |        |        |        | 0.003  | -0.018 | 0.000  |
| $\rho^{(3)}$        |        |        |        | 0.001  | -0.001 | 0.000  |
| RMSE                |        |        |        |        |        |        |
| $\beta_0$           | 0.644  | 0.191  | 0.152  | 0.254  | 0.190  | 0.130  |
| $\beta_x$           | 0.440  | 0.287  | 0.229  | 0.333  | 0.284  | 0.194  |
| $\beta_t$           | 0.108  | 0.070  | 0.056  | 0.073  | 0.070  | 0.043  |
| $\beta_{xt}$        | 0.137  | 0.099  | 0.080  | 0.105  | 0.098  | 0.061  |
| $\rho^{(1)}$        |        |        |        | 0.026  | 0.029  | 0.020  |
| $\rho^{(2)}$        |        |        |        | 0.032  | 0.031  | 0.020  |
| $\rho^{(3)}$        |        |        |        | 0.048  | 0.043  | 0.028  |
| Relative efficiency |        |        |        |        |        |        |
| $\beta_0$           | 0.288  | 0.970  | 1.217  | 0.730  | 0.973  | 1.423  |
| $\beta_x$           | 0.619  | 0.948  | 1.191  | 0.818  | 0.960  | 1.404  |
| $\beta_t$           | 0.630  | 0.977  | 1.223  | 0.932  | 0.979  | 1.569  |
| $\beta_{xt}$        | 0.696  | 0.965  | 1.198  | 0.908  | 0.977  | 1.560  |
| $\rho^{(1)}$        |        |        |        | 0.858  | 0.781  | 1.100  |
| $\rho^{(2)}$        |        |        |        | 0.808  | 0.845  | 1.271  |
| $\rho^{(3)}$        |        |        |        | 0.862  | 0.973  | 1.499  |

### 7.3 Analysis of the Analgesic Trial

As a further illustration of the singly robust and doubly robust versions of PL and GEE, we analyzed the analgesic trial introduced in Section 2.2 and analyzed in Chapter 6. As discussed in Section 6.4, the dichotomized version of the response, ‘Global Satisfaction Assessment’, is of interest. Results of fitting singly robust and doubly robust versions of GEE and PL are given in Table 7.5.

We note that there are similarities and differences between the parameter estimates and standard errors of GEE and PL analyses. The WGEE has larger standard errors than the other methods. For ease of comparison, the WPL reported in this chapter is  $U_{IPWCC}$  of equation 5.27. It has slightly larger coefficient estimates. To get a better estimates and efficiency one could consider the other singly and doubly robust versions (5.28–5.32), studied in length in Chapters 5 and 6.

Parameter estimates and standard errors of the DR-PL studied in this chapter are comparable to those obtained in Section 6.4. This implies that, multiple imputation based approach is another route for the predictive model. Combined with insights from the simulation study and attractive feature of PL for the estimation of the association parameters, we recommend for use of the PL methods.

Table 7.5: *Analgesic Trial. Parameter estimates (empirically-corrected standard errors) for singly and doubly robust versions of GEE and pairwise Pseudo-likelihood.*

|                   |              | GEE               |               |               |
|-------------------|--------------|-------------------|---------------|---------------|
| Effect            | Par.         | WGEE              | MI-GEE        | DR-GEE        |
| Inter.            | $\beta_0$    | 2.460(0.646)      | 2.773(0.593)  | 2.564(0.452)  |
| Time              | $\beta_1$    | -0.517(0.414)     | -0.694(0.410) | -0.652(0.343) |
| Time <sup>2</sup> | $\beta_2$    | 0.118(0.085)      | 0.147(0.083)  | 0.118(0.073)  |
| PCA <sub>0</sub>  | $\beta_3$    | -0.196(0.126)     | -0.234(0.109) | -0.202(0.095) |
|                   |              | Pseudo-likelihood |               |               |
| Effect            | Par.         | WPL               | MI-PL         | DR-PL         |
| Inter.            | $\beta_0$    | 3.079 (0.299)     | 2.674(0.578)  | 2.714(0.331)  |
| Time              | $\beta_1$    | -0.999 (0.200)    | -0.695(0.365) | -0.694(0.239) |
| Time <sup>2</sup> | $\beta_2$    | 0.194 (0.039)     | 0.148(0.074)  | 0.148(0.044)  |
| PCA <sub>0</sub>  | $\beta_3$    | -0.193 (0.061)    | -0.209(0.120) | -0.215(0.053) |
| corr <sub>1</sub> | $\rho^{(1)}$ | 0.258 (0.028)     | 0.298(0.049)  | 0.247(0.028)  |
| corr <sub>2</sub> | $\rho^{(2)}$ | 0.252 (0.031)     | 0.280(0.053)  | 0.202(0.029)  |
| corr <sub>3</sub> | $\rho^{(3)}$ | 0.284 (0.041)     | 0.305(0.079)  | 0.186(0.039)  |

## 7.4 Discussion

In this chapter, we evaluate the relative merits of pseudo-likelihood (PL) and GEE for incomplete data and illustrate them using simulation studies and a practical case study. In addition to the singly robust and doubly robust estimators of PL introduced in Chapters 5 and 6, here we combine MI with PL. We handle MAR missingness by multiply imputing the missing outcomes using a parametric model, followed by analyzing the resulting complete data sets using pseudo-likelihood. Inverse probability weighting-based, multiple imputation-based and doubly robust PL and GEE are compared under correctly specified, partially misspecified and severely misspecified dropout and imputation models.

Pseudo-likelihood (PL) is shown to have some advantages over GEE. A pseudo-likelihood (PL) function is constructed by modifying a joint density. On the contrary, generalized estimating equations (GEE) are resulting from modifying the score equations from the likelihood function. PL is advisable for the estimation of the association parameter. Geys, Molenberghs, and Lipsitz (1998) reported that the efficiency of PL

ranges from acceptable good to excellent. GEE has computational advantage when scientific interest is restricted in the estimation of marginal mean parameters.

WGEE is found to be inefficient and sensitive to the choice of the dropout model. In the various simulation scenarios we considered, compared to GEE, PL is as efficient as full likelihood, while the additional computational burden is minor. Since complete data methods (GEE and PL in this case) are performed on the multiply imputed data sets, MI-GEE and MI-PL provided very comparable results for the mean parameters. Based on the simulation study we conducted, the doubly robust estimators offer not only efficiency improvement over the singly robust methods, but also yield the least biased estimates.

Further investigation is recommended to fully understand situations where doubly robust estimators may go off the cliff and severe biases may occur under misspecification of both the dropout and predictive models.





## Chapter 8

# A Multiple Imputation Based Approach to Sensitivity Analysis and Effectiveness Assessments in Incomplete Longitudinal Data

It has been debated whether the primary analysis in longitudinal clinical trials should focus on efficacy or effectiveness. An important aspect of this debate is the impact of missing data arising from patient discontinuation. From the extensive literature on missing data, it is clear that the meaning and consequences of missing data depends on the situation. Permutt and Pinheiro (2009) illustrate five realistic clinical examples where the meaning and consequences of missing data differ. The situations range from that in which a patient discontinuation prior to the planned endpoint of the trial does not result in any loss of information because the dropout is itself an outcome, where in other situations there is indeed a loss of information, and in some instances the idea of a value being missing does not even make sense. For example, a patient is enrolled in a trial for prevention of premature ventricular contractions that is expected to reduce risk of sudden cardiac death. If the patient died of sudden cardiac arrest would it make sense to impute or estimate the number of premature ventricular contraction (PVC) that could have been observed had the patient not died of sudden cardiac arrest? Our view is that both efficacy and effectiveness are essential in understanding

the usefulness of a drug, and that both should be assessed in some manner whenever possible. However the relevant questions are when to put the greatest emphasis on each, and what designs and analysis are most appropriate for so doing. Arguably, this will depend, first and foremost, on the research question.

The setting addressed in this chapter is that of phase II or phase III clinical trials for investigational drugs to treat the symptoms of chronic illnesses, such as depression, pain, or diabetes. In such settings efficacy may be viewed as the effects of the drug if taken as directed; that is, the benefit of the drug expected at the endpoint of the trial assuming patients stayed on drug, counter to the fact that some dropped out. Effectiveness in these same settings may be viewed as the effects of the drug as actually taken. Specific definitions of effectiveness may vary but in general imply some type of benefit and risk assessment, some measure of the benefit minus the liability recognizing that patients who discontinue the drug, particularly because of safety or tolerability issues, are unlikely to have lasting benefit from it. Carpenter, Roger and Kenward (2011) refer to hypotheses about efficacy and effectiveness as the *de-jure* and *de-facto* hypotheses, respectively.

It is important to understand both what happens when a drug is evaluated as actually taken and when taken as directed, especially when including safety assessments in the scope of inference. And while it is important to consider when to put the greatest emphasis on which research question (Mallinckrodt and Kenward, 2009), this chapter focuses on what endpoints and analyses are most appropriate for each. Table 8.1, which borrows heavily from introductory chapters in the recent National Academy of Science guidance on the prevention and treatment of missing data (National Research Council, 2010) summarizes the estimands and estimators that may be associated with efficacy and effectiveness hypotheses.

Effectiveness of the initial randomized medication at the planned endpoint of the trial is essentially the maintained benefit at the planned endpoint attributable to the randomized medication for the period of time in which it was taken. For testing this hypothesis, it is not adequate to assess patients only until they drop out of the trial, follow-up data from the time of dropout until the planned endpoint of the trial are needed. However, ethical considerations often mandate that alternative medication be allowed after patients discontinue randomized study medication.

