Statistical Methods for Incomplete Covariates and Two-Phase Designs

by

Michael McIsaac

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2012

© Michael McIsaac 2012

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Incomplete data is a pervasive problem in health research, and as a result statistical methods enabling inference based on partial information play a critical role. This thesis explores estimation of regression coefficients and associated inferences when variables are incompletely observed. In the later chapters, we focus primarily on settings with incomplete covariate data which arise by design, as in studies with two-phase sampling schemes, as opposed to incomplete data which arise due to events beyond the control of the scientist. We consider the problem in which "inexpensive" auxiliary information can be used to inform the selection of individuals for collection of data on the "expensive" covariate. In particular, we explore how parameter estimation relates to the choice of sampling scheme. Efficient sampling designs are defined by choosing the optimal sampling criteria within a particular class of selection models under a two-phase framework. We compare the efficiency of these optimal designs to simple random sampling and balanced sampling designs under a variety of frameworks for inference.

As a prelude to the work on two-phase designs, we first review and study issues related to incomplete data arising due to chance. In Chapter 2, we discuss several models by which missing data can arise, with an emphasis on issues in clinical trials. The likelihood function is used as a basis for discussing different missing data mechanisms for incomplete responses in short-term and longitudinal studies, as well as for missing covariates. We briefly discuss common *ad hoc* strategies for dealing with incomplete data, such as complete-case analyses and naive methods of imputation, and we review more broadly appropriate approaches for dealing with incomplete data in terms of asymptotic and empirical frequency properties. These methods include the EM algorithm, multiple imputation, and inverse probability weighted estimating equations. Simulation studies are reported which demonstrate how to implement these procedures and examine performance empirically. We further explore the asymptotic bias of these estimators when the nature of the missing data mechanism is misspecified. We consider specific types of model misspecification in methods designed to account for the missingness and compare the limiting values of the resulting estimators.

In Chapter 3, we focus on methods for two-phase studies in which covariates are incomplete by design. In the second phase of the two-phase study, subject to correct specification of key models, optimal sub-sampling probabilities can be chosen to minimise the asymptotic variance of the resulting estimator. These optimal phase-II sampling designs are derived and the empirical and asymptotic relative efficiencies resulting from these designs are compared to those from simple random sampling and balanced sampling designs. We further examine the effect on efficiency of utilising external pilot data to estimate parameters needed for derivation of optimal designs, and we explore the sensitivity of these optimal sampling designs to misspecification of preliminary parameter estimates and to the misspecification of the covariate model at the design stage. Designs which are optimal for analyses based on inverse probability weighted estimating equations are shown to result in efficiency gains for several different methods of analysis and are shown to

be relatively robust to misspecification of the parameters or models used to derive the optimal designs. Furthermore, these optimal designs for inverse probability weighted estimating equations are shown to be well behaved when necessary design parameters are estimated using relatively small external pilot studies. We also consider efficient two-phase designs explicitly in the context of studies involving clustered and longitudinal responses. Model-based methods are discussed for estimation and inference. Asymptotic results are used to derive optimal sampling designs and the relative efficiencies of these optimal designs are again compared with simple random sampling and balanced sampling designs. In this more complex setting, balanced sampling designs are demonstrated to be inefficient and it is not obvious when balanced sampling will offer greater efficiency than a simple random sampling design. We explore the relative efficiency of phase-II sampling designs based on increasing amounts of information in the longitudinal responses and show that the balanced design may become less efficient when more data is available at the design stage. In contrast, the optimal design is able to exploit additional information to increase efficiency whenever more data is available at phase-I.

In Chapter 4, we consider an innovative adaptive two-phase design which breaks the phase-II sampling into a phase-IIa sample obtained by a balanced or proportional sampling strategy, and a phase-IIb sample collected according to an optimal sampling design based on the data in phases I and IIa. This approach exploits the previously established robustness of optimal inverse probability weighted designs to overcome the difficulties associated with the fact that derivations of optimal designs require *a priori* knowledge of parameters. The efficiency of this hybrid design is compared to those of the proportional and balanced sampling designs, and to the efficiency of the true optimal design, in a variety of settings. The efficiency gains of this adaptive two-phase design are particularly apparent in the setting involving clustered response data, and it is natural to consider this approach in settings with complex models for which it is difficult to even speculate on suitable parameter values at the design stage.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Richard J. Cook, for his immeasurable support. His mentorship extended beyond the composition of this thesis; Dr. Cook's insight and guidance were instrumental in the creation of this doctoral thesis, in the completion of my master's research essay, and in my preparation for a life in academia. Thank you for all of your time and advice.

I would like to thank the members of my examining committee – Drs. Jerry Lawless, Grace Yi, Scott Leatherdale, and Shelley Bull – for their time and for their valuable comments on this thesis. I am especially grateful to Drs. Lawless and Yi for being part of my stage II comprehensive examination committee and for providing me with important feedback as my thesis progressed.

I am also grateful to the Natural Sciences and Engineering Council of Canada (NSERC) for their financial support. This research was completed because of the funding provided by an Alexander Graham Bell Canada Graduate Scholarship and research studentships with Dr. Cook; thank you both.

Finally, I would like to thank my family for a lifetime of support. In particular, I would like to thank my wife, Emily, for being understanding, encouraging, and generally amazing. This thesis is dedicated to you.

Contents

List of Tables **xi**

List of Figures **xiii**

1 Introduction **1**

1.1 Statistical Methods for Incomplete Data 1

1.1.1 Terminology for Incomplete Data 1

1.1.2 Methods Based on Imputation 2

1.1.3 Inverse Probability Weighting 3

1.2 Incomplete Data Arising by Two-Phase Designs 4

1.3 Motivating Settings and Studies 6

1.3.1 Incomplete Data in Randomised Dermatology Trials 6

1.3.2 Biomarker Studies in Psoriatic Arthritis 9

1.3.3 Canadian Longitudinal Study on Aging 11

1.4 Outline of Thesis Research 12

1.4.1 Methods for Data Missing by Happenstance 12

1.4.2 Two-Phase Sampling Design 13

1.4.3 Adaptive Stratified Two-Phase Sampling 14

2 Statistical Methods for Data Missing By Happenstance **15**

2.1 Introduction 15

2.2 Methods for Incomplete Binary Responses in Randomised Trials 16

2.2.1 Models and Measures of Treatment Effect 16
2.2.2 Parameter Estimation with Incomplete Response Data 18
2.2.3 An Illustrative Simulation Study 23
2.2.4 Further Remarks 25
2.3 Methods for the Analysis of Incomplete Longitudinal Data 26
2.3.1 Notation and Terminology 26
2.3.2 Likelihood-Based Methods of Estimation and Inference 27
2.3.3 Generalised Estimating Equations 29
2.3.4 Naive Methods of Imputation for Incomplete Longitudinal Data 31
2.4 Methods for Dealing with Incomplete Covariate Data in Observational Studies . . 31
2.4.1 Likelihood Analyses 31
2.4.2 An EM Algorithm 32
2.4.3 Multiple Imputation with Missing Covariates 33
2.4.4 Inverse Probability Weighted Estimating Functions 34
2.4.5 An Illustrative Simulation Study 34
2.4.6 A Note on Generalisability 38
2.5 Issues of Model Misspecification with Incomplete Data 38
2.5.1 Asymptotic Behaviour of Estimators from Estimating Functions 39
2.5.2 Misspecified Models for Handling Incomplete Responses 40
2.5.3 Misspecified Models for Incomplete Covariate Data 48
2.6 Discussion 51

3 Two-Phase Sampling Designs **55**
3.1 Introduction 55
3.1.1 Study Design under Two-Phase Sampling Schemes 55
3.2 Design Of Studies With Two-Phase Sampling 57
3.3 Frameworks For Analysis 59
3.3.1 Maximum Likelihood Estimation with Complete Data 59
3.3.2 Maximum Likelihood With Covariates Missing At Random 59

3.3.3 Inverse Probability Weighted Estimating Equations 61
3.3.4 Inverse Probability Weighted Estimating Equations With Estimated Weights 62
3.3.5 Augmented Inverse Probability Weighted Estimating Equations 63
3.4 Relative Efficiency Of Phase-Two Designs 65
3.4.1 Derivation of Phase-Two Sampling Designs 65
3.4.2 Empirical Properties of Phase-Two Designs 68
3.5 Misspecification in Optimal Two-Phase Designs 73
3.5.1 Sensitivity Analyses of Optimal Designs with Binary X 75
3.5.2 Sensitivity Analyses of Optimal Designs with Continuous X 82
3.6 Recommendations for Selection of PsA Patients 84
3.7 Response-Dependent Sampling with Clustered Binary Data 88
3.7.1 The Response Model for Clustered Data 88
3.7.2 The Selection Model 88
3.7.3 Asymptotic Relative Efficiencies 90
3.8 Response-Dependent Sampling with Longitudinal Binary Data 93
3.8.1 The Response Model for Longitudinal Data 93
3.8.2 The Selection Model 93
3.8.3 Asymptotic Relative Efficiencies 94
3.9 Discussion 95

4 Adaptive Stratified Two-Phase Sampling **100**
4.1 Inverse Probability Weighting with Basic Stratified Sampling 100
4.1.1 Asymptotic Variance of IPW under BSS 101
4.2 IPW Analysis for Multi-Phase Sampling 102
4.3 Adaptive Multi-Phase Sampling 103
4.4 Empirical Properties of Adaptive Multi-phase Designs 105
4.4.1 Simulation Results for a Binary X 105
4.4.2 Simulation Results for a Continuous X 112
4.4.3 Simulation Results for Clustered Responses 117
4.5 Considerations of Studies Involving Greater Numbers of Stages 127
4.6 Conclusion 127

5 Future Work **131**

5.1 Implications of Research Findings 131

5.2 Ongoing and Future Research Plans 133

5.2.1 Repeated Measure Analyses of Longitudinal Data 133

5.2.2 Current Status Failure Time Data 136

5.2.3 Dealing with Data Missing by Design and Chance 139

Bibliography **141**

List of Tables

1.1 Summary of rates of missing data and strategies employed for dealing with missing data in examples of phase 3 randomised controlled trials in psoriasis 8

2.1 Simulation results of naive and adjusted analyses using inverse weighting and multiple imputation with incomplete response data 24

2.2 Simulation results of naive and adjusted analyses using inverse weighting, EM, and multiple imputation to account for an ignorably incomplete covariate X . . . 36

2.3 Simulation results of naive and adjusted analyses using inverse weighting, EM, and multiple imputation to account for a non-ignorably incomplete covariate X . 37

3.1 Simulation results for four different phase-II sampling designs with binary covariates and response 71

3.2 Simulation results for different phase-II sampling designs with $X|V$ following a gamma distribution 74

3.3 Parameter estimates from the PsA pilot data and the associated optimal designs . 76

3.4 Asymptotic relative efficiencies from potentially misspecified optimal designs . . 78

3.5 Simulation results for different phase-II sampling designs with $X|V$ following a log-normal distribution, but Opt_{ipw} derived using a misspecified covariate distribution 83