In the Intention-to-Treat (ITT) framework where inferences are drawn based on the originally assigned treatment, including follow-up data when alternative medications are allowed can mask or exaggerate both the efficacy and safety effects of the initially assigned treatments, thereby invalidating causal inferences for the effectiveness of the originally assigned medication (Mallinckrodt and Kenward, 2009).

Table 8.1: *Estimands and estimators commonly used to assess efficacy and effectiveness in clinical trials*

| Hypothesis        | Estimands                                    | Estimators | Data included in analysis  |
|-------------------|--|------------|--|
| Efficacy          | Mean change to planned endpoint <sup>a</sup> | DL         | Observed data while on drug  |
|                   |  | MI         | Observed data while on drug  |
|                   |  | GLM        | Observed data while on drug + LOCF imputation                          |
| Effectiveness     | Mean change to last observation              | GLM        | Observed data while on drug  |
|                   | Mean change to planned endpoint <sup>b</sup> | GLM        | Observed data while on drug + BOCF imputation                          |
|                   | Mean change to planned endpoint              | GLM        | Observed data while on drug + follow up data (rescue meds not allowed) |
| Treatment regimen | Mean change to planned endpoint <sup>c</sup> | Various    | Observed data while on drug + follow up data (rescue meds allowed)     |

<sup>a</sup>is the mean change assuming all patients taken drug as directed up to the endpoint

<sup>b</sup>is the mean change for all patients regardless of taken drug or not and assuming no alternative medications are used from before endpoint

<sup>c</sup>is the mean change for all patients regardless of compliance of test drug or alternative medications

Therefore, it has been proposed in the National Academy of Sciences guidance (National Research Council, 2010) and elsewhere (Fleming, 2011), that the hypothesis of interest is that of a treatment regimen, that is, initiating treatment with a particular intervention. However, the treatment regimen hypothesis is not useful in the situations of interest here as it is unlikely an investigational medication can be approved for use as part of a regimen unless it has first been proven safe and effective on its own.

As discussed in Section 3.6.2, a number of techniques have been used to impute the missing (follow-up) data to circumvent problems from the confounded follow-up data. Last and baseline observation carried forward (LOCF and BOCF) are perhaps the two most commonly used methods. Although the acronyms imply truly carrying observations forward in time, an LOCF result can be interpreted as either the change observed while actually taking drug, or as the change to the designed endpoint of the trial assuming the patients' condition would not have changed after discontinuing the drug. With BOCF, it is assumed that patients who discontinue drug received no lasting benefit, so the change from baseline after stopping study medication should be zero and thus the values after discontinuation should equal the baseline values.

However, the assumption that patients' condition would return to the baseline state after ceasing study medication is questionable in many situations as study effects, placebo effects, and natural time evolution also influence outcomes. Therefore, if patients receive no pharmacological benefit from a drug, either because it has no effect or because they discontinue taking the medication, their outcomes would be equal to their baseline values only if the study effect and the placebo effect were zero.

Alternatively, the placebo group provides an estimate of no pharmacological benefit of the drug that reflects the study effect and placebo effect. Hence, information from the placebo group may provide a better estimate of effectiveness for patients who discontinue drug than using patients' last or baseline observation.

Carpenter, Roger and Kenward (2011) define and illustrate a family of multiple imputation based approaches for assessing sensitivity in testing *de-jure* (efficacy) and *de-facto* (effectiveness) hypotheses. Using the placebo group to impute missing values for both the placebo and drug groups is a specific form of their "jump to reference" approach. With placebo Multiple Imputation (pMI), placebo is considered as the reference and drug patients jump to the placebo group after dropout. However, one may want placebo to jump to a standard of care, or placebo and drug to jump to standard of care (Little and Yau, 1996), for example, if interested in ITT estimand from Table 8.1. The advantage of the proposed method is that these and other choices are perfectly feasible. Although the principles and assumptions underlying the

jump to reference approach are clear and easy to understand (Carpenter, Roger and Kenward, 2011), the performance of the method has not been rigorously evaluated. Also, specific algorithms for such placebo-based imputation schemes and software for their implementation were not duly presented in the literature, leaving clinical statisticians without clear guidance and tools for conducting such analyses. The work presented in this chapter tries to bridge this gap by proposing and evaluating the performance of a novel placebo-based multiple imputation method (a specific form of jump to reference) that we refer to as placebo multiple imputation (pMI). The procedure is implemented using available software (SAS PROC MI).

Therefore, the primary objective of this chapter is to assess the statistical performance of pMI as an estimator of effectiveness (as actually taken hypothesis). The behavior of pMI was also considered in the context of a sensitivity analysis in testing the efficacy (taken as directed) hypothesis. In this context, pMI assumes the statistical behavior of drug-treated patients after dropout is the statistical behavior of all placebo-treated patients including dropouts. Thus, pMI can also be interpreted as a specific form of an MNAR analysis expected to yield a conservative estimate of efficacy.

Section 8.1 is devoted for specific details about the pMI approach. Design and analysis of the simulation study is presented in Section 8.2. Section 8.3 details analysis of the case study, Depression Trials (Section 2.3). Finally, Section 8.4 discusses these results.

## 8.1 Placebo Multiple Imputation

The concept of MI with “copy reference” or “jump to reference” has been proposed in the literature (Little and Yau, 1996; Carpenter, Roger and Kenward, 2011). In the placebo multiple imputation approach, placebo is considered as the reference and drug patients jump to the placebo group after dropout. Placebo multiple imputation (pMI) estimates visit-wise means or mean changes assuming that the statistical behavior of drug treated patients who discontinue becomes that of placebo-treated patients after the time of dropout. Two views may be taken of this estimand: 1) as an assessment of effectiveness, assuming patients who discontinue before the endpoint receive no pharmacological benefit after dropout; and 2) as a worst reasonable case assessment of efficacy– the outcome that would have been observed had the patient stayed on drug.

To implement this approach, multiple imputation is used to replace missing outcomes for drug-treated subjects who discontinued using multiple draws from the

posterior predictive distribution estimated from subjects who were randomized to the placebo arm in that same trial. To set up the imputation model, define observed subject-specific covariates ( $\mathbf{X}_i$ ) and partially observed outcomes ( $\mathbf{Y}_i^o$ ) whose joint distribution drives the imputation mechanism for missing outcomes. Let  $\mathbf{Y}_i = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)$ , be the  $1 \times T$  outcome vector containing, for the  $i^{\text{th}}$  subject,  $k_i$  observed outcomes and  $T - k_i$  unobserved outcomes; and,  $\mathbf{X}_i$  be a  $1 \times P$  vector of fully observed covariates.

Most missing values in the clinical trial settings addressed here are caused by dropouts resulting in a monotone pattern of missingness. Therefore, Bayesian regression employing factorization of the multivariate normal density for the data with monotone missingness pattern (Rubin, 1987, pp. 166-167), using the procedure that is available in SAS PROC MI, provides an easy and fast way to impute the missing values. Bayesian regression is based on the regression model fitted to placebo outcomes, relating a future outcome  $Y_t$  to the earlier outcomes  $Y_{t-2}, \dots, Y_0$ . Then a new regression model is simulated from the posterior predictive distribution of the parameters (using non-informative Jeffrey's priors for the regression coefficients and the error variance) and is used to impute the missing values for each variable (Rubin, 1987, pp. 166-167). This method of sampling from the Bayesian predictive distribution of  $\mathbf{Y}_i^m | \mathbf{Y}_i^o$  does not require MCMC iterations, but rather takes advantage of the monotone missingness pattern and the fact that the multivariate normal density can be factored into a sequence of conditional normal densities (Schafer, 1997, pp. 218-220). The basic idea is to estimate the parameters for the imputation model using only data from the placebo arm and then use those parameters to impute missing values for both the drug-treated and placebo-treated patients. Partially observed outcomes from treated subjects are used when imputing their missing outcomes only as arguments that are passed to the "placebo-estimated" imputation model when applying it to generate missing values for treated subjects.

Data are processed sequentially by repeatedly calling SAS PROC MI to impute missing outcomes at visits  $t = 1, \dots, T$  as described in detail in the following steps.

1. *Initialization.* Set  $t = 0$  (baseline visit).
2. *Iteration.* Set  $t = t + 1$ . Create a data set combining records from placebo and treated subjects with columns for covariates  $\mathbf{X}_i$  and outcomes at visits  $1, \dots, t$  with outcomes for all treated subjects set to missing at visit  $t$  and set to observed or imputed values at visits  $1, \dots, t - 1$ .
3. *Imputation.* Run Bayesian regression in SAS PROC MI on this data to impute

missing values for visit  $t$  using previous outcomes for visits 1 to  $t-1$  and baseline covariates. Note that only placebo data will be used to estimate the imputation model since no outcome is available for treated subjects at visit  $t$ .

4. Replace imputed data for all treated subjects at visit  $t$  with their observed values, whenever available. If  $t < T$  then go to Step 2, otherwise proceed to Step 5.
5. Repeat steps 1-4,  $m$  times with different seed values to create  $m$  imputed data sets. While  $m = 5$  imputations are frequently chosen, different, in particular higher number of imputations,  $m = 10$  or  $m = 20$  say, are no problem at all with current-day computational resources, especially not for commonly encountered clinical-trial sample sizes.
6. *Analysis.* For each completed data set, evaluate treatment difference at the last scheduled visit using a likelihood-based normal repeated measures model as would have been applied had the data been complete (implemented via SAS PROC MIXED).
7. *Combined Inference.* Compute pMI-based estimate and associated confidence interval (CI) for the treatment contrasts at last scheduled visit using Rubin's combining rules (Rubin, 1987), as implemented in SAS PROC MIANALYZE.