4.1 Simulation results for adaptive phase-II sampling designs with binary covariates and response when $n = 250$ and $N = 1000$ 110

4.2 Simulation results for adaptive phase-II sampling designs with binary covariates and response when $n = 500$ and $N = 1000$ 111

4.3 Simulation results for adaptive phase-II sampling designs with a continuous expensive covariate when $n = 250$ and $N = 1000$ 115

4.4 Simulation results for adaptive phase-II sampling designs with a continuous expensive covariate when $n = 500$ and $N = 1000$ 116

4.5 Simulation results for adaptive phase-II sampling designs with clustered responses when $\psi = 1.25$, $n = 250$, and $N = 1000$ 123

4.6 Simulation results for adaptive phase-II sampling designs with clustered responses when $\psi = 1.25$, $n = 500$, and $N = 1000$ 124

4.7 Simulation results for adaptive phase-II sampling designs with clustered responses when $\psi = 3.00$, $n = 250$, and $N = 1000$ 125

4.8 Simulation results for adaptive phase-II sampling designs with clustered responses when $\psi = 3.00$, $n = 500$, and $N = 1000$ 126

4.9 Simulation results for multi-stage adaptive phase-II sampling designs with binary covariates and response when $n = 250$ and $N = 1000$ 130

List of Figures

1.1 Timeline for collection of information on disease progression over a fixed follow-up in the Psoriatic Arthritis setting with a univariate response 10

1.2 Timeline for collection of information on disease progression over a fixed follow-up in the Psoriatic Arthritis setting with clustered responses 11

1.3 Timeline for collection of longitudinal information on disease progression 12

2.1 Limiting values of naive complete-case estimators of relative risk and odds ratio . 20

2.2 Schematic of schedule of assessments in longitudinal study with $K = 5$ for an individual with $K_i = 3$ 27

2.3 Limiting values of estimators of relative risk from analyses with misspecified models when responses are incomplete 48

2.4 Limiting values of estimators of odds ratios from analyses with misspecified models when responses are incomplete 49

2.5 Limiting values of estimators of relative risk from analyses with misspecified models when covariates are incomplete 52

2.6 Limiting values of estimators of odds ratios from analyses with misspecified models when covariates are incomplete 53

3.1 Asymptotic relative efficiency of estimators for α_x under different designs as a function of the phase-II sampling fraction 77

3.2 Boxplots of optimal selection probabilities derived using parameter estimates from pilot studies with binary covariates and response 80

3.3 Asymptotic relative efficiencies of designs as a function of the size of the pilot studies used to estimate the design parameters with binary covariates and response 81

3.4 Boxplots of optimal selection probabilities derived using parameter estimates from pilot studies when $X|V$ followed a gamma distribution 85

3.5 Boxplots of asymptotic relative efficiencies of the Opt_{ipw} design as a function of the size of the pilot studies used to estimate the design parameters when $X|V$ followed a gamma distribution 86

3.6 Asymptotic relative efficiencies of IPW_2 estimators of α_x using balanced (8 and 6 strata) and Opt_{ipw} designs with clustered responses 91

3.7 Asymptotic relative efficiencies of ML estimators of α_x using balanced (8 and 6 strata) and Opt_{ipw} designs with clustered responses 92

3.8 Asymptotic relative efficiency from different phase-II designs when using IPW_2 analysis to estimate transitional effects 96

3.9 Asymptotic relative efficiency from different phase-II designs when using ML analysis to estimate transitional effects 97

4.1 Boxplots of the difference in sampling fractions between optimal and other designs with binary covariates and response when $n = 250$ and $N = 1000$ 107

4.2 Boxplots of the difference in sampling fractions between optimal and other designs with binary covariates and response when $n = 500$ and $N = 1000$ 108

4.3 Boxplots of the difference in sampling fractions between optimal and other designs with a continuous expensive covariate when $n = 250$ and $N = 1000$ 113

4.4 Boxplots of the difference in sampling fractions between optimal and other designs with a continuous expensive covariate when $n = 250$ and $N = 1000$ 114

4.5 Boxplots of the difference in sampling fractions between optimal and other designs with clustered responses when $\psi = 1.25$, $n = 250$ and $N = 1000$ 119

4.6 Boxplots of the difference in sampling fractions between optimal and other designs with clustered responses when $\psi = 1.25$, $n = 500$ and $N = 1000$ 120

4.7 Boxplots of the difference in sampling fractions between optimal and other designs with clustered responses when $\psi = 3.00$, $n = 250$ and $N = 1000$ 121

4.8 Boxplots of the difference in sampling fractions between optimal and other designs with clustered responses when $\psi = 3.00$, $n = 500$ and $N = 1000$ 122

4.9 Boxplots of the difference in sampling fractions between optimal and multistage adaptive Opt_{ipw} designs with binary covariates and response when $n = 250$ and $N = 1000$ 128

5.1 Timeline diagram for current status data on disease progression 136

5.2 Kaplan-Meier estimate of the distribution of the time from disease diagnosis to first damaged joint for clinic patients with stored sera based on current status data at clinic entry 137

Chapter 1

Introduction

This thesis is concerned with statistical methods for the analysis of incomplete data. Attention is given to data that are incomplete due to unforeseen reasons and data that may be intentionally incomplete because of decisions made at the planning stage of a study. Such decisions are often made in epidemiological research with a view to constraining costs (time, effort, or financial costs). We begin our discussion with a review of the terminology and approaches for dealing with data which are incomplete by happenstance, with a focus on the clinical trial setting.

1.1 Statistical Methods for Incomplete Data

1.1.1 Terminology for Incomplete Data

The statistical and clinical literature have adopted a standard nomenclature to characterise the mechanisms giving rise to missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Little and Rubin, 2002; Sterne et al., 2009). If the missingness (whether a variable is observed or not) does not depend on the response or other observed variates, then data are said to be MCAR. If missingness depends only on fully observed variables, then responses are said to be MAR. Otherwise, the missing data mechanism is called MNAR.

The term *ignorable* missing data is used when consistent analyses can be based on only those individuals with complete data (Little and Rubin, 2002); typically this encompasses MCAR mechanisms and MAR mechanisms when the response model controls for all prognostic factors related to missingness. Non-ignorable missing data mechanisms then typically corresponds to MNAR mechanisms and to MAR mechanisms when the response model does not control for all prognostic factors related to missingness. In the case of this non-ignorable MAR mechanism,

inconsistent estimates may arise from common methods of analysis (Joseph et al., 2004), but there are practical methods to correct for this which we discuss later; when data are MNAR, there are considerable challenges in trying to conduct appropriate analyses (Allison, 2001). The terms used to describe the nature of incomplete data must be used with caution, however, since they are to be interpreted in the context of the missing data mechanism, the response model, and the framework for the analysis (Little and Rubin, 2002). Determining if the missing data mechanism is MCAR, MAR, or MNAR is not trivial and usually cannot be done using observed study data (Beunckens et al., 2005); sensitivity analyses often play an important role (Molenberghs et al., 2004).

Perhaps the simplest strategy for dealing with missing data is to restrict attention to individuals with complete data. This is referred to as a *complete-case* analysis (CCA) (Demissie et al., 2003), or sometimes *observed-case* analysis (Wilcock et al., 2000). This approach has appeal from the standpoint that implementation is simple. However, standard complete-case analyses can incur substantial bias in estimates of event rates and associated treatment effects if the missing data mechanism is non-ignorable (Knol et al., 2010). This occurs because the sub-sample of completely observed individuals is not necessarily representative of the original sample.

1.1.2 Methods Based on Imputation

The CONSORT (Consolidated Standards of Reporting Trials) group recommends that the data from a clinical trial be analysed by retaining patients in their originally randomised assigned groups (Schulz et al., 2010) following the intent to treat (ITT) principle (Sackett and Gent, 1979). Such an ITT analysis needs to be clearly defined by the analyst, but typically involves all patients who were randomised; that is, it includes patients who yield complete data, as well as patients for whom data are missing because they prematurely terminated the study (Schulz et al., 2010). An imputation strategy facilitates this approach.

An immediate and obvious challenge lies in determining what values should be used in place of the missing variables. The National Research Council Panel on Handling Missing Data states that there is no ideal method of replacing missing values (National Research Council, 2010). Ad hoc methods of imputation can lead to biased estimates and standard errors may likewise be invalid. Multiple imputation offers a more rigorous framework for imputation as well as strategies for ensuring valid comparison of treatment effects. Analysts must be explicit in the assumptions made to use this method, but it can be used effectively in many settings.

Ad Hoc Methods of Imputation

If the outcome is a simple indicator of the success or "response" to treatment, then an imputation of "failure" is often made for missing responses (Alirezai et al., 2007). This is sometimes called non-responder imputation (NRI) and is used extensively in dermatology trials (Gottlieb et al., 2004; Gordon et al., 2006b; Menter et al., 2007; Leonardi et al., 2008, 2011). It is viewed

as a conservative way of handling missing data since it may be rationalised by the assumption that subjects prematurely withdraw from trials due to lack of efficacy (Saurat et al., 2008). It is important to note, however, that NRI is not necessarily conservative for estimation of treatment effects if the termination rate differs between treatment arms (McIsaac et al., 2013). In dermatology trials, researchers often replace missing measurements with their last observed value (Gottlieb et al., 2004; Menter et al., 2008; Gordon et al., 2011; Leonardi et al., 2011; Reich et al., 2011). This last observation carried forward (LOCF) imputation implicitly assumes that variables remain constant after their last observed values has been recorded (Papp et al., 2008a).

These crude methods of single imputation can produce results which are significantly biased or which are no longer interpretable in terms of the study objective. Interestingly, these methods remain popular for the analysis of clinical trials despite the facts that their validity depends on implausible restrictive assumptions (Cook et al., 2004; Papp et al., 2008b; Prakash et al., 2008), and that analyses based on these two different methods of imputation can lead to very different inferences in practice (Roy et al., 2011).

Multiple Imputation
There has been considerable interest in recent years in a more sophisticated imputation approach called multiple imputation (Rubin, 1987; Little and Rubin, 2002) and its implementation in several statistical software packages (Mayer et al., 2012) has contributed to its growing popularity. Multiple imputation involves randomly generating a value for each missing variable to create a pseudo-complete dataset. This is repeated a number of times to create multiple pseudo-complete datasets. Each dataset is then analysed separately to obtain several estimates of the treatment effect. These estimates are then appropriately combined to obtain a final summary estimate. Standard errors for the final summary estimate are computed in a way that recognises that the pseudo-complete datasets were obtained by imputation of data.

The advantage of multiple imputation over the cruder approaches to imputation discussed earlier arises from the specification of a model to generate the imputed values. Since the missing values can never be truly known, the model accounts of the uncertainty in predicting the imputed values. This model must be specified by the analyst (either explicitly or implicitly), but it can exploit available observed information to make the imputation procedure more informed than NRI or LOCF. Of course the validity of this approach hinges on correct specification of the imputation model; if the imputation model is incorrect, inferences will be invalid (Carpenter et al., 2006).