## 8.2 Simulation Study

A simulation study was conducted to assess the properties of pMI. Although the simulations were not intended to mimic any particular clinical setting, many input parameters for the simulation study were taken from the depression trial described in Section 2.3. The variance-correlation matrix (variances on the diagonal, correlations off-diagonally) used in the simulation study are provided in Table 8.2. This variance-correlation matrix is common for both treatment groups.

For the simulation study, we generated an outcome at five time points. A total of  $S = 1,000$  complete data sets were simulated from multivariate normal distribution with realistic parameters, similar to those estimated from recent clinical trials in depression. A sample of size  $N=200$  was equally divided between two treatment arms and five imputations were used in the simulation study.

Scenarios included two trajectories of patient response: 1) Improvement (IMP), where the mean trends were for patients to improve over time, such as would often be the case for symptomatic treatments of chronic illnesses; 2) Worsening (WOR),

Table 8.2: *Variance-correlation matrix of the outcome over time (variances on the diagonal, correlations off-diagonally)*

|        |        |        |        |        | Visit |
|--------|--------|--------|--------|--------|-------|
| 1      | 2      | 3      | 4      | 5      |       |
| 21.005 | 0.6905 | 0.558  | 0.4323 | 0.3761 |       |
|        | 24.907 | 0.724  | 0.564  | 0.470  |       |
|        |        | 30.679 | 0.715  | 0.627  |       |
|        |        |        | 31.644 | 0.781  |       |
|        |        |        |        | 30.346 |       |

where the mean trends were for patient to worsen over time, such as would often be the case for disease modification treatments in progressive illnesses such as Alzheimer's disease. Scenarios also included three dropout patterns, all with an overall dropout rate of 30%: 1) equal rates (30%) in the drug and placebo groups (=); 2) higher dropout in the drug group (HD), 40% dropout in the drug group versus 20% in the placebo group; 3) higher dropout in the placebo group (HP), 20% dropout in the drug group versus 40% in the placebo group. Lastly, scenarios included three levels of treatment effects.

Dropout was induced by deleting values according to a logistic model relating probability of dropout at a particular visit with changes from baseline to the previous visit in the simulated efficacy outcomes. Specific values for the logistic model were chosen so as to yield the desired dropout rates in the various scenarios. Of particular note, however, is that the dropout mechanism was missing at random (MAR), given that probabilities of dropout were based only on outcome values observed at earlier time points. In the case of MAR, the estimation and inferential methodology is valid for the entire class of MAR, regardless of the particular form of the missing data mechanism. While the methodology could be extended towards MNAR, then an explicit form for an MNAR mechanism needs to be assumed, both for generating data and for subsequent estimation, and combined with the pMI assumption.

For efficacy, the difference between drug and placebo in mean change to endpoint was a standardized effect size (ES) of 0.5, 0.3, or 0.0. For effectiveness, the mean difference at endpoint resulted from a mixture distribution where the effect size of completers was 0.5, 0.3, or 0.0, as described for efficacy and the effect size for patients who dropped out was 0.0. Therefore, the true value for the endpoint contrasts was



the weighted mean of the two groups. For example, with  $ES = 0.5$  and improvement direction of change, the estimate of treatment effect is 0.3875, and the estimate for the interaction of treatment with time is -0.7855. Hence, the true advantage of drug over placebo for efficacy at the last endpoint is 2.7545. For effectiveness, the true values would be the product of 2.7545 and probability of stay. With 20%, 30%, and 40% dropout in the drug group, the corresponding true advantage of drug over placebo for effectiveness were 2.2405, 2.0502, and 1.6842. The true values for the placebo group are summarized in Table 8.3. Given these trajectories, the assumptions for BOCF and LOCF were not valid.

Table 8.3: *Visit-wise population means in the placebo group*

| Visit |             |           |
|-------|-------------|-----------|
| Visit | Improvement | Worsening |
| 1     | 18.8        | 18.8      |
| 2     | 16.8        | 20.8      |
| 3     | 14.8        | 22.8      |
| 4     | 12.8        | 24.8      |
| 5     | 10.8        | 26.8      |

Results from pMI were compared with results from LOCF, BOCF, direct likelihood (DL) and standard multiple imputation (MI) in 18 scenarios that were arranged as a  $2 \times 3 \times 3$  factorial. Focus was on comparing pMI versus BOCF and LOCF in regards to the effectiveness estimand whereas focus was on comparing pMI versus DL and MI in regards to the efficacy estimand.

Note that pMI is different in nature from, say, LOCF and BOCF. The latter are explicit, and very strong, MNAR type assumptions about the data generating process that are very often violated in practice, as has been reported repeatedly in the literature (Kenward and Molenberghs, 2009). pMI is a way to make operational the effectiveness estimand. This implies that it may have regulatory and other relevance, regardless of whether or not it is the true data generating mechanism. Furthermore, while pMI in the way used here requires MAR, it does not require the explicit specification of the missing data mechanism but rather is valid over the entire class.

### 8.2.1 Results

Simulation results were summarized in terms of Bias ( $\text{Bias}(\hat{\beta})$ ), Relative Bias, Variance ( $\text{Var}(\hat{\beta})$ ), Mean Square Error (MSE), Confidence Interval (CI) coverage and

Rejection rates.

Bias is defined as the difference between mean estimate and true value of the parameter whereas relative bias is the percentage difference between the mean estimate and the true value of the parameter. Confidence interval (CI) coverage is estimated by the percentage of the retained data sets in which the Wald 95 percent confidence interval for  $\hat{\beta}$  included the true value of the parameter. Rejection rates is estimated by the proportion of data sets where the two-sided Wald test of null hypothesis ( $H_0$ ) of no treatment difference was rejected at 5% level of significance.

$$\text{Bias}(\hat{\beta}) = \bar{\beta} - \beta, \quad \bar{\beta} = \sum_{i=1}^S \frac{\hat{\beta}_i}{S}, \quad \beta \text{ represent true parameter.}$$

$$\text{MSE} \equiv \text{MSE}_{MC}(\bar{\beta}) = \text{Bias}^2(\bar{\beta}) + \text{Var}(\bar{\beta}) \quad ,$$

and

$$\text{Relative bias} = \frac{\text{bias}}{\text{true parameter}}.$$

Positive bias indicates that the average estimate of the treatment contrast is larger than true value, when  $\text{ES} > 0$ , and that the average contrast favors placebo, when  $\text{ES} = 0$ . Negative bias indicates that the average estimate of the treatment contrasts is smaller than the true value, when  $\text{ES} > 0$ , and it favors drug when  $\text{ES} = 0$ .

### 8.2.1.1 Effectiveness

Tables 8.4 through 8.8 summarize results from testing the effectiveness hypothesis. Regarding bias, pMI had minimal to no bias in all scenarios; BOCF and LOCF had large biases in almost all scenarios, with the bias in BOCF favoring drug effectiveness in 5/18 scenarios. DL and MI were biased in favor of drug whenever  $\text{ES} > 0$  but unbiased when  $\text{ES} = 0$ .

Table 8.4: *Bias in estimates of the effectiveness estimand for the analysis of change from baseline*

| Scenarios  |     |                 | Method |        |        |        |        |
|------------|-----|-----------------|--------|--------|--------|--------|--------|
| Trajectory | ES  | dropout pattern | BOCF   | LOCF   | DL     | MI     | PMI    |
| IMP        | 0.5 | =               | 0.727  | 0.625  | -0.661 | -0.659 | 0.059  |
|            |     | HD              | 1.243  | 0.925  | -1.031 | -1.031 | -0.096 |
|            |     | HP              | 0.253  | 0.363  | -0.474 | -0.473 | -0.037 |
| IMP        | 0.3 | =               | 0.362  | 0.300  | -0.470 | -0.468 | 0.009  |
|            |     | HD              | 1.025  | 0.757  | -0.627 | -0.624 | -0.015 |
|            |     | HP              | -0.093 | 0.0408 | -0.284 | -0.279 | 0.002  |
| IMP        | 0   | =               | 0.012  | 0.018  | 0.039  | 0.044  | 0.105  |
|            |     | HD              | 0.681  | 0.512  | 0.041  | 0.038  | 0.132  |
|            |     | HP              | -0.665 | -0.469 | 0.037  | 0.036  | 0.067  |
| Worse      | 0.5 | =               | 0.852  | 0.599  | -0.808 | -0.807 | -0.136 |
|            |     | HD              | -0.265 | 0.089  | -1.035 | -1.036 | -0.178 |
|            |     | HP              | 1.622  | 1.034  | -0.527 | -0.521 | -0.076 |
| Worse      | 0.3 | =               | 0.500  | 0.367  | -0.471 | -0.462 | -0.035 |
|            |     | HD              | -0.551 | -0.094 | -0.633 | -0.629 | -0.069 |
|            |     | HP              | 1.287  | 0.739  | -0.301 | -0.293 | -0.137 |
| Worse      | 0   | =               | 0.008  | 0.022  | 0.033  | 0.034  | 0.115  |
|            |     | HD              | -0.836 | -0.269 | 0.042  | 0.037  | 0.123  |
|            |     | HP              | 0.839  | 0.302  | 0.046  | 0.045  | -0.029 |

Figure 8.1 shows relative bias in estimates of the effectiveness estimand by direction of change and dropout pattern for the 0.3 effect size of the drug. pMI has the smallest relative bias in both, improvement and worsening, trajectories.

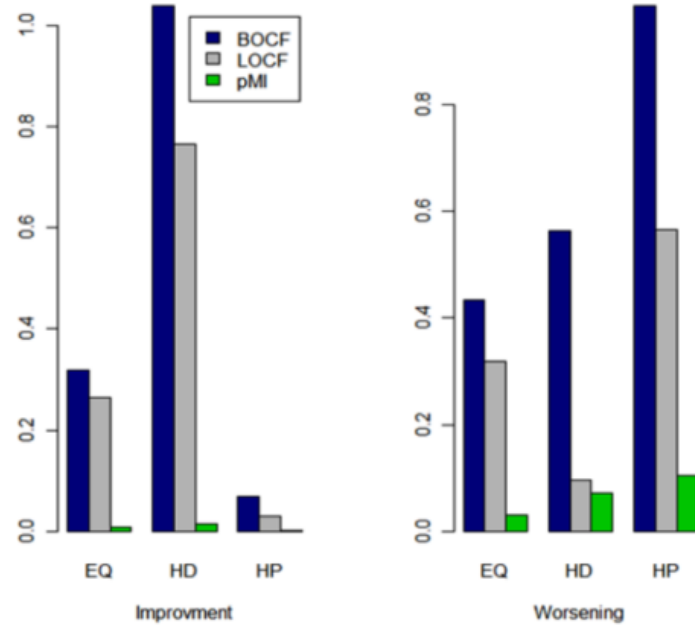


Figure 8.1: *Relative bias in estimates of the effectiveness estimand by direction of change and dropout pattern for the 0.3 effect size of the drug.*

The variance (results not shown) in estimates from LOCF and BOCF was lower than from DL or MI. The variance in estimates from pMI was generally intermediate to those from LOCF / BOCF versus DL and MI. The variance in estimates from pMI varied according to how much drug group data were replaced by placebo data; as the proportion of drug treated data being replaced by placebo increases the sample becomes more homogenous and variance in treatment contrasts decreases. In addition, the MSE for pMI was fairly consistent across scenarios and often smaller than the corresponding MSE for other methods. In contrast, MSEs from the other methods varied across scenarios and were often greater than the MSE from pMI.

Table 8.5: Mean square error in estimates of the effectiveness estimand for the analysis of change from baseline.

| Scenarios  |     |                 | Method |       |       |       |        |
|------------|-----|-----------------|--------|-------|-------|-------|--------|
| Trajectory | ES  | dropout pattern | BOCF   | LOCF  | DL    | MI    | pMI    |
| IMP        | 0.5 | =               | 0.886  | 0.856 | 1.137 | 1.185 | 0.479  |
|            |     | HD              | 1.838  | 1.189 | 1.754 | 1.781 | 0.400  |
|            |     | HP              | 0.334  | 0.458 | 0.907 | 0.960 | 0.543  |
| IMP        | 0.3 | =               | 0.477  | 0.544 | 0.926 | 0.982 | 0.444  |
|            |     | HD              | 1.334  | 0.901 | 1.092 | 1.138 | 0.372  |
|            |     | HP              | 0.275  | 0.326 | 0.761 | 0.791 | 0.526  |
| IMP        | 0   | =               | 0.313  | 0.424 | 0.691 | 0.717 | 0.433  |
|            |     | HD              | 0.731  | 0.583 | 0.702 | 0.735 | 0.376  |
|            |     | HP              | 0.706  | 0.543 | 0.687 | 0.724 | 0.528  |
| Worse      | 0.5 | =               | 0.933  | 0.635 | 1.337 | 1.382 | 0.489  |
|            |     | HD              | 0.275  | 0.282 | 1.749 | 1.789 | 0.430  |
|            |     | HP              | 2.826  | 1.339 | 0.957 | 0.976 | 70.542 |
| Worse      | 0.3 | =               | 0.469  | 0.411 | 0.903 | 0.903 | 0.452  |
|            |     | HD              | 0.506  | 0.281 | 1.090 | 1.118 | 0.385  |
|            |     | HP              | 1.851  | 0.815 | 0.771 | 0.788 | 0.542  |
| Worse      | 0   | =               | 0.228  | 0.275 | 0.675 | 0.692 | 0.441  |
|            |     | HD              | 0.900  | 0.341 | 0.686 | 0.727 | 0.380  |
|            |     | HP              | 0.900  | 0.358 | 0.682 | 0.703 | 0.522  |

The range in CI coverage (Table 8.6) for the effectiveness estimand ranged from 2% to 85% for BOCF, from 24% to 81% for LOCF, from 76% to 95% for DL, from 77% to 95% for MI, and from 98% to 99% for pMI.

Table 8.6: *Confidence interval coverage for the effectiveness estimand for the analysis of change from baseline*

| Scenarios  |     |                 | Method |      |      |      |      |
|------------|-----|-----------------|--------|------|------|------|------|
| Trajectory | ES  | dropout pattern | BOCF   | LOCF | DL   | MI   | pMI  |
| IMP        | 0.5 | =               | 51.0   | 59.0 | 88.3 | 89.3 | 99.1 |
|            |     | HD              | 17.4   | 38.3 | 76.1 | 78.7 | 99.3 |
|            |     | HP              | 79.1   | 72.5 | 91.7 | 91.8 | 98.1 |
| IMP        | 0.3 | =               | 72.2   | 73.2 | 92.0 | 91.8 | 99.4 |
|            |     | HD              | 28.2   | 50.5 | 87.9 | 89.2 | 99.5 |
|            |     | HP              | 84.1   | 80.4 | 93.1 | 94.3 | 97.7 |
| IMP        | 0   | =               | 79.7   | 76.6 | 94.4 | 94.9 | 99.4 |
|            |     | HD              | 52.5   | 63.3 | 95.5 | 94.9 | 99.5 |
|            |     | HP              | 50.5   | 63.6 | 95.5 | 94.9 | 97.5 |
| Worse      | 0.5 | =               | 32.6   | 54.9 | 85.0 | 85.4 | 98.8 |
|            |     | HD              | 77.9   | 80.0 | 76.0 | 76.9 | 99.0 |
|            |     | HP              | 1.9    | 24.7 | 90.6 | 91.3 | 98.2 |
| Worse      | 0.3 | =               | 62.1   | 69.9 | 91.3 | 92.5 | 98.6 |
|            |     | HD              | 59.3   | 81.1 | 86.6 | 88.7 | 99.5 |
|            |     | HP              | 7.3    | 45.9 | 93.4 | 93.6 | 97.8 |
| Worse      | 0   | =               | 84.8   | 80.9 | 94.9 | 94.3 | 99.0 |
|            |     | HD              | 35.8   | 75.4 | 95.3 | 95.0 | 99.7 |
|            |     | HP              | 36.5   | 75.1 | 95.2 | 94.9 | 97.5 |

Although pMI's simulation variance was smaller, its corresponding model-based variance obtained using Rubin's combination rules in PROC MIANALYZE was overestimating the "true" variance of PMI estimator. Table 8.7 depicts the ratio of simulation variance/Model based Variance. Unlike other models, pMI has larger model based variance. The ratio ranges between 0.401 and 0.778. Comparing the ratio among the three dropout levels, HD (Higher dropout in the drug group) exhibits the worst. The inflation of the model-based variance seems directly proportional to the proportion of drug treated data to be replaced by placebo data.

Table 8.7: *Ratio of Simulation Variance versus Model-Based Variance of the effectiveness estimand for the analysis of change from baseline*

| Scenarios  |     |                 | Method |       |       |       |       |
|------------|-----|-----------------|--------|-------|-------|-------|-------|
| Trajectory | ES  | dropout pattern | BOCF   | LOCF  | DL    | MI    | pMI   |
| IMP        | 0.5 | =               | 2.478  | 2.795 | 0.966 | 0.957 | 0.545 |
|            |     | HD              | 2.021  | 2.282 | 1.004 | 0.963 | 0.426 |
|            |     | HP              | 1.886  | 2.236 | 0.984 | 0.986 | 0.758 |
| IMP        | 0.3 | =               | 2.522  | 2.841 | 0.961 | 0.953 | 0.509 |
|            |     | HD              | 2.057  | 2.337 | 1.007 | 0.994 | 0.41  |
|            |     | HP              | 1.946  | 2.301 | 0.978 | 0.962 | 0.752 |
| IMP        | 0   | =               | 2.445  | 2.802 | 0.986 | 0.954 | 0.507 |
|            |     | HD              | 2.093  | 2.400 | 1.008 | 0.972 | 0.401 |
|            |     | HP              | 2.055  | 2.417 | 0.982 | 0.955 | 0.778 |
| Worse      | 0.5 | =               | 1.919  | 2.417 | 0.986 | 0.989 | 0.597 |
|            |     | HD              | 1.825  | 2.349 | 0.999 | 0.980 | 0.462 |
|            |     | HP              | 1.761  | 2.351 | 0.990 | 0.967 | 0.754 |
| Worse      | 0.3 | =               | 1.937  | 2.362 | 0.999 | 0.949 | 0.595 |
|            |     | HD              | 1.756  | 2.299 | 1.008 | 0.991 | 0.443 |
|            |     | HP              | 1.688  | 2.285 | 0.992 | 0.962 | 0.712 |
| Worse      | 0   | =               | 1.906  | 2.260 | 1.001 | 0.961 | 0.583 |
|            |     | HD              | 1.659  | 2.198 | 1.003 | 0.989 | 0.440 |
|            |     | HP              | 1.610  | 2.182 | 0.994 | 0.951 | 0.719 |

Power for pMI was close to the power from DL and MI when dropout rate was equal in the drug and control groups or when dropout was higher on placebo (HP); however, when dropout was higher on drug (HD) power from pMI was appreciable lower than from DL or MI. When ES =0, BOCF and LOCF provided the desired control of false positive (FP) results in only 2 of 6 scenarios, with at least triple the desired rate of FP (2.5%) in 4 of 6 scenarios, with maximum rates of 64% for BOCF and 34% for LOCF. DL and MI always provided the desired control. For pMI, the FP rate was always lower than the desired rate of 2.5%.