1.1.3 Inverse Probability Weighting

Another approach for addressing biases that may arise from data which are MAR involves weighting observations from individuals who have provided complete information so that the resulting weighted complete-case analysis furnishes estimates compatible with the complete sample (Whittemore and Halpern, 1997). Missing data are not imputed when inverse probability

weighting (IPW) is used, rather the complete cases are reweighed to reflect the fact that they are potentially also representing several unobserved cases. Available information on incomplete cases can be exploited in IPW to model the probability that an individual will be completely observed; the weight for each individual with complete data is the inverse of this probability. Thus the IPW approach simply requires a model for the probability of missingness which can be fitted and implemented in many statistical packages (Hogan et al., 2004; van der Wal and Geskus, 2011).

Inverse probability weighting, as described above, eliminates the potentially significant biases of standard complete-case analyses in MAR data, but does not optimally exploit individuals with missing response data. Augmented inverse probability weighted approaches (Robins et al., 1994; Tsiatis, 2006) are an extension of IPW which allow for greater use of information from individuals with incomplete data and, as a result, do not suffer from as much loss of power as IPW may suffer from. Augmented inverse probability weighting (AIPW) requires the specification of a second model, but consistent estimators may be found if either of the models is correctly specified (Robins et al., 1994; Carpenter et al., 2006; Tsiatis, 2006).

1.2 Incomplete Data Arising by Two-Phase Designs

Consider the problem of estimating a particular finite population characteristic such as the mean of a random variable X, denoted $\overline{X}$. If X is highly variable, then a simple random sample would need to be quite large in order for a precise estimate of $\overline{X}$ to be obtained. Now suppose that it is very expensive to observe X, where this expense may be either in the actual fiscal cost of observing X, or in terms of the time and resources required to obtain accurate measurements. Budgetary constraints make it impossible to observe X in large samples. Further suppose an auxiliary variable V is available which is strongly associated with X, relatively inexpensive to measure, and of a known range. Two-phase sampling approaches, which were formalised by Neyman (1938), use the inexpensive information contained in V to select samples with which one can efficiently estimate $\overline{X}$.

As the name suggests, there are two phases to these sampling procedures. The first phase (phase I) involves collecting a large sample of individuals and measuring the cheap auxiliary variable V for all individuals in this sample. This primary sample can then be stratified based on the values of V. The second phase (phase II) involves using stratified sampling to select a small subset of the primary sample. Values of X are measured for individuals in this secondary subsample and these are used to estimate $\overline{X}$.

Strata of the primary sample are constructed so that individuals within the same stratum have similar values of V. Since X and V are highly correlated, the values of X within strata will also be fairly homogenous. This two-phase sampling approach improves efficiency over simple random sampling by sampling a small number of individuals from each stratum. The difference between

variances resulting from an estimator based on simple random sampling ($\widehat{\overline{X}}_{SRS}$) and an estimator based on, for example, proportionately allocated stratified sampling ($\widehat{\overline{X}}_{strat}$) is asymptotically

$$var(\widehat{\overline{X}}_{SRS}) - var(\widehat{\overline{X}}_{strat}) \doteq \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{h=1}^{H} \left(\overline{X}_h - \overline{X}\right)^2 W_h, \tag{1.1}$$

where $\overline{X}_h$ is the mean for X in stratum h, W_h is the proportion of the population in stratum h, n is the sample size, and N is the total population size (Lohr, 2009). Note that (1.1) is always non-negative, and that it increases as the $\left(\overline{X}_h - \overline{X}\right)^2$ increase. Thus, there is a gain in efficiency if X values tend to be more similar within-strata than between-strata, as they are in two-phase sampling. Therefore, when an auxiliary variable is available to guide selection at the second stage, estimators based on proportional two-phase sampling should be more efficient than those based on simple random sampling. The size of the gain in efficiency over simple random sampling depends on how strata are defined and on the stratum-specific phase-II sampling probabilities that are employed; in most cases, it is possible to achieve greater efficiency through designs other than proportional sampling.

Two-phase sampling (or double sampling) has been widely used to increase the precision of study results while keeping study costs low (Pickles et al., 1995; Whittemore and Halpern, 1997; Chatterjee et al., 2003). This sampling framework can be effective whenever the cost of measuring a specific covariate is expensive, relative to the cost of measuring some associated auxiliary information. Two-phase sampling requires the collection of inexpensive auxiliary information in a large phase-I sample, so that this inexpensive information can be exploited in determining which individuals should be selected into a phase-II subsample where the expensive covariates are measured (Reilly and Pepe, 1995; Zhao et al., 2009). The efficiency gain that comes from such a two-phase sampling framework depends on the parameter of interest, the method of analysis, and the way in which the phase-I data are exploited in the design of the phase-II selection probabilities (Reilly, 1996). Here we are concerned with regression analyses using such two-phase data and the efficiency that can be gained by using optimal phase-II selection probabilities.

Consider a study in which interest lies in estimation of an element of the parameter vector α which describes the conditional distribution of a response Y given covariates X and V, as in

$$f(Y|X,V;\alpha).$$

Data arise according to the joint model $f(Y|X,V;\alpha)g(X|V;\beta)h(V;\gamma)$, but the parameters β and γ, which describe the distributions of the covariates, are typically treated as nuisance parameters.

Further suppose that these data are gathered in from a two-phase study where the response Y and inexpensive covariate V are collected for all individuals in a large phase-I sample of size N. The expensive covariate X, on the other hand, is only measured for individuals selected into the phase-II subsample. Selection into the phase-II sample will be governed by the phase-II selection

probabilities

$$\pi(Y,V;\delta) = P(R = 1|Y,V;\delta),$$

where $R_i = 1$ if individual i is selected for inclusion in the phase-II sample and $R_i = 0$ otherwise. The data are ultimately incomplete; all N individuals provide information on Y and V, but measurements of X will be missing for the $\sum_{i=1}^{N}(1-R_i)$ individuals not selected into the phase-II sample. Note that within this two-phase sampling framework, we consider missingness which occurs by design so we can be confident that missingness is not governed by the latent variables – i.e. $P(R = 1|Y,X,V;\delta) = P(R = 1|Y,V;\delta)$ – so data will be at most MAR in the terminology of Little and Rubin (2002). If the phase-II selection probabilities do not exploit the phase-I data – i.e. $P(R = 1|Y,V;\delta) = P(R = 1;\delta)$ – then the data will be MCAR. Note that analyses which assume the data arise as a simple random sample can introduce biases unless data are MCAR (Lawless et al., 1999; Little and Rubin, 2002). However, using phase-II selection probabilities which result in data that are MCAR precludes one from gaining efficiency through exploitation of the phase-I data.

1.3 Motivating Settings and Studies

1.3.1 Incomplete Data in Randomised Dermatology Trials

In many areas of clinical research, it is difficult to collect complete data for all individuals. Studies in psychiatric conditions such as depression routinely feature high drop-out rates and it can be difficult to collect detailed clinical information in studies off many other mental health conditions. We discuss issues in the conduct of dermatology trials in detail here, partially motivated by collaborations with researchers in psoriatic arthritis. A detailed survey of the literature is given to help understand the scope of the issues and get a sense of the methods routinely used in practice.

In many randomised controlled dermatology trials, an experimental treatment is compared to standard care for its effect on symptoms of a particular disease or condition. The treatments usually take some time to alleviate the symptoms and so studies routinely involve a baseline assessment of the condition, usually at the time of randomisation, and assessment at one or more follow up visits. Tests for the effect of the experimental treatment may be based on a measure of the severity of the condition at a particular time, the percentage change in the severity from baseline, an indicator of whether a minimum percent improvement has been achieved, or an indicator of whether symptoms have been eliminated or not. Studies which explore treatment effect on psoriasis, for example, often incorporate measures of Psoriasis Area and Severity Index (PASI), Physician Global Assessment (PGA), and Dermatology Life Quality Index (DLQI). Studies can be found which explore treatment effect in terms of PGA scores (Leonardi et al., 2011) after

a certain number of weeks, the change in PASI or DLQI (Griffiths et al., 2010; Menter et al., 2007) over a period of time, or total clearance of symptoms by a set follow up time (Gordon et al., 2006b). The most common primary outcome in psoriasis trials is the binary PASI 75, an indicator of whether a decrease in PASI score of at least 75% from baseline has been achieved.

We present here an overview of the rates of missing data encountered in randomised controlled trials in psoriasis, the common strategies used for dealing with missing data in analyses, and the reasons given for incomplete data. Table 1.1 summarises this information for twenty-two randomised, placebo-controlled, double-blind phase 3 clinical trials of biologics (adalimumab, alefacept, briakinumab, efalizumab, etanercept, infliximab, and ustekinumab) in the treatment of plaque psoriasis published from 2000 to 2011. Studies were excluded if they compared use of concomitant topical treatments, explored immunogenicity only, or primarily explored safety.

The reported rate of missing responses in primary analyses ranged from 2.7%-47.7%; two studies did not report the rate of missing data. Fifteen of the twenty-two studies reported that NRI was used in the primary analyses to deal with missing response data. Five studies indicated that missing data were at least partially imputed using LOCF for primary analyses. Three of the studies did not indicate how missing data were handled. In secondary analyses, approaches for dealing with missing data included NRI, LOCF, CCA, and BOCF (a variation on LOCF that involves imputing baseline observations). Only three studies examined the sensitivity of primary analyses to their choice of missing data strategy. Papp et al. (2005) used LOCF in primary analyses and then assessed the robustness of their findings by performing a sensitivity analysis using NRI; Gordon et al. (2011) and Strober et al. (2011) used NRI for their primary analysis and used LOCF in a sensitivity analysis. Saurat et al. (2008) used NRI in their primary analyses and called this approach "the generally more conservative approach for analysing data". They used LOCF, however, when considering secondary analyses with continuous responses as they "considered it excessively conservative to impute a value of zero for missing patients", as would be done with NRI.

When data are MAR, standard complete-case analyses yields biased estimates of the response rate in the population of patients. The extent of the bias depends on the strength of the dependence of missingness on the observed covariates, as well as the serial association between responses over time. The estimates from an NRI analysis are also biased. When data are MAR, the degree to which "success rate" estimates are conservative will differ between subgroups defined by treatment and observed prognostic factors. The cumulative effect of these differing biases when comparing the probability of response between treatment groups will not necessarily be conservative; the direction of the bias could actually differ for different measures of treatment effect. Thus, despite being referred to as a "conservative approach" (Saurat et al., 2008), NRI analyses could, in fact, result in under-estimation or over-estimation of treatment effects. Biases here will be most common if the proportion of missing data differs greatly between treatment arms. LOCF and BOCF are valid if responses do not change over time. The size of any bias resulting from LOCF will depend on the amount of missing data and the (unknown) trend of the

Table 1.1: Summary of rates of missing data and strategies employed for dealing with missing data in examples of phase 3 randomised controlled trials in psoriasis

	Study Name	Percent missing for primary analysis	Strategies for primary analysis (and sensitivity)	Strategies for secondary analyses	Reported reasons for missing response data[†]
Leonardi et al. (2011)	REACH	18.1%	NRI	LOCF	LTF, LOE, AE, PtD, O.
Strober et al. (2011)		7.4%	NRI	NRI	LTF, AE, PTD, O.
Gordon et al. (2011)		47.7%	NRI(LOCF)	NRI, LOCF	LTF, LOE, AE, PtD, O.
Reich et al. (2011)		not reported	NRI(LOCF)	NRI, CCA	LTF, LOE, AE, PtD, O.
Griffiths et al. (2010)	ACCEPT	2.7%	not reported	not reported	LOE, AE, O.
Menter et al. (2008)		6.1%	NRI	LOCF	LTF, LOE, AE, PtD, O.
Van de Kerkhof et al. (2008)		11.9%	LOCF	LOCF	LOE, AE, O.
Leonardi et al. (2008)	PHOENIX 1	3.0%	NRI	BOCF[‡]	LOE, AE, O.
Papp et al. (2008c)	PHOENIX 2	2.7%	NRI	BOCF	LOE, AE, O.
Saurat et al. (2008)	CHAMPION	5.5%	NRI	LOCF	LTF, LOE, AE, PtD, O.
Tyring et al. (2007)		3.4%	LOCF	LOCF	LTF, LOE, AE, PtD, O.
Menter et al. (2007)	EXPRESS II	6.8%	NRI	BOCF	LTF, LOE, AE, O.
Gordon et al. (2006b)		5.4%	NRI	not reported	LTF, LOE, AE, PtD, O.
Dubertret et al. (2006)	CLEAR	8.8%	NRI	CCA, LOCF	LTF, LOE, AE, PtD, O.
Menter et al. (2005)		6.4%	NRI	LOCF	LTF, AE, PtD, O.
Papp et al. (2005)		8.5%	LOCF(NRI)	LOCF	LTF, LOE, AE, PtD, O.
Reich et al. (2005b)	EXPRESS	6.6%	NRI	BOCF, NRI	LOE, AE, O
Leonardi et al. (2005)		10.6%	NRI	NRI	LTF, LOE, AE, PtD, O.
Gottlieb et al. (2004)		not reported	NRI and LOCF	BOCF, NRI, LOCF	LOE, AE, O.
Lebwohl et al. (2003)		3.6%	not reported	not reported	LTF, LOE, AE, PtD, O.
Leonardi et al. (2003)		6.0%	LOCF	LOCF	LOE, AE
Krueger et al. (2002)		11.4%	not reported	not reported	LTF, LOE, AE, PtD, O.

[‡]Baseline Observation Carried Forward (BOCF) is closely related to LOCF imputation; here missing post-baseline continuous variables had baseline values imputed for subjects who prematurely terminated due to lack of efficacy, treatment failure, or need for alternate treatment.

[†]Reported reasons for missingness include LTF (lost to follow up); LOE (lack of efficacy, including worsening disease, inadequate response, unsatisfactory therapeutic effect, disease progression); AE (adverse event); PtD (patient decision, including withdrawn consent); O (Other, including pregnancy, administrative decision, death).

missing responses (Cook et al., 2004; Beunckens et al., 2005). Multiple imputation analyses will be valid whenever a correct model can be specified. IPW analyses can also achieve consistent estimators since the weighting can eliminate the bias inherent in standard complete-case analyses. The validity of the IPW approach hinges on correct specification of the model for missingness, but unlike the missing response model needed for valid imputation, here we need only model a binary indicator of missingness, which can be simpler. More efficient estimates can be obtained through the use of augmented inverse probability weighted estimating equations, but these have not yet seen wide-spread use in the clinical literature.

All of the trials reported at least some reasons for the incomplete response data. Reasons commonly given for missing data included lack of efficacy, adverse events, loss to follow up, as well as others such as pregnancy and administrative decision. Obtaining the cause of patient drop-out is important. The more information that is recorded on the reasons for incomplete data, the easier it is to select a model for the missing data mechanism and to carry out corresponding corrective analyses. For example, if it is known that loss to follow up occurred solely due to of administrative reasons unrelated to the response (e.g. the patient moved), then such responses are MCAR and the missing data are ignorable. If all incomplete data were due to such reasons, the reported results from standard analyses are easily interpreted. Drop-outs due to lack of efficacy could result in responses that are MAR or MNAR depending on whether this missingness is sufficiently explained by the treatment and observed prognostic factors. If an individual drops out of the study because their response is unfavourable to an extent which cannot be explained by the treatment assignment or other observed prognostic factors, then responses are MNAR and valid analyses are not straightforward. However, if responses are missing differentially by treatment arm due to a lack of efficacy of the experimental treatment, then responses could be MAR and valid analyses can be achieved through appropriate weighting or imputation as long as the necessary additional models are correctly specified. Obviously missing data are a pervasive problem in dermatology clinical trials and the methods being used to analyse these data in practice are not sufficient to account for a MAR mechanism.

1.3.2 Biomarker Studies in Psoriatic Arthritis

Psoriatic arthritis (PsA) is an immunological disease associated with considerable joint pain and inflammation which can ultimately lead to serious disability and poor quality of life (Langley et al., 2005; Chandran et al., 2010a). The disease course is complex and heterogeneous; some patients experience rapid joint destruction, and some exhibit little evidence of progression even after considerable time with the disease (Sutradhar and Cook, 2009; Gladman and Chandran, 2011). Identification of patients at high risk of progression is therefore critical to ensure timely intervention early in the course of disease for those who need it, and to avoid unnecessary use of expensive, powerful, but potentially toxic biologic therapies. Due to the important role of serum biomarkers and genetic factors in disease progression (Rahman et al., 2008; Gladman and

Chandran, 2011), there is increased interest in the design and analysis of biomarker studies.

The Centre for Prognosis Studies in the Rheumatic Diseases was established in 1976 as a tertiary referral centre affiliated with the University Health Network of the University of Toronto. It now maintains the largest cohort of patients with psoriatic arthritis in the world (University Health Network Centre For Prognosis Studies In The Rheumatic Diseases, 2007). Patients in this cohort attend the clinic for annual examinations during which the extent of damage of sixty-four individual joints is assessed and graded using a 5-point modified Steinbrocker scale (van der Heijde et al., 2005). In addition, levels of a well-established systemic marker of inflammation, the erythrocyte sedimentation rate (ESR), is recorded at clinic visits (Gladman and Chandran, 2011), and blood samples are collected and stored for use in future studies. Radiographic examinations are scheduled every two years. As a result of this careful followup, there is considerable information on the presence of inflammation and rates of disease progression.

Of course, disease progression can be modelled in a number of ways including the development of newly damaged joints (Sutradhar and Cook, 2009), the involvement of new types of joints (Tolusso and Cook, 2009; Chandran et al., 2010b), and the onset of a particular condition. The main focus of this thesis will be set in the context where disease progression is modelled as a binary indicator of one of these events (see Figure 1.1).

Figure 1.1: Diagram displaying the timeline for collection of information on disease progression over a fixed follow-up time in the Psoriatic Arthritis setting with a univariate response. The value of the expensive covariate X can be determined through retrospective examination of the sera stored from the time of clinic entry. So, the specification of $R_i = I(X_i \text{ obs.})$ can be made after the follow-up visit.

We will also consider bivariate responses (see Figure 1.2), which represent the development of involvement of the two sacroiliac (SI) joints, an important concern because it represents the onset of spondyloarthritis. Damage of the SI joints is determined by clinical and radiological examination with the extent of damage in each joint graded using a standardised scale (Rahman et al., 1998). Serum biomarkers and genetic factors can play important roles in identifying patients at high risk for developing psoriatic spondyloarthritis (Rahman et al., 1998; Gladman and Chandran, 2011), and as a consequence, biomarker studies are of considerable importance.

A major stream of research in this centre now involves the identification of serum biomarkers prognostic for disease progression. The enzyme MMP-3 is the biomarker of greatest current interest because it is thought to play an important role in the destruction of cartilage and bone in rheumatic diseases characterised by synovitis (Chandran et al., 2010a). A current goal is to study

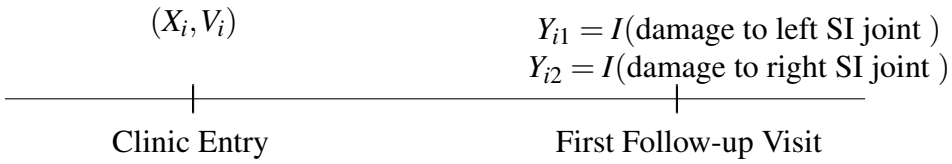

Figure 1.2: Diagram displaying the timeline for collection of information on disease progression over a fixed follow-up time in the Psoriatic Arthritis setting with clustered responses. The value of the expensive covariate X can be determined through retrospective examination of the sera stored from the time of clinic entry. So, the specification of $R_i = I(X_i \text{ obs.})$ can be made after the follow-up visit.

the effect of the biomarker MMP-3 on progression of PsA while accounting for the effect of the inexpensive and easy to measure ESR (Gladman and Chandran, 2011), to establish the added value of MMP-3 over the marker most commonly used in this setting. The stored blood samples can be used to obtain measures of MMP-3, but measurement of MMP-3 level is expensive and cannot be carried out for all clinic patients. Given the availability of the data on joint damage over assessments and the inexpensive ESR data, we aim to use a two-phase study design to derive optimal sampling probabilities for the selection of the phase-II sample (i.e. the patients on whom MMP-3 will be measured).

Since different selection procedures result in different levels of efficiency in parameter estimation, it is of interest to determine how to best exploit the available clinic data in selecting individuals for measurement of the biomarker. By developing optimal phase-II selection strategies, biomarker effects can be estimated with greater precision and associated tests will be more powerful. These efficiency gains are possible without increasing the study budget – the idea is not to select more individuals for biomarker analysis, but rather to select individuals in a more informed way.

1.3.3 Canadian Longitudinal Study on Aging

The Canadian Longitudinal Study on Aging (CLSA) is a pan-Canadian longitudinal study of disease incidences and associated risk factors involving 50,000 individuals aged 45 to 85 years old who are to be followed for 20 years or death. All participants in the CLSA will provide some information to the study, while a subset of 30,000 will be chosen for additional, in-depth examination. This sub-cohort will undergo a more intensive clinical examination, providing imaging data, and giving biological specimens every three years; specimens will be stored in biobanks in a controlled environment to facilitate subsequent testing. Thus the biobank will serve as a valuable resource for affiliated investigators to study risk factors predictive of disease onset and progression (CLSA, 2009; Raina et al., 2009).

Samples will be expensive to process for all 30,000 individuals in the cohort undergoing intensive

follow-up, so it will be of central importance to determine how individuals should be selected for testing of stored specimens. We therefore explore the extension of the two-phase sampling problem to longitudinal data. Since interest lies in the onset of disease, we focus on transitional models and formulate the exposure effects on transition probabilities. We study various designs for sampling and analysis to investigate how optimal selection procedures can be derived at a particular time point given the partial histories of individuals data.