Table 8.8: *Rejection rates in assessing null hypothesis of no treatment difference for the analysis of change from baseline*

| Scenarios  |     |                 | Method |      |      |      |      |
|------------|-----|-----------------|--------|------|------|------|------|
| Trajectory | ES  | dropout pattern | BOCF   | LOCF | DL   | MI   | pMI  |
| IMP        | 0.5 | =               | 83.5   | 82.4 | 88.6 | 85.5 | 55.7 |
|            |     | HD              | 28.7   | 52.6 | 90.1 | 86.4 | 41.1 |
|            |     | HP              | 99.2   | 97.8 | 90.6 | 86.3 | 79.2 |
| IMP        | 0.3 | =               | 53.5   | 53.5 | 46.4 | 41.5 | 14.3 |
|            |     | HD              | 6.6    | 18.3 | 49.4 | 45.8 | 7.9  |
|            |     | HP              | 91.7   | 82.5 | 47.8 | 44.4 | 31.3 |
| IMP        | 0   | =               | 9.5    | 11.0 | 2.3  | 2.1  | 0.1  |
|            |     | HD              | 0.3    | 1.5  | 1.9  | 2.2  | 0.2  |
|            |     | HP              | 49.0   | 34.4 | 1.8  | 2.2  | 0.8  |
| Worse      | 0.5 | =               | 82.0   | 88.9 | 90.3 | 87.8 | 64.8 |
|            |     | HD              | 99.6   | 95.8 | 90.5 | 87.4 | 48.0 |
|            |     | HP              | 42.3   | 83.2 | 90.3 | 87.8 | 79.9 |
| Worse      | 0.3 | =               | 50.3   | 58.8 | 50.5 | 45.1 | 20.3 |
|            |     | HD              | 97.4   | 76.7 | 49.7 | 46.1 | 11.2 |
|            |     | HP              | 6.9    | 43.2 | 48.0 | 45.5 | 36.0 |
| Worse      | 0   | =               | 7.5    | 7.9  | 2.3  | 2.7  | 0.6  |
|            |     | HD              | 64.1   | 21.0 | 2.0  | 2.4  | 0.2  |
|            |     | HP              | 0.1    | 2.4  | 1.9  | 2.1  | 1.0  |

### 8.2.1.2 Efficacy

Tables A.1-A.4 summarize simulation results (bias, relative bias, MSE and CI coverage) from tests of the efficacy hypothesis. The Monte Carlo variance in the estimates and rejection rates apply to both efficacy and effectiveness as the same analysis is interpreted in two contexts that only vary by what is considered the true value for the treatment difference.

As expected, DL and MI provided unbiased estimates of efficacy with confidence interval (CI) coverage essentially equal to the nominal rates. In contrast BOCF and LOCF were biased, with the direction of bias varying by scenario, leading to poor CI



coverage and large MSE. pMI provided bias and MSE that are generally smaller than those in BOCF and LOCF. However, the CI coverage is greater than the nominal coverage rate in scenarios with no or smaller benefit of drug after discontinuation. In 17/18 scenarios the bias from pMI was conservative as the mean estimate of efficacy was smaller than the corresponding true value; in the 18th scenario pMI was essentially unbiased.

### 8.3 Analysis of the Depression Trial

Results from analysis of the depression trial data (Section 2.3) are summarized in Table 8.9. The mean change to the endpoint visit on placebo was approximately 8 points compared with approximately 10 points on drug. Therefore, as is typically the case in depression clinical trials, an appreciable placebo response was observed, thereby invalidating the assumptions for BOCF and LOCF.

The endpoint contrast from pMI was 1.54 compared with 1.08 from BOCF, 2.13 from DL, 2.09 from MI, and 1.75 from LOCF. Therefore, in the effectiveness context, the pMI result suggested that the effectiveness of the drug was approximately 73% the magnitude of the efficacy, as estimated by DL and MI.

Table 8.9: *Endpoint treatment contrasts by analytic method from the actual clinical trial dataset*

| Method | Endpoint Contrast | Standard error | P-value |
|--------|-------------------|----------------|---------|
| DL     | 2.13              | 0.97           | 0.030   |
| MI     | 2.09              | 0.92           | 0.023   |
| pMI    | 1.54              | 0.94           | 0.102   |
| LOCF   | 1.75              | 0.88           | 0.047   |
| BOCF   | 1.08              | 0.89           | 0.253   |

In the efficacy context, the pMI result can form the lower bound, or worst reasonable estimate of efficacy, to be combined with other sensitivity analyses to define a “region of ignorance”. That is, a region wherein the true value almost certainly lies, but exactly where is not certain.

The SE from pMI was slightly less than the SE from MI and DL and greater than the SE from BOCF and LOCF. Given that the CI coverage in the simulation results

for effectiveness was greater than the nominal coverage, the marginally significant  $p$  value from pMI is of less interest. However, it is relevant to note that the  $p$  value from pMI was considerably smaller than the  $p$  value from BOCF.

## 8.4 Discussion

To our knowledge, the simulation study in the present chapter is the first rigorous evaluation of pMI, a specific form of the jump to reference imputation approach detailed by Little and Yau (1996) and Carpenter, Roger and Kenward (2011). The pMI approach generally provided unbiased estimates of effectiveness in these simulations where there was no benefit from drug after discontinuation. However, the confidence interval coverage was consistently greater than the nominal coverage rate due to the upward bias of Rubin's variance estimator. In fact, Rubin's variance estimator is known for overestimating the true variance in some cases (Wang and Robins, 1998; Robins and Wang, 2000). Its performance is context-sensitive and one has to be careful, especially when the imputation and analysis models are discrepant. Overestimating the "true" variance of pMI estimator results in less efficient although still valid conservative confidence intervals, which might still be acceptable, given that in clinical research towards effectiveness/efficacy, one would rather err on the conservative side. Note that imputing missing values for both treated and untreated subjects using the same imputation model induces positive correlation between the estimated means in the two treatment groups. This result in smaller variance of the treatment contrast, however, is not captured by Rubin's model-based estimator. Hence, lower ratios of simulated variance versus model-based variance are observed for pMI estimates of treatment effect. This explains the relatively lower efficiency of pMI based inference, as indicated by larger coverage rates, compared to MI and DL methods.

In addition, pMI yielded conservative estimates of efficacy in all scenarios, whereas LOCF and BOCF were conservative in some scenarios and anti-conservative in others. As expected, DL and MI yielded unbiased estimates of efficacy and tended to overestimate effectiveness in those scenarios where a drug effect existed. However, in those scenarios where there was no drug effect, and therefore the true values for both efficacy and effectiveness were zero, DL and MI yielded unbiased estimates of efficacy and effectiveness.

These results should be viewed in light of the strengths and limitations of the present investigation. With the 18 scenarios, the simulation study was, on the one hand, comprehensive, but still narrow in scope relative to the vast array of clinical situations. Moreover, several implementations of pMI may be worth considering.

For example, we also considered an imputation approach similar to what Carpenter, Roger and Kenward (2011) refer to as “copy to reference”. In our implementation of copy to reference only baseline severity was used in the imputation model and all post-baseline data were replaced by imputed values for those patients that dropped out. Detailed results are not reported as this method did not provide unbiased or nearly unbiased effectiveness estimates in all scenarios. Independent replication of the simulation results and more experience with pMI in actual settings would be useful.

In the study, reasons for dropout were not differentiated as all dropouts resulted from the same model. In practice, it would be useful to separately consider dropouts and their impact by reasons for dropout. For example, it has been suggested that dropouts due to adverse events were the main area of concern and that methods like DL or MI provided reasonable estimates of effectiveness for other reasons of discontinuation (Kim, 2011). Therefore, rather than applying pMI to all dropouts, it may be useful to impute missing values from the placebo group only for drug treated patients that drop out due to adverse events. However, for initial assessments, the approach in the present work of applying pMI to all drug treated dropouts was useful in that it tested the method with high and differential rates of dropout, thereby allowing assessment of performance under extreme conditions.

Note that pMI, along with direct likelihood, multiple imputation, and their Bayesian and semi-parametric counterparts, is based on the assumption of MAR. This makes the latter methods very suitable as the preferred analysis for efficacy and the former for effectiveness. Of course, in all cases, it is possible for an MNAR mechanism to be operating. While contrasting pMI and the results of the other methods would then provide a partial response to the sensitivities engendered by this violation, extensive simulations under the assumption of MNAR would be needed, a subject of subsequent research.

Given these results, further investigation of pMI in scenarios not covered in the present work is warranted and use of pMI as an a priori specified sensitivity analysis in situations similar to those investigated in this study is justified.