Specifically, we examine the improved precision in estimation that can result when more information is used in deriving optimal selection probabilities. We suppose that longitudinal data arise as in Figure 1.3 and consider different designs for selecting individuals to undergo an expensive genotyping process where the goal is efficient estimation of the transitional effect of a genetic marker. Phase-II sampling designs for selecting individuals for measurement of the marker will be considered at different time points. By allowing these designs to depend on the varying amount of information that is available at different time points, we can explore how the efficiency of these phase-II designs changes as more phase-I data are available for exploitation.

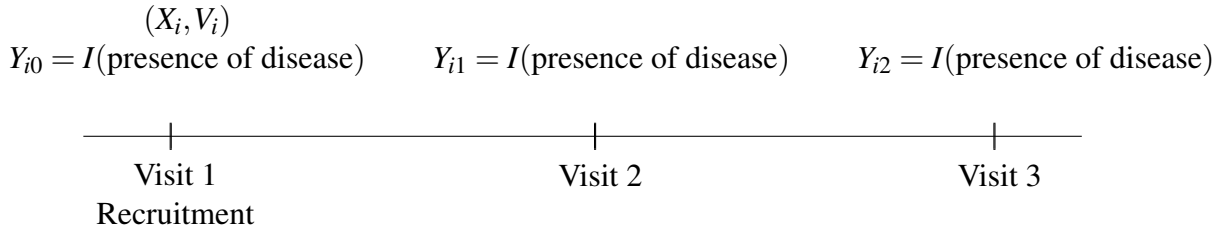

Figure 1.3: Diagram of timeline for collection of longitudinal assessments of disease activity or progression. Here the selection of individuals for measurement of X can be made at any time point t through specification of $P(R_i = I|V_i, Y_{i0}, \ldots, Y_{it})$, so that more information is available for exploitation in the sampling designs at later time points.

1.4 Outline of Thesis Research

The remainder of the thesis is organised as follows.

1.4.1 Methods for Data Missing by Happenstance

In Chapter 2 attention is directed at data which are incomplete by chance. The likelihood function is used as a basis for discussing different missing data mechanisms for incomplete responses in short-term and longitudinal studies, as well as for missing covariates. In Section 2.2 we discuss the problem of incomplete binary responses with an emphasis on issues in clinical trials.

We discuss the case of incomplete longitudinal data in Section 2.3, and the problem of incomplete covariates in Section 2.4. We briefly discuss common *ad hoc* strategies for dealing with incomplete data, such as complete-case analyses and the naive methods of imputation introduced in Section 1.3.1. We also review more broadly appropriate approaches for dealing with incomplete data: the EM algorithm, multiple imputation, and inverse probability weighted estimating equations. Simulation studies are reported which demonstrate how to implement these procedures and present asymptotic and empirical frequency properties for these estimators in a variety of settings. In Section 2.5 we explore the asymptotic biases of multiple imputation and inverse probability weighted estimators when the respective modelling assumptions are misspecified due to an incorrect formulation of models for imputation or the missing data process.

1.4.2 Two-Phase Sampling Design

Chapter 3 is devoted to the study of issues in observational studies in which a covariate is incomplete by design. Two-phase studies are reviewed and the asymptotic efficiencies of several methods of analysis are presented (Section 3.3). These asymptotic efficiencies are used to derive optimal phase-II sampling designs which select informative sub-samples in the second phase of sampling so that the asymptotic variance of a resulting estimator is minimised (Section 3.4). Results from simulation studies demonstrate the empirical efficiencies resulting from these optimal designs in the context of psoriatic arthritis data. Optimal designs are compared to those based on simple random sampling and balanced sampling to quantify the potential gains in terms of efficiency when optimal designs are employed. Use of optimal designs requires *a priori* knowledge of certain parameters, however in Section 3.5, we explore the sensitivity of optimal designs to misspecification of design parameters, and consider the robustness of an optimal design to misspecification of the nuisance covariate model. We further examine the effect on efficiency of utilising external pilot data to estimate the design parameters needed for derivation of optimal designs. Designs which are optimal for analyses based on inverse probability weighted estimating equations are shown to result in efficiency gains for several different methods of analysis and are shown to be robust to misspecification of the parameters or models used to derive the optimal designs. Furthermore, these optimal designs for inverse probability weighted estimating equations are shown to be well behaved when necessary design parameters are estimated using relatively small external pilot studies. In the simulation studies in Sections 3.1-3.6 we focus on a binary, univariate response and binary auxiliary covariate, and consider both a binary and a continuous expensive covariate.

In Section 3.7 we consider asymptotic relative efficiencies of two-phase sampling designs with clustered responses and cluster-level exposure and auxiliary variables. Marginal models (Liang and Zeger, 1986) are adopted in this setting with analysis frameworks based on maximum likelihood and inverse probability weighted pseudo-likelihood. We demonstrate the potential inefficiency that can result from use of balanced designs in this setting. Furthermore, we show the

efficiency gains that can be achieved through the use of an optimal design for IPW analysis, whether or not the resulting clustered data are actually analysed using IPW; this is the first foray into this area that we are aware of. In Section 3.8 we give a framework for two-phase designs in longitudinal studies, another relatively unexplored area, where interest lies in modelling the effect of an exposure variable on the onset of disease under a first-order Markov model. We explore the relative efficiency of phase-II sampling designs based on increasing amounts of information in the longitudinal responses and show that the balanced design may become less efficient when more data is available at the design stage. In contrast, the optimal design is able to exploit additional information to increase efficiency whenever more data is available at phase I.

1.4.3 Adaptive Stratified Two-Phase Sampling

In Chapter 4, we consider an innovative adaptive two-phase design which breaks the phase-II sampling into a phase-IIa sample obtained by a balanced or proportional stratified sampling strategy, and a phase-IIb sample collected according to an optimal sampling design for IPW analysis. Optimal phase-IIb sampling probabilities are derived for given phase-IIa sample sizes. These optimal phase-IIb designs can be implemented using initial parameter estimates that can be found by analysing the phase-I and phase-IIa data (an 'internal' pilot study). This approach exploits the previously established robustness of optimal inverse probability weighted designs to overcome the difficulties associated with the fact that derivations of optimal designs require *a priori* knowledge of parameters. The efficiency of this adaptive design is compared to those of the proportional and balanced sampling designs, and to the efficiency of the true optimal design, in a variety of settings reflecting the datasets considered in Chapter 3: a univariate response and binary covariates, a continuous expensive covariate X, and clustered responses. It is natural to consider this approach in settings with complex models for which it is difficult to even speculate on suitable parameter values at the design stage. The adaptive design is important in that it allows for near-optimal selection probabilities to be used without requiring the added costs associated with collected external pilot data. The efficiency gains of this adaptive two-phase design are especially obvious in the setting involving clustered response data.

These results are summarised in Chapter 5, which also contains a description of important problems that remain unsolved and directions for future work.

Chapter 2

Statistical Methods for Data Missing By Happenstance

2.1 Introduction

In well-conducted clinical trials, randomisation eliminates the possible effect of confounding variables in the assessment of treatment effects. That is, when the assignment of the treatment to patients is carried out by random allocation, different treatment groups will have similar distributions of demographic and clinical features, so any differences seen in the distribution of responses between the treatment groups are attributable to the different treatments they receive. There are a number of other rationale put forward for use of randomisation in health research (Sprott and Farewell, 1993), but it is the elimination of the effect of confounding variables and facilitation of causal inference that has had the most profound impact in advancing scientific understanding.

Following recruitment and randomisation, however, participants in clinical trials often withdraw before completion of follow-up, leading to incomplete outcome data. Incomplete data can of course arise for a variety of reasons; many illustrative examples can be seen in the second chapter of Molenberghs and Kenward (2007). Depending on the reasons for withdrawal, the individuals who remain in the study may no longer form groups with similar distributions of the demographic and clinical features, which compromises the validity of causal inferences. The purpose of this chapter is to discuss models and mechanisms by which incomplete data can arise in clinical trials, the consequences missing data can have on the interpretation of study results, and methods which can be employed to minimise the effect of these consequences. The impact of model misspecification in methods for dealing with incomplete data is also studied based on the limiting behaviour of resulting estimators. A clear understanding of the practical and statistical issues involved with incomplete response data will improve ability to critically appraise the clinical

literature and help guide selection of methods of analysis.

The remainder of this chapter is organised as follows. In Section 2.2 we discuss the problem of incomplete binary responses. We restrict attention to the case of a binary treatment indicator and a single binary auxiliary variable to simplify the discussion, calculations, and empirical studies, but we remark on practical issues with more complex settings at the end of this section. We discuss the case of incomplete longitudinal data in Section 2.3, and the problem of incomplete covariates in Section 2.4. In Section 2.5 we explore the limiting values of estimators obtained from analyses which aim to address the incomplete nature of data, but which do so based on one or more misspecified models. Concluding remarks are made in Section 2.6.

2.2 Methods for Incomplete Binary Responses in Randomised Trials

2.2.1 Models and Measures of Treatment Effect

Consider a balanced two-arm clinical trial in which patients are randomised to receive either an experimental treatment or standard care. Let $X = 1$ indicate that a patient was allocated to receive experimental therapy and $X = 0$ otherwise, where $P(X = 1) = 0.5$. Suppose the outcome of interest is whether the patient had a successful response; we let $Y = 1$ if this is the case and $Y = 0$ otherwise. We illustrate the problem of dependently missing data by considering a situation with a single additional binary variable V, where $V = 1$ indicates the presence of a particular feature and $V = 0$ otherwise; $P(V = 1) = p$. Suppose that the variable V is an effect modifier (Rothman and Greenland, 1998) so that the treatment has a different effect for individuals with and without the risk factor. This may be represented by the logistic model

$$P(Y = 1|X,V;\eta) \quad = \quad \text{expit}(\eta_0 + \eta_x X + \eta_v V + \eta_{xv} XV) \,, \tag{2.1}$$

where $\eta = (\eta_0, \eta_x, \eta_v, \eta_{xv})'$. In most situations there will be sub-populations between which there is variation in the event rate and the effect of treatment; (2.1) is the simplest model which accommodates this phenomenon.

While (2.1) may reflect a simple reality, in clinical trials we typically aim to assess treatment effects based on marginal models (i.e. models that do not condition on prognostic variables such as V); indeed provided X is independent of V, the causal effect of treatment is typically defined in terms of such a model. Thus the logistic model used for treatment comparisons is formulated as

$$P(Y = 1|X;\alpha) \quad = \quad \text{expit}(\alpha_0 + \alpha_x X) \,, \tag{2.2}$$

where $\alpha = (\alpha_0, \alpha_x)'$. Of course,

$$P(Y = 1|X;\alpha) \quad = \quad E_V\left[P(Y = 1|X,V;\eta);p\right] \ , \tag{2.3}$$

since V is independent of X due to randomisation, and so it is possible to obtain the functional form of α in terms of $(\eta', p)'$.