## Chapter 9

# General Conclusions and Future Research

In applied research such as clinical trials, two general lines of attack have been employed to address the problem of incomplete data. The first is simply to design and carry out the study in a manner that limits the amount of incomplete data. Better implementation of a more appropriate design could reduce the frequency of missing values substantially. A variety of techniques for doing this has been proposed for clinical trials (National Research Council, 2010). The second line of attack for the treatment of incomplete data, and the focus of this thesis, is to apply analysis methods that exploit partial information in the observed data about the missing data to reduce the potential bias created by the missing data.

The area of missing data analysis has grown substantially over the past few decades. Concern has been raised about simple methods such as complete case (CC) analysis and last observation carried forward (LOCF) (Little and Rubin, 2002; Molenberghs and Kenward, 2007; Kenward and Molenberghs, 2009; National Research Council, 2010). Their use is decreasing and more principled, MAR-based methods increase in use; these include multiple imputation strategies Rubin (1987) and so-called direct-likelihood or direct Bayesian analysis. These are based on the property of *ignorability*, which ensures that such analyses are valid under MAR, supplemented with mild regularity conditions, even without explicitly modeling the missing data mechanism, provided that all incomplete sequences are subjected to analysis (Rubin, 1976; Little and Rubin, 2002; Molenberghs and Kenward, 2007; Fitzmaurice *et al.*, 2009). While *ignorability* would follow under likelihood inference, this is not generally true for non-likelihood approaches such as GEE and PL.

Likelihood methods enjoy many desirable properties, such as efficiency under appropriate regularity conditions and the ability to calculate functions of interest based on the proposed parametric model. However, for non-Gaussian outcomes in contrast, not only can the specification of the likelihood function be cumbersome, but also estimation of the parameters can be computationally intensive. In addition, fully specifying the joint probability model comes with the risk of possible misspecifications. Therefore, the difficulty in evaluating the likelihood for models with discrete correlated data has motivated alternative methods of estimation, the popular ones being GEE and PL.

While GEE methods replace score equations with alternative functions, in pseudo-likelihood, the likelihood itself is replaced by a more tractable expression. When attention is restricted to specification of the first moments (i.e., the mean structure) of the outcome sequence only, GEE leads to valid inferences by circumventing the need to address the association structure. Because of its frequentist nature, GEE in its basic form, as applied to incomplete data, is valid only under MCAR. To allow valid use of GEE under MAR, GEE has been extended to weighted generalized estimating equations (Robins, Rotnitzky and Zhao, 1995) and doubly robust GEE (Scharfstein, Rotnitzky and Robins, 1999; Bang and Robins, 2005; Tsiatis, 2006; Carpenter, Kenward and Vansteelandt, 2006; Molenberghs and Kenward, 2007; Rotnitzky, 2009; Birhanu *et al.*, 2011).

In contrast with GEE, PL methods can easily accommodate association (Yi, Zeng and Cook, 2011; He and Yi, 2011). Broadly speaking, one might consider marginal or conditional pseudo-likelihood. Pseudo-likelihood is closely related to but different from full likelihood and therefore not guaranteed to be valid under MAR, even though in some specific cases it might, because Rubin (1976) provided conditions for *ignorability* that are sufficient but not always necessary.

A substantial part of our work (Chapters 4, 5, 6 and 7) was devoted to the aforementioned alternatives to full likelihood, PL and GEE, with incomplete data. In their basic form, both GEE and PL, are valid only under the strongest MCAR mechanism. The aim of our work was to study in more depth the extension needed to ensure the validity of these methods under the less strong missing data mechanism, MAR. MCAR is a sufficient condition to the validity of GEE and PL. A number of extensions and modifications of GEE and PL, such as WGEE, MI-GEE, DR-GEE, and the singly robust and doubly robust version of PL are studied in the thesis. A brief overview of the resulting conclusions for the pertinent chapters is now presented.

In Chapter 4, we investigated and compared robustness of weighted GEE (WGEE), multiple imputation GEE (MI-GEE) and doubly robust GEE (DR-GEE). Advantages

and drawbacks of inverse probability weighting (IPW) methods with respect to multiple imputation have been the subject of some debate (Scharfstein, Rotnitzky and Robins, 1999; Clayton *et al.*, 1998; Carpenter, Kenward and Vansteelandt, 2006; Beunckens, Sotto and Molenberghs, 2008). Limitations of weighted GEE include: (a) the need to correctly specify the missingness model, and (b) potential instabilities associated with very large weights, leading to inefficient estimation and high variance in finite samples. Double robust (DR) estimators have the potential to alleviate both of these two limitations. In line with the literature, results from our extensive small-sample simulation studies corroborated the weakness of WGEE: WGEE is inefficient, especially for small sample sizes, and sensitive to misspecification in the dropout model. For WGEE, we have proposed a variance estimation that accounts for the variability in the weights (Birhanu *et al.*, 2011). MI-GEE proves to be robust under misspecification of the imputation model. The DR-GEE offer not only efficiency improvement over WGEE, but also bias protection against misspecification of the dropout model. Consequently, we advice to use MI-GEE and DR-GEE over WGEE in practice.

Although several authors have already used PL estimation (le Cessie and van Houwelingen, 1994; Geys, Molenberghs, and Lipsitz, 1998; Geys, Molenberghs, and Ryan, 1999; Aerts *et al.*, 2002), little work has been done for PL estimation with incomplete data. In this thesis, pseudo-likelihood methods were investigated in the marginal and conditional modeling of incomplete non-Gaussian and Gaussian data. While the numerical and computational issues accompanying the likelihood expressions of the models are circumvented by means of substituting pairwise pseudo-likelihood expressions for their full likelihood counterparts, the incompleteness in the data are addressed using inverse probability weighting ideas, thereby yielding singly and doubly robust estimators. This broadens the tool base for the obtaining models for incomplete data.

In Chapter 5, we have laid out a general framework for handling incomplete data, predominantly within the pseudo-likelihood setting, and formulated several alternative corrections employing inverse probability weighting ideas to ensure validity under MAR. These corrections follow both single and double robustness ideas, making use of inverse probability weighting (IPW), possibly supplemented with a predictive model for the unobserved outcomes given the observed ones, wherever appropriate. We developed the methodology, indicated how a conventional sandwich-type estimator can be used, and illustrated case applications for multivariate normal outcomes and a conditional pseudo-likelihood model for a binary outcome.

In Chapter 6, we extended the ideas in Molenberghs *et al.* (2011), using inverse

probability weighting and double robustness ideas (Scharfstein, Rotnitzky and Robins, 1999; Van der Laan and Robins, 2003; Bang and Robins, 2005; Rotnitzky, 2009). While Molenberghs *et al.* (2011) considered specific case studies involving Gaussian and exchangeable binary outcomes (Chapter 5), we supplemented the latter with a simulation study and considered more general types of correlation structure. Singly robust estimators with correctly specified dropout model and our doubly robust estimators without weights are at least as efficient as full likelihood. Furthermore, under full or near exchangeability, the naive available case version is as efficient as the doubly robust estimators. This is a very appealing property, because double robustness can be invoked without having to use weights or expectations.

In Chapter 7, the efficiency and robustness of various version of GEE and PL are investigated and compared focusing on marginal models for non-Gaussian longitudinal data with dropout. Inverse probability weighting-based, multiple imputation-based and doubly robust PL and GEE are compared under correctly specified, partially misspecified and severely misspecified dropout and imputation models. Compared to GEE, PL is as efficient as full likelihood, while the additional computational burden is minor. When the scientific interest lies in the estimation of the association parameter as well we advocate the use of PL.

Kang and Schafer (2007) showed empirically that there exist situations where severe biases may occur even when both weight and predictive models are only slightly misspecified. These authors also showed that widely varying weights are a potential risk for bias as well. This underscores that, like any tool in statistics, the user ought to be aware of the relative merits and advantages of the doubly robust method. In this respect, it is highly relevant that, in a number of settings we considered in Chapter 5 and Chapter 6, such as (5.33), the weights cancel from the estimating equations, thereby increasing robustness. Our doubly robust estimators, discussed in Chapter 5 and Chapter 6, are more efficient than the classical doubly robust estimators, because the models for the missingness mechanism (weight) cancels out.

In Chapter 8, we assessed the statistical performance of a method referred to as placebo multiple imputation (pMI) as an estimator of effectiveness and as a worst reasonable case sensitivity analysis in assessing efficacy. The pMI method assumes the statistical behavior of placebo- and drug-treated patients after drop out is the statistical behavior of placebo-treated patients. Thus, in the effectiveness context pMI assumes no pharmacological benefit of the drug after dropout. In a simulation study with 18 scenarios the pMI approach generally provided unbiased estimates of effectiveness and conservative estimates of efficacy. However, the confidence interval coverage was consistently greater than the nominal coverage rate. In contrast, LOCF



and BOCF were conservative in some scenarios and anti-conservative in others with respect to efficacy and effectiveness. As expected, direct likelihood (DL) and standard multiple imputation (MI) yielded unbiased estimates of efficacy and tended to over-estimate effectiveness in those scenarios where a drug effect existed. However, in scenarios with no drug effect, meaning the true values for both efficacy and effectiveness were zero, DL and MI yielded unbiased estimates of efficacy and effectiveness.

Our work has concentrated on a limited, yet important, aspect of incomplete data modeling however, and there are a number of issues not covered in this thesis and need further examination if we wish to use our PL-based doubly robust estimation equation routinely. These are briefly discussed below.

While our methodologies are general, their implementation for general missing data patterns is more complicated than when missingness is confined to dropout, or arises in a clustered-data setting. Robust methodologies, like those discussed in the thesis could be used and further research on extending the discussed PL-based estimating equations to non-monotone case are recommended. While our work mostly focuses on the MAR setting, in practice one cannot rule out the possibility of an MNAR mechanism. Furthermore, even when MAR is deemed plausible, it is of interest to conduct some form of sensitivity analysis (Molenberghs and Kenward, 2007; Fitzmaurice *et al.*, 2009). Obvious routes include: (1) the extension of Pseudo-likelihood (PL) to MNAR, advocated by Parzen *et al.* (2006); and (2) the extension of our results along the lines of Vansteelandt, Rotnitzky, and Robins (2007).

The advantage of a variety of proposals (singly robust and doubly robust) is that the user has freedom of selection. Of course, more work is needed to provide further guidance toward such a choice. In Chapter 5, we have indicated, for some specific cases, how standard errors can be derived. It is important to consider methods that do not involve tedious analytical considerations, such as, for example, the jackknife-based method of Heagerty and Lele (1998) need to be undertaken for a variety of other choices.



# Bibliography

- Aerts, M., Geys, H., Molenberghs, G. and Ryan, L.M. (2002). *Topics in Modelling of Clustered Data*. London: Chapman & Hall/CRC.
- Arnold, B.C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhya B*, **53**, 233-243.
- Bahadur, R.R., (1961). *A representation of the joint distribution of responses to  $n$  dichotomous items. In: Studies in Item Analysis and Prediction, H. Solomon (Ed.). Stanford Mathematical Studies in the Social Sciences VI*. Stanford, CA: Stanford University Press.
- Bang, H. and Robins, J.M.(2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**, 962–972.
- Beunckens, C., Sotto, C. and Molenberghs, G. (2008). A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal data. *Computational Statistics and Data Analysis*, **52**, 1533–1548.
- Birhanu, T., Lipkovich, I, Molenberghs, G, and Mallinckrodt, H. C.(2013). A Multiple Imputation Based Approach to Sensitivity Analyses and Effectiveness Assessments in Longitudinal Clinical Trials. *Journal of Biopharmaceutical Statistics*, accepted.
- Birhanu, T., Molenberghs, G., Sotto, C. and Kenward, M.G. (2011). Doubly robust and multiple-imputation-based generalized estimating equations. *Journal of Biopharmaceutical Statistics*, **21**, 202–225.
- Birhanu, T., Sotto, C., Molenberghs, G., Kenward, M.G. and Verbeke, G. (2012). Doubly Robust Composite Likelihood for Hierarchical Categorical Data. *submitted for publication*.
- Carey, V.C., Zeger, S.L. and Diggle, P.J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**, 517–526.

- Carpenter, J.R., Kenward, M.G. and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Royal Statistics Society Series A* **3**, 571–584.
- Carpenter, J.R, Roger, J, and Kenward, M.G. (2011). Analysis of Longitudinal Trials with Missing Data: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Submitted for publication*.
- Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998). Analysis of longitudinal binary data from multi-phase sampling (with discussion). *Royal Statistics Society Series B* **60**, 71–87.
- Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.
- Dale, J.R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.
- Daniel, R.M. (2008). On aspects of robustness and sensitivity in missing data methods. *Unpublished PhD thesis*. London School of Hygiene and Tropical Medicine, UK.
- De Backer, M., De Keyser, P., De Vroey, C. and Lesaffre, E. (1996). A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250mg/day vs. itraconazole 200mg/day—a double-blind comparative trial. *British Journal of Dermatology*, **134**, 16–17.
- Dempster, A.P., Laird, N.M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.
- Diggle, P.J., Heagerty, P., Liang, K-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. New York: Oxford University Press.
- Faes, C., Aerts, M., Molenberghs, G., Geys, H., Teuns, G. and Bijneens, L. (2008). A high-dimensional joint model for longitudinal outcomes of different nature. *Statist. Med.*, **27**, 4408–4427.
- Fahrmeir, L. and Tutz, G. (2002). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Heidelberg: Springer.
- Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2009). *Longitudinal Data Analysis*. Boca Raton: CRC/ Chapman & Hall

- Fleming, T.R. (2011). Addressing missing data in clinical trials. *Annals of Internal Medicine* **154**,113–117.
- Geys, H., Molenberghs, G. and Lipsitz, S.R. (1998). A note on the comparison of pseudo-likelihood and generalized estimating equations for marginal odds ratio models. *Journal of Statistical Computation and Simulation*, **62**, 45-72.
- Geys, H., Molenberghs, G. and Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, **94**, 734-745.
- Hartley, H.O. and Hocking, R. (1971). The analysis of incomplete data. *Biometrics*, **27**, 7783–808.
- He, W. and Yi, G.Y. (2011). A pairwise likelihood method for correlated binary data with/without missing observations under generalized partially linear single-index models. *Statistica Sinica*, **21**, 207–229.
- Heagerty, P.J. and Lele, S.R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, **93**, 1099–1111.
- Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G. and Mallinckrodt, C. (2006). Analyzing incomplete binary longitudinal clinical trial data. *Statistical Science*, **21**, 52–69.
- 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 discussions). *Statist Sci*, **22**, 523–539.
- Kenward, M.G. and Molenberghs, G. (2009). Last observation carried forward: a crystal ball? *Journal of Biopharmaceutical Statistics*, **19**, 872–888.
- Kim, Y.(2011) Missing data handling in chronic pain trials. *Journal of Biopharmaceutical Statistics*, **21(2)**, 311–325.
- Kleinman, J. (1973). Proportions with extraneous variance: single and independent samples. *Journal of the American Statistical Association*, **68**, 46–54.
- Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

- Lang, J.B. and Agresti, A. (1994). Simultaneously modelling joint and marginal distributions of multivariate categorical responses. *Journal of American Statistical Association*, **89**, 625–632.
- le Cessie, S. and van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Applied Statistics*, **43**, 95–108.
- Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
- Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**, 3–40.
- Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.
- Little, R.J.A. (1994a). A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–483.
- Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.
- Little, R.J.A. and Yau, L. (1996). Intent-to-Treat Analysis in Longitudinal Studies with Drop-Outs. *Biometrics*, **52**, 1324–1333.
- Liu and Gould (2002). Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *Journal of Biopharmaceutical Statistics*, **12 (2)**, 207–226.
- Mallinckrodt, C.H. and Kenward, M.G. (2009). Conceptual considerations regarding choice of endpoints, hypotheses, and analyses in longitudinal clinical trials. *Drug Information Journal*, **43**, 449–458.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman & Hall.
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, **9**, 538–558.
- Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons.
- Molenberghs, G., Kenward, M.G. and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with non-random dropout. *Biometrika* **84**, 33–44.

- Molenberghs, G., Kenward, M.G., Verbeke, G. and Birhanu, T. (2011). Pseudo-likelihood estimation for incomplete data. *Statistica Sinica*, **21**, 187–206.
- Molenberghs, G. and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of American Statistical Association*, **89**, 633–644.
- Molenberghs, G. and Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine*, **18**, 2237–2255.
- Molenberghs, G. and Ryan, L.M. (1999). Likelihood inference for clustered multivariate binary data. *Environmetrics*, **10**, 279–300.
- 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. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.
- Molenberghs, G., Verbeke, G., and Beunckens, C. (2008). Formal and informal model selection with incomplete data. *Statistical Science*, **23**, 201–218.
- National Research Council. (2010). *The prevention and Treatment of Missing Data in Clinical Trials*. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of American Statistical Association* **92**, 1320-1329.
- Parzen, M., Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J.G. and Troxel, A. (2006). Pseudo-likelihood methods for longitudinal binary data with non-ignorable missing responses and covariates. *Statistics in Medicine*, **25**, 2784-2796.
- Permutt, T and Pinheiro, J.(2009). Dealing with the missing data challenge in clinical trials. *Drug Information Journal*, **43(4)**, 403-408.
- Plackett, R.L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association*, **60**, 516–522.

- Renard, D., Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, **44**, 649-667.
- Roberts, D.T. (1992). Prevalence of dermatophyte onychomycosis in the United Kingdom: Results of an omnibus survey. *British Journal of Dermatology*, **126 Suppl. 39**, 23-27.
- 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**, 106-121.
- Robins, J.M. and Wang, N. (2000). Inference for imputation estimators, *Biometrika* **87**, 113-124.
- Rotnitzky, A. (2009). Inverse probability weighted methods. In: *Longitudinal Data Analysis*. G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs (eds.). Boca Raton: CRC/ Chapman & Hall, pp. 453-476.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.
- Rubin, D.B. (1978). Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. In: *Imputation and Editing of Faulty or Missing Survey Data*. Washington, DC: U.S. Department of Commerce, pp. 1-23.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Schafer, J.L. (1997). *em Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Schafer, J.L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**, 3-15.
- Schafer, J. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, **57**, 19-35.
- Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semi-parametric nonresponse models. *Journal of the American Statistical Association*, **94**, 1096-1146 (with discussion).