The resulting response rates in the control and treatment arms are $p_C = P(Y = 1|X = 0) = \text{expit}(\alpha_0)$ and $p_T = P(Y = 1|X = 1) = \text{expit}(\alpha_0 + \alpha_x)$, respectively. Some common measures of treatment effect include the absolute difference $\text{AD} = p_T - p_C$, the number needed to treat $\text{NNT} = (p_T - p_C)^{-1}$, the relative risk $\text{RR} = p_T/p_C$, and the odds ratio $\text{OR} = [p_T/(1-p_T)]/[p_C/(1-p_C)]$ (Matthews and Farewell, 1996; Laupacis et al., 1998). When the experimental treatment has a higher response rate, the AD and NNT measures are positive and the RR and OR are greater than one.

Let $I(A)$ be an indicator function such that $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. If response data are incomplete, in order to thoroughly discuss modelling issues it is necessary to introduce a new random variable $R = I(Y \text{ observed})$, so $R = 1$ if Y is observed and $R = 0$ otherwise. The biases that result from incomplete data arise if there is an association between the response (Y) and whether we observe it or not (R). There are a variety of ways of introducing an association between Y and R including through bivariate binary models (Cox, 1972) and shared random effect models (Albert and Follmann, 2009). Here we consider the setting in which both Y and R are associated with the covariates X and V. When V is unknown, an association between Y and R exists because of the omission of V from the analysis. We adopt this framework because when V is known, there are a variety of approaches to incorporating information about V into the analyses to mitigate problems, as we discuss in the following sections.

Suppose that the missing data model is

$$P(R = 1|X,V;\delta) = \text{expit}(\delta_0 + \delta_x X + \delta_v V + \delta_{xv} XV) \ , \tag{2.4}$$

where $\delta = (\delta_0, \delta_x, \delta_v, \delta_{xv})'$. This model accommodates a different dependence on V in the two treatment arms. We assume in this idealised setting that $R \perp Y|X,V$. Since $X \perp V$ by randomisation, the marginal proportion of missing data is

$$\begin{aligned} p_R = P(R = 1;\delta,p) \quad &= \quad E_X\left\{E_V\left[P(R = 1|X,V)\right]\right\} \\ &= \quad \sum_{x=0}^{1}\sum_{v=0}^{1} P(R = 1|X = x,V = v;\delta)\ P(V = v;p)\ P(X = x) \ , \end{aligned}$$

where $P(V = v;p) = p^v(1-p)^{1-v}$, and $P(X = x) = 1/2$ due to balanced randomisation. The

joint probability mass function for $Y,R|X$ is

$$\begin{aligned} P(Y,R|X;\Omega) &= E_V\left[P(Y|X,V;\eta)\,P(R|X,V;\delta)\right] \\ &= \sum_{v=0}^{1} P(Y|X,V=v;\eta)\,P(R|X,V=v;\delta)P(V=v;p)\,, \end{aligned} \tag{2.5}$$

where $\Omega = (\eta', p, \delta')'$. From (2.5) we can derive the conditional odds ratio for the association between Y and R given X as

$$OR_{Y,R|X} = \frac{P(Y=1,R=1|X;\Omega)}{P(Y=1,R=0|X;\Omega)} \Big/ \frac{P(Y=0,R=1|X;\Omega)}{P(Y=0,R=0|X;\Omega)}\,,$$

and we can calculate the conditional probability

$$P(Y|X,R;\Omega) = \frac{P(Y,R|X;\Omega)}{P(R|X;\Omega)} = \frac{P(Y,R|X;\Omega)}{\sum_{y=0}^{1} P(Y=y,R|X;\Omega)}\,. \tag{2.6}$$

So, thus far we have defined a simple model for $Y|X,V$ and $R|X,V$ under the assumption that Y and R are conditionally independent given (X,V). When we condition on X but not V, the response Y and the missing data indicator R are associated (i.e. dependent). We have mentioned that this setting was problematic, but here we will explore why this is the case.

2.2.2 Parameter Estimation with Incomplete Response Data

Complete-Case Analyses

Complete-Case Analyses When Covariate V is Unknown

The likelihood function is perhaps the most fruitful starting point when considering inference based on parametric models (Sprott, 2000). When data may be incomplete, the availability of the response of interest is stochastic, and hence the observed data likelihood is

$$L \;\propto\; P(Y,R=1|X)^R\,P(R=0|X)^{1-R}\,.$$

Noting that $P(Y,R=1|X) = P(Y|R=1,X)P(R=1|X)$, this may be re-expressed as $L_{Y|R=1,X} \cdot L_{R|X}$ where

$$L_{Y|R=1,X} = \left[P(Y=1|R=1,X;\Omega)^Y\,P(Y=0|R=1,X;\Omega)^{1-Y}\right]^R \tag{2.7}$$

is obtained from $P(Y|R=1,X)^R$ by considering the two possible realisations of Y, and

$$L_{R|X} = P(R=1|X;\Omega)^R \, P(R=0|X;\Omega)^{1-R} \, . \tag{2.8}$$

When responses are not available from all individuals in a sample, it is tempting to restrict attention to individuals with complete data and base analyses on this subset. This restriction, however, implicitly conditions on $R=1$ so that a complete-case maximum likelihood analysis actually maximises the partial likelihood (2.7). It appears that (2.8) does not contain information about the parameters we are interested in because it relates to the missing data process alone. Note however that while (2.7) is indexed by Ω, the quantities estimated by standard analyses based on available data (i.e. the sub-sample of individuals with $R=1$) are

$$\alpha_0^\dagger = \text{logit}\, P(Y=1|X=0,R=1;\Omega)$$

and

$$\alpha_x^\dagger = \text{logit}\, P(Y=1|X=1,R=1;\Omega) - \alpha_0^\dagger \, .$$

These parameters differ from α_0 and α_x whenever $P(Y|X,R=1) \neq P(Y|X)$, which will occur here if $P(Y|X,V) \neq P(Y|X)$ and $P(R|X,V) \neq P(R|X)$. Using (2.6), we can compute the naive measures of treatment effect which are actually being estimated from complete-case analyses: $\text{AD}^\dagger = P(Y=1|X=1,R=1) - P(Y=1|X=0,R=1)$, $\text{NNT}^\dagger = 1/\text{AD}^\dagger$, $\text{RR}^\dagger = P(Y=1|X=1,R=1)/P(Y=1|X=0,R=1)$, and $\text{OR}^\dagger = [P(Y=1|X=1,R=1)/P(Y=0|X=1,R=1)]/[P(Y=1|X=0,R=1)/P(Y=0|X=0,R=1)]$.

To explore this more fully, we consider here some specific parameter configurations. Let $P(X=1)=0.5$ and $P(V=1)=0.5$. In the response model (2.1), we let $\eta_v = 0$ and $\eta_{xv} = \log(2)$ so the odds ratio characterising the treatment effect is twice as big for those with $V=1$ compared to those with $V=0$. We set $\alpha_x = \log(1.5)$ in (2.2), so the marginal odds ratio of the treatment effect is 1.5, and we solve for η_0 and η_x so that $P(Y=1|X=0) = \text{expit}(\alpha_0) = 0.5$ (i.e. the probability of response is 0.5 in the control arm). The marginal relative risk is therefore 1.2. In the missing data model (2.4) we set $\delta_x = \delta_v = 0$ and for each δ_{xv} we solve for δ_0 so that $P(R=1)=0.5$.

Figure 2.1 displays a plot of $RR^\dagger$ and $OR^\dagger$, the limiting values of complete-case estimators of RR and OR, as a function of δ_{xv}. When $\delta_{xv}=0$, the probability of the response being missing is the same for all individuals regardless of their covariates (data are *missing completely at random*, in the terminology of Little and Rubin (2002)), so $P(R|X,V) = P(R|X) = P(R)$. In this case, $RR^\dagger = RR = 1.2$ and $OR^\dagger = OR = 1.5$. When $\delta_{xv} < 0$, complete-case estimators of these effect measures will be too small and hence correspond to a understatement of the effect of treatment. Conversely, when $\delta_{xv} > 0$, the inferences regarding the benefit of treatment are anti-conservative.

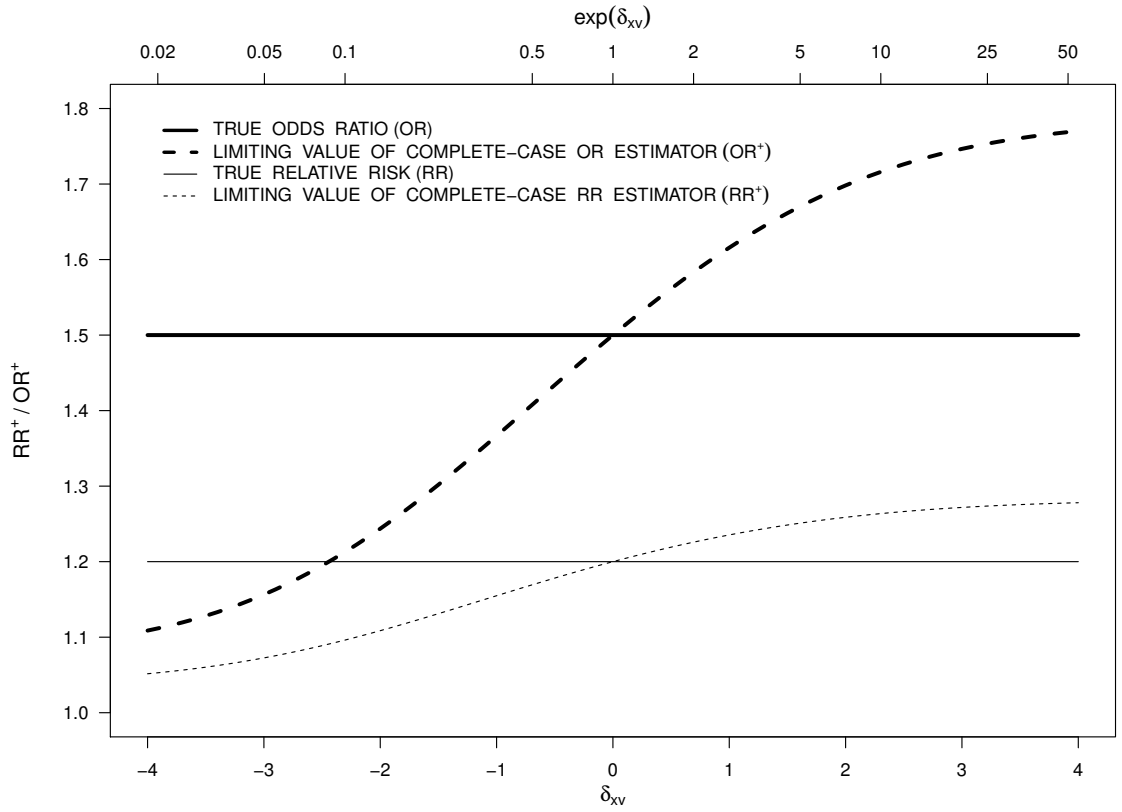

Figure 2.1: Limiting values of naive complete-case estimators of the relative risk ($RR^{\dagger}$) and odds ratio ($OR^{\dagger}$) as a function of δ_{xv}.