- Skellam, J.G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, **10**, 257–261.
- Tsiatis, Anastasios A. (2006). *Semiparametric theory and missing data*. New York: Springer.
- Tyl, R.W., Price, C.J., Marr, M.C., and Kimmel, C.A. (1988). Developmental toxicity evaluation of dietary di(2-ethylhexyl)phthalate in Fischer 344 rats and CD-1 mice. *Fundamental and Applied Toxicology*, **10**, 395–412.
- Van der Laan, M.J. and Robins, J.M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.
- Vansteelandt, S., Rotnitzky, A., and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone non-response. *Biometrika*, **94**, 841–860.
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.
- Wang, N. and Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, **85**, 935–948.
- Windholz, M. (1983). *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals (10th ed.)* Rahway, NJ: Merck and Co.
- Wu, M.C. and Bailey, K.R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, **45**, 939–955.
- 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**, 175–188.
- Yi, G.Y., Zeng, L. and Cook, R.J. (2011). A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. *The Canadian Journal of Statistics*, **39**, 34–51.
- Zhao, L.P. and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–648.



## Appendix A

# Simulation Results for Efficacy Estimand

In this appendix, we present tables for the simulation studies discussed in Section 8.2.1.2 of Chapter 8. Table A.1 present bias from test of efficacy estimand for the analysis of change from baseline. Table A.2 depicts relative bias in estimates of the efficacy estimand for the analysis of change from baseline. Mean Square Error in estimates of the efficacy estimand for the analysis of change from baseline are shown in Table A.3 while results for the Confidence Interval (CI) coverage in estimates of the efficacy estimand for the analysis of change from baseline are given in Table A.4.

Table A.1: *Bias in estimates of the efficacy estimand for the analysis of change from baseline*

| Scenarios  |     |                 | Method |        |       |       |        |
|------------|-----|-----------------|--------|--------|-------|-------|--------|
| Trajectory | ES  | dropout pattern | BOCF   | LOCF   | DL    | MI    | pMI    |
| IMP        | 0.5 | =               | 1.431  | 1.329  | 0.043 | 0.045 | 0.764  |
|            |     | HD              | 2.314  | 1.995  | 0.039 | 0.039 | 0.974  |
|            |     | HP              | 0.767  | 0.877  | 0.039 | 0.041 | 0.477  |
| IMP        | 0.3 | =               | 0.876  | 0.815  | 0.044 | 0.047 | 0.524  |
|            |     | HD              | 1.690  | 1.422  | 0.038 | 0.041 | 0.650  |
|            |     | HP              | 0.228  | 0.362  | 0.037 | 0.042 | 0.323  |
| IMP        | 0   | =               | 0.012  | 0.018  | 0.038 | 0.044 | 0.105  |
|            |     | HD              | 0.681  | 0.512  | 0.041 | 0.038 | 0.132  |
|            |     | HP              | -0.665 | -0.470 | 0.037 | 0.036 | 0.067  |
| Worse      | 0.5 | =               | 1.694  | 1.442  | 0.034 | 0.035 | 0.706  |
|            |     | HD              | 0.813  | 1.167  | 0.043 | 0.042 | 0.899  |
|            |     | HP              | 2.193  | 1.605  | 0.045 | 0.051 | 0.496  |
| Worse      | 0.3 | =               | 0.998  | 0.865  | 0.027 | 0.037 | 0.463  |
|            |     | HD              | 0.122  | 0.579  | 0.041 | 0.043 | 0.604  |
|            |     | HP              | 1.632  | 1.084  | 0.044 | 0.052 | 0.207  |
| Worse      | 0   | =               | 0.008  | 0.022  | 0.033 | 0.034 | 0.115  |
|            |     | HD              | -0.836 | -0.269 | 0.042 | 0.037 | 0.123  |
|            |     | HP              | 0.839  | 0.302  | 0.046 | 0.045 | -0.028 |

Table A.2: *Relative bias in estimates of the efficacy estimand for the analysis of change from baseline*

| Scenarios  |     |                 | Method |       |       |       |       |
|------------|-----|-----------------|--------|-------|-------|-------|-------|
| Trajectory | ES  | dropout pattern | BOCF   | LOCF  | DL    | MI    | pMI   |
| IMP        | 0.5 | =               | 0.519  | 0.483 | 0.016 | 0.016 | 0.277 |
|            |     | HD              | 0.840  | 0.724 | 0.014 | 0.014 | 0.354 |
|            |     | HP              | 0.278  | 0.318 | 0.014 | 0.015 | 0.173 |
| IMP        | 0.3 | =               | 0.530  | 0.493 | 0.027 | 0.028 | 0.317 |
|            |     | HD              | 1.023  | 0.860 | 0.023 | 0.025 | 0.393 |
|            |     | HP              | 0.138  | 0.219 | 0.022 | 0.025 | 0.195 |
| Worse      | 0.5 | =               | 0.615  | 0.523 | 0.012 | 0.013 | 0.256 |
|            |     | HD              | 0.295  | 0.424 | 0.016 | 0.015 | 0.326 |
|            |     | HP              | 0.796  | 0.583 | 0.016 | 0.018 | 0.180 |
| Worse      | 0.3 | =               | 0.604  | 0.524 | 0.017 | 0.022 | 0.280 |
|            |     | HD              | 0.074  | 0.351 | 0.025 | 0.026 | 0.366 |
|            |     | HP              | 0.987  | 0.656 | 0.027 | 0.032 | 0.126 |

Table A.3: Mean square error in estimates of the effectiveness estimand for the analysis of change from baseline

| Scenarios  |     |                 | Method |       |       |       |       |
|------------|-----|-----------------|--------|-------|-------|-------|-------|
| Trajectory | ES  | dropout pattern | BOCF   | LOCF  | DL    | MI    | pMI   |
| IMP        | 0.5 | =               | 2.406  | 2.232 | 0.701 | 0.751 | 1.059 |
|            |     | HD              | 5.646  | 4.314 | 0.693 | 0.720 | 1.340 |
|            |     | HP              | 0.858  | 1.095 | 0.684 | 0.738 | 0.769 |
| IMP        | 0.3 | =               | 1.114  | 1.117 | 0.707 | 0.765 | 0.719 |
|            |     | HD              | 3.139  | 2.349 | 0.701 | 0.749 | 0.794 |
|            |     | HP              | 0.318  | 0.455 | 0.681 | 0.715 | 0.630 |
| IMP        | 0   | =               | 0.313  | 0.424 | 0.691 | 0.717 | 0.433 |
|            |     | HD              | 0.731  | 0.583 | 0.702 | 0.735 | 0.376 |
|            |     | HP              | 0.706  | 0.543 | 0.687 | 0.724 | 0.528 |
| Worse      | 0.5 | =               | 3.077  | 2.354 | 0.685 | 0.732 | 0.968 |
|            |     | HD              | 0.866  | 1.636 | 0.680 | 0.717 | 1.207 |
|            |     | HP              | 5.007  | 2.847 | 0.681 | 0.707 | 0.782 |
| Worse      | 0.3 | =               | 1.216  | 1.025 | 0.681 | 0.691 | 0.665 |
|            |     | HD              | 0.217  | 0.607 | 0.692 | 0.724 | 0.745 |
|            |     | HP              | 2.857  | 1.443 | 0.682 | 0.705 | 0.566 |
| Worse      | 0   | =               | 0.228  | 0.275 | 0.675 | 0.692 | 0.441 |
|            |     | HD              | 0.900  | 0.341 | 0.686 | 0.727 | 0.380 |
|            |     | HP              | 0.900  | 0.358 | 0.682 | 0.703 | 0.522 |

Table A.4: Confidence Interval coverage in estimates of the efficacy estimand for the analysis of change from baseline

| Scenarios  |     |                 | Method |      |      |      |      |
|------------|-----|-----------------|--------|------|------|------|------|
| Trajectory | ES  | dropout pattern | BOCF   | LOCF | DL   | MI   | pMI  |
| IMP        | 0.5 | =               | 12.2   | 22.7 | 94.2 | 94.9 | 94.5 |
|            |     | HD              | 0.2    | 1.3  | 95.2 | 95.3 | 93.4 |
|            |     | HP              | 49.3   | 42.7 | 95.4 | 95.2 | 94.0 |
| IMP        | 0.3 | =               | 40.1   | 47.0 | 94.8 | 94.4 | 97.4 |
|            |     | HD              | 3.9    | 10.7 | 95.4 | 94.7 | 98.5 |
|            |     | HP              | 79.2   | 71.8 | 96.0 | 94.6 | 96.1 |
| IMP        | 0   | =               | 79.7   | 76.6 | 94.4 | 94.9 | 99.4 |
|            |     | HD              | 52.5   | 63.3 | 95.5 | 94.9 | 99.5 |
|            |     | HP              | 50.5   | 63.6 | 95.5 | 94.9 | 97.5 |
| Worse      | 0.5 | =               | 0.6    | 6.1  | 94.6 | 94.3 | 93.7 |
|            |     | HD              | 36.3   | 17.1 | 95.2 | 95.2 | 93.3 |
|            |     | HP              | 0.1    | 2.8  | 95.3 | 95.5 | 93.9 |
| Worse      | 0.3 | =               | 23.2   | 36.8 | 95.0 | 94.4 | 96.6 |
|            |     | HD              | 84.7   | 57.6 | 95.2 | 95.2 | 97.7 |
|            |     | HP              | 1.9    | 21.5 | 95.3 | 95.6 | 96.6 |
| Worse      | 0   | =               | 84.8   | 80.9 | 94.9 | 94.3 | 99.0 |
|            |     | HD              | 35.8   | 75.4 | 95.3 | 95.0 | 99.7 |
|            |     | HP              | 36.5   | 75.1 | 95.2 | 94.9 | 97.5 |