Complete-Case Analyses when Covariate V is Known

If we are able to identify the variable V which renders Y and R conditionally independent (i.e., $Y \perp R|X,V$), another option is to write the observed data likelihood based on the conditional model as

$$L \propto P(Y,R=1|X,V)^R \left[P(R=0|X,V)\right]^{1-R}.$$

Since $P(Y,R=1|X,V) = P(Y|X,V)P(R=1|Y,X,V)$ and $P(R=1|Y,X,V) = P(R=1|X,V)$ this can in turn be written as $L_{Y|X,V} \cdot L_{R|X,V}$ where $L_{Y|X,V} \propto P(Y|X,V)$ and $L_{R|X,V} \propto P(R|X,V)$. In practice one would naturally restrict attention to the partial likelihood $L_{Y|X,V}$, since we are not typically interested in modelling the missing data process unless it is necessary. As seen above, a complete-case analysis with restriction to individuals with $R=1$ yields inconsistent estimators of α when we just condition on X, however when we condition on V as well, a complete-case analysis gives consistent estimators for η. Identification of variables like V which are prognostic for Y and associated with the missing data process is therefore key to ensure consistent estimation of

parameters. It is not sufficient for these variables to be associated with the response alone or the missing data status alone since in either case such variables cannot render Y and R conditionally independent.

While conditioning on a suitable V seems to have solved our problem, the catch is that we did not want to condition on V in our assessment of the treatment effect – we are estimating η instead of α, so we are estimating the wrong thing! We do have the option of modelling $V|X$, which amounts to modelling the marginal distribution of V since X was determined by randomisation, and given an estimate of p as $\widehat{p}$, we can compute a crude estimate by solving for α in

$$\tilde{P}(Y=1|X;\tilde{\alpha}) = \sum_{v=0}^{1} P(Y=1|X,V=v;\widehat{\eta})\ \widehat{p}^{v}(1-\widehat{p})^{1-v}\,.$$

This procedure has been called the expected conditional mean (ECM) approach (Chen and Cook, 2012). Due to the so-called curse of dimensionality, this process is considerably more challenging and undesirable when V is high dimensional (i.e. a vector) (Robins and Ritov, 1997). A very convenient and more direct approach to estimating α is obtained using inverse probability weights as we describe in the next sub-section.

Use of Inverse Probability Weights

Suppose we have a sample of size N giving data $\{(Y_i,X_i,V_i), i=1,\ldots,N\}$. The score function for the logistic regression model in (2.2), resulting from (2.7) can be written as

$$S(\alpha) = \sum_{i=1}^{N} R_i\ (Y_i - E(Y_i|X_i;\alpha))\ \ [1,\ X_i]'\ .$$

With complete data (i.e. if $P(R_i=1)=1,\ i=1,\ldots,N$) this has expectation zero and hence yields a consistent estimator for α (McCullagh and Nelder, 1989). With incomplete data however,

$$\begin{aligned} E[S_i(\alpha)] &= E_X\left\{E_{Y|X}\left\{E_{R|Y,X}\left[S(\alpha)\right]\right\}\right\} \\ &= E_X\left\{E_{Y|X}\left[P(R_i=1|Y_i,X_i)\ (Y_i - E(Y_i|X_i;\alpha))\,[1,\ X_i]'\right]\right\}\ , \end{aligned}$$

which does not in general equal zero. If the probability of a response being missing depends on Y given X, then inconsistent estimators are obtained for α; the corresponding limiting values are the $\alpha^{\dagger}$ given in the previous section.

Now again suppose we are able to identify V as a covariate which renders $Y \perp R|X,V$. In this case we can employ the model for $P(R=1|Y,X,V) = P(R=1|X,V;\delta)$ in an *inverse probability*

weighted estimating function defined as

$$U(\alpha) = \sum_{i=1}^{N} \frac{R_i}{P(R_i = 1|X_i, V_i; \delta)} \ (Y_i - E(Y_i|X_i; \alpha)) \ [1, \ X_i]' \tag{2.9}$$

(Robins et al., 1995). Taking the expectation of an entry of (2.9) as before yields

$$\begin{aligned} E[U_i(\alpha)] &= E_{X,V}\left\{E_{Y|X,V}\left[E_{R|Y,X,V}\left(\frac{R_i}{P(R_i = 1|X_i, V_i)}(Y_i - E(Y_i|X_i; \alpha))\,[1, \ X_i]'\right)\right]\right\} \\ &= E_{X,V}\left\{E_{Y|X,V}\left[(Y_i - E(Y_i|X_i; \alpha))\,[1, \ X_i]'\right]\right\} \\ &= E_X\left\{E_{V|X}\left\{(E(Y_i|X_i, V_i) - E(Y_i|X_i; \alpha))\,[1, \ X_i]'\right\}\right\} \\ &= E_X\left[(E(Y_i|X_i; \alpha) - E(Y_i|X_i; \alpha))\,[1, \ X_i]'\right] = 0 \end{aligned} \tag{2.10}$$

and so a consistent estimator of α is obtained from (2.9).

Note that in practice the parameters in the model $P(R|X,V;\delta)$ must be estimated and this can easily be carried out via logistic regression since R is a binary variable. Naive standard errors which do not recognise that the weights have been estimated can lead to invalid tests (with incorrect type I error rates) and invalid confidence intervals (with coverage rates not compatible with the normal level).

Multiple Imputation

Multiple imputation is, in its simplest implementation, a simulation-based approach to creating complete data from an incomplete dataset. Again suppose that we have identified a covariate V which renders $Y \perp R|X,V$, and the model for $Y|X,V$ is given by (2.1). A multiple imputation approach involves fitting a model to $Y|X,V$ based on individuals with complete data, even though $Y|X$ is the model of interest. The fitted model would give a consistent maximum likelihood estimator $\widehat{\eta}$, along with the asymptotic covariance matrix for $\widehat{\eta}$, $\mathscr{I}^{-1}(\widehat{\eta})$, where $\mathscr{I}(\eta)$ is the expected information matrix from an analysis based on (2.1). Since η is not of interest, this fitted model is simply used to generate complete data which are then analysed with the model of interest. The particular steps in such analyses are described in the following paragraphs.

The approach has a Bayesian flavour in that after fitting $Y|X,V$ we sample from MVN($\widehat{\eta}$, $\mathscr{I}^{-1}(\widehat{\eta})$) to obtain another realisation of the 4×1 parameter vector $\widehat{\eta}$ which we denote by g^1. If the response for any individual is missing, then we simulate the binary response as a Bernoulli variate with probability expit$(g^1_0 + g^1_x X + g^1_v V + g^1_{xv} XV)$ using the respective covariate values. This yields *the first imputed value* for each individual with missing data, and we label the realised response y^1. After each individual with incomplete data in the dataset has a response simulated based on g^1, a second sample is drawn from MVN($\widehat{\eta}$, $\mathscr{I}^{-1}(\widehat{\eta})$) and labelled g^2. Using this value, one

samples a second value $Y^2 \sim \text{Bern}(\text{expit}(g_0^2 + g_x^2 X + g_v^2 V + g_{xv}^2 XV))$ for each person with a missing response data. This procedure is repeated m times until we have m "complete" datasets. For each of the m "complete" datasets we then fit the model of interest given by (2.2).

Let $\widehat{\alpha}_x^r$ denote the estimate of α_x from the rth imputed dataset and $\omega^r = \widehat{\text{var}}(\widehat{\alpha}_x^r)$ be the naive variance estimate ignoring the fact that some data had been imputed by simulation. The combined estimate of α_x obtained by multiple imputation is simply the average, so $\bar{\widehat{\alpha}}_x = \sum_{r=1}^m \widehat{\alpha}_x^r / m$ is the reported point estimate from multiple imputation. Let $\bar{\omega} = \sum_{r=1}^m \omega^r / m$ denote the average of the naive (within imputation) variance estimates, and let $\omega^* = (m-1)^{-1} \sum_{r=1}^m (\widehat{\alpha}_x^r - \bar{\widehat{\alpha}}_x)^2$ denote the variation between imputation samples. Rubin (1987) argues that the asymptotic variance of $\bar{\widehat{\alpha}}_x$ is $\text{var}(\bar{\widehat{\alpha}}_x) = \bar{\omega} + (1 + m^{-1})\omega^*$ and

$$\frac{\bar{\widehat{\alpha}}_x - \alpha_x}{\sqrt{\text{var}\left(\bar{\widehat{\alpha}}_x\right)}} \sim t_{u_m}$$

approximately, where the degrees of freedom are given by

$$u_m = (m-1)\left[1 + \frac{m\bar{\omega}}{(1+m)\omega^*}\right]^2 .$$

Wang and Robins (1998) prove consistency and derive the large sample properties of estimators arising from multiple imputation under correct model specification. More refinements to the estimated degrees of freedom have since been made (Barnard and Rubin, 1999; Lipsitz et al., 2002) and are implemented in SAS. We will not get into these issues here, but remark simply that one appeal of multiple imputation is the ability to make use of auxiliary variables such as V when constructing the imputation model. In the context of longitudinal data with missing at random processes (see Section 2.3), this can be achieved by adopting a joint model for the responses over time (e.g., a mixed model) and, while the primary analysis is to be based only on a final response, intermediate values can ensure a more suitable imputation process which may translate to more precise estimates of treatment effects and more powerful tests.

In the present setting with incomplete response data, if the selection model in an inverse weighted analysis and the imputation model for a multiple imputation analysis are correct, one would not expect the resulting estimators to differ much since they are both essentially generating pseudo-complete data in slightly different ways.

2.2.3 An Illustrative Simulation Study

Here we report on a simple simulation study to illustrate these methods. We let $p_R = 0.50$, $P(V = 1) = p = 0.5$, $\alpha_x = \log 1.5$, $\eta_v = \log 0.5$ and $\eta_{xv} = \log 2$. These specifications can be

used to obtain values for η_0 and η_x. Note that the true odds ratio $\exp(\alpha_x)$, which would be consistently estimated in the absence of missing data, is 1.5 in this formulation ($\alpha_x \approx 0.4055$). We then specify the missing data model as $\delta_x = 0$, $\delta_v = \log 2$, $\delta_{xv} = \log 4$, and ensure that $P(R = 1) = p_R = 0.50$, so 50% of subjects will have incomplete response data and there is a differential degree of association between Y and R in the control and treatment arms. The limiting value of a naive estimate of α_x is 0.4831 based on the earlier calculations, giving an asymptotic bias of approximately 0.0777.

Table 2.1: Simulation results of naive and adjusted analyses using inverse weighting (known and estimated weights) and multiple imputation with incomplete response data; $P(X = 1) = 0.5$; $P(V = 1) = 0.5$; $p_R = 0.5$; $\alpha_0 = 0$, $\alpha_x = \log(1.5)$ $\eta_0 = 0.347$, $\eta_x = 0.059$, $\eta_v = \log(0.5) = -0.693$, $\eta_{xv} = \log(2) = 0.693$; $p_R = 0.5$; $\delta_0 = -0.654$, $\delta_x = 0$, $\delta_v = \log(2) = 0.693$, $\delta_{xv} = \log(4) = 1.386$; Number of subjects = 500; Number of simulations = 2000

Method of Analysis	Parameter	Bias	ESE	ASE	ECP
CCA	α_0	-0.072	0.201	0.196	93.3
	α_x	0.076	0.268	0.260	93.1
IPW	α_0	-0.005	0.204	0.199	95.1
	α_x	0.009	0.278	0.274	94.1
IPW_2	α_0	-0.004	0.203	0.200	95.2
	α_x	0.008	0.279	0.275	94.3
MI	α_0	-0.004	0.203	0.195	94.2
	α_x	-0.004	0.281	0.277	94.2

CCA is Complete-Case Analysis; IPW is Inverse Probability Weighted analysis with known weights; IPW_2 is Inverse Probability Weighted analysis with estimated weights; MI is Multiple Imputation, which here involved the creation of $m = 20$ pseudo-complete datasets

Two thousand datasets of $N = 500$ individuals were simulated and the following analyses were carried out: i) a complete-case likelihood analysis using (2.7), ii) an inverse weighted analysis using (2.9) with weights known, iii) an inverse weighted analysis with weights estimated via logistic regression, and iv) multiple imputation with $m = 20$ and the imputation model based on $Y|X,V$. In all cases the response model was simply based on $Y|X$. The empirical biases, empirical standard errors (ESE), average asymptotic standard errors (ASE), and empirical coverage of nominal 95% confidence intervals (ECP) are reported in Table 2.1.

The empirical biases of the complete-case analyses (expected since $\eta_{xv} \neq 0$ and $\delta_{xv} \neq 0$) are apparent, and this leads to empirical coverage probabilities which are less than the nominal 95% level. The empirical biases from the inverse weighted analyses with known and estimated weights are negligible and the empirical coverage probabilities are compatible with the 95% level. The biases are similarly small for the estimators based on multiple imputation and the empirical coverage probabilities are compatible with the 95% level for these as well. Also noteworthy is the similarity in the standard errors of the estimates based on inverse weighting and multiple imputation.

2.2.4 Further Remarks

In many clinical settings there are a number of *ad hoc* alternative approaches for dealing with missing response data. In dermatology trials, for example, it is common to use so-called "non-responder" imputation (Gordon et al., 2006a; Reich et al., 2005a). If, as we have described here, the response $Y = 1$ indicates a successful response to treatment (e.g. alleviation of symptoms), then in non-responder imputation (NRI), individuals who do not provide a response are assigned a value $Y = 0$ (i.e. they did not remain in the trial and report an alleviation of symptoms). The rationale for this crude form of imputation may arise from the notion that anything other than completing the course of treatment and exhibiting a good clinical response is undesirable and hence should be treated as a failure. An intuitively appealing aspect of this form of imputation is that all patients randomised are utilised in the analysis. However with NRI, a naive estimator of the probability of a successful response given X is, in fact, consistent for the joint probability $P(Y = 1, R = 1|X)$; this reflects that individuals must both provide a response and the response must be successful. The validity of estimates achieved through this method depends, therefore, on the process giving rise to the missing data. If $R \perp Y, X$ estimates of response rates within treatment arms (and therefore also estimates of AD) are conservative in that they are down-weighted by the probability of a response being observed (in fact, we are consistently estimating $P(Y = 1|X) \cdot P(R = 1)$). When data are not missing completely at random, NRI analyses will not yield consistent estimates of RR, OR, or AD. Depending on the mechanism giving rise to the missing data (which is generally unknown), NRI analyses can lead to conservative (too small) or anti-conservative (too large) estimates of treatment effect. Despite this, NRI is commonly assumed to be a conservative method of analysis (Saurat et al., 2007).

When responses are continuous, the calculations discussed in previous sections can be carried out following similar principles; to make this clear we wrote the expressions in a general form using expectations and explicit probability statements in key places. With continuous responses, however, another common crude method of imputation is often used called "mean value" imputation. In this case the average value of the response (perhaps for that particular treatment arm, or overall) is assigned to individuals with missing responses. This strategy can also lead to conservative or anti-conservative estimates of treatment effect depending on the particular setting,

and naive standard errors will not typically reflect the effect of imputation.

The discussion of multiple imputation given earlier is often referred to as “parametric" multiple imputation since it relies on the explicit specification of a parametric model to simulate the imputed data for each dataset. Other versions of multiple imputation are often adopted which employ implicit models to exploit the data observed in the sample (Herzog and Rubin, 1983; Little and Rubin, 2002). If we consider a particular individual with a missing response, “non-parametric" multiple imputation involves finding individuals in the dataset who are similar to this particular person with respect to key attributes or a summary measure, and randomly selecting from the responses in this set of similar individuals (Schenker and Welsh, 1988; Reilly and Pepe, 1997). This sampling is done with replacement to make up multiple complete datasets. Here judgement is not required to specify a probability model for imputation of the response, but rather to identify the set of “similar" individuals for each individual with a missing response (Rubin, 1987). Matching, stratification or use of propensity scores are useful for this goal, and several procedures are available in common statistical packages to facilitate this.

2.3 Methods for the Analysis of Incomplete Longitudinal Data

2.3.1 Notation and Terminology

Consider a longitudinal study in which the plan is to assess each of n individuals over K distinct assessment times. Let $Y_i = (Y_{i1}, \ldots, Y_{iK})'$ denote the random variable corresponding to the response vector for individual i over the K assessments. Suppose that every individual under study has measurements taken on p baseline covariates so that subject i has baseline covariate vector $X_i = (X_{i1}, \ldots, X_{ip})'$. We assume X_i is completely observed, and let $P(Y_i|X_i)$ denote the probability model of interest.

We restrict attention to incomplete longitudinal data due to drop-out, and suppose that the last time an observation for individual i occurred was at time K_i; this is a random variable and we let k_i denote its realisation, as illustrated in Figure 2.2. We can then partition the response vector as $Y_i = (\bar{Y}_i, Y_i^-)$, where $\bar{Y}_i = (Y_{i1}, \ldots, Y_{iK_i})'$ is observed and $Y_i^- = (Y_{i,K_i+1}, \ldots, Y_{iK})$ is missing. Let $R_i = (R_{i1}, \ldots, R_{iK})'$ be the corresponding vector of missing data indicators, where $R_{ik} = I(k \leq K_i)$, $k = 1, \ldots, K$. We can therefore equivalently think of R_i as a random vector or K_i as a random variable. Little and Rubin (2002) and Rubin (1976) define three classes of missing data mechanisms for this context.

Data are said to be *missing completely at random (MCAR)* if missingness (failing to observe a value) does not depend on any observed or unobserved measurements, i.e. $P(R_i|Y_i, X_i) = P(R_i)$. Data are said to be *missing at random (MAR)* if, conditional on the observed data, missingness does not depend on the data that are unobserved; that is, $P(R_i|Y_i, X_i) = P(R_i|\bar{Y}_i, X_i)$. Data are

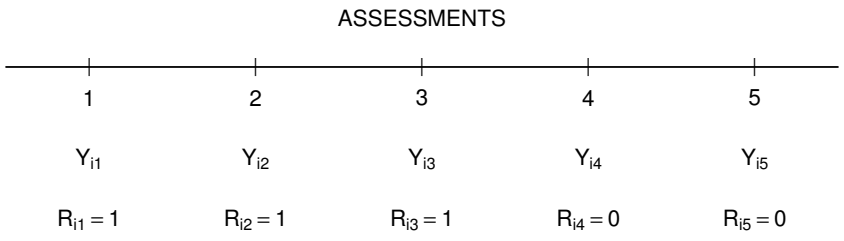

Figure 2.2: Schematic of schedule of assessments in longitudinal study with $K = 5$ for an individual with $K_i = 3$.

said to be *missing not at random (MNAR)* (or sometimes *not missing at random*) if missingness depends on the value of the realised (but unobserved) response, i.e. $P(R_i|Y_i, X_i)$ cannot be simplified. It is perhaps worth emphasising that these terms must be used and interpreted in the context of the available information (or at least the information being used); MNAR mechanism can become a MAR mechanism in light of additional information used judiciously.

2.3.2 Likelihood-Based Methods of Estimation and Inference

As in the univariate case, the likelihood for incomplete longitudinal data is developed by specifying the joint distribution of response variable Y_i and the missing data indicators R_i (or equivalently K_i), given the covariates X_i. Two classes of models have been proposed based on alternative factorisations of the joint distribution of $(Y_i, R_i)|X_i$ (Little, 1995): one is based on *selection models* (Little and Rubin, 2002), the other is based on *pattern mixture models* (Glynn et al., 1993; Little, 1993).

With selection models, the joint distribution of Y_i and R_i is factored as

$$P(R_i, Y_i|X_i; \alpha, \delta) = P(R_i|Y_i, X_i; \delta)\ P(Y_i|X_i; \alpha)\ , \tag{2.11}$$

where the distribution of R_i, $P(R_i|Y_i, X_i; \delta)$, is indexed by a vector of parameters δ and the distribution of Y_i, $P(Y_i|X_i; \alpha)$, is indexed by a vector of α.

With pattern-mixture models, the factorisation of the joint distribution is

$$P\left(R_i, Y_i|X_i; \alpha, \delta\right) = P\left(Y_i|X_i, R_i; \tau\right)\ P\left(R_i|X_i; \kappa\right)\ , \tag{2.12}$$

where in $P(Y_i|X_i, R_i; \tau)$, the distribution of Y_i, is defined separately for each missing data configuration and indexed by parameters τ, and the distribution of R_i, $P(R_i|X_i; \kappa)$, is known up to parameters κ.

When we are concerned with the parameters of the marginal distribution of Y, averaged over the missing data patterns, it is in many senses more natural to use selection models, because people

do not want to make inference conditional on the missing data indicators. In the following, we focus on selection models.

To describe the likelihood based approach we derive the joint density of the observed data $(\bar{Y}_i, R_i)$ by integrating out the missing data Y_i^- in the selection model of the joint distribution as

$$P(R_i, \bar{Y}_i | X_i; \delta, \alpha) = \int P(R_i | \bar{Y}_i, Y_i^-, X_i; \delta)\ P(\bar{Y}_i, Y_i^- | X_i; \alpha)\ dY_i^- .$$

Let $\bar{Y} = \{\bar{Y}_i, i = 1, 2, \ldots, N\}$ and $R = \{R_i, i = 1, 2, \ldots, N\}$ for a sample of size N. Then the observed-data joint likelihood for $(\delta', \alpha')'$ is

$$L(\delta, \alpha; \bar{Y}, R) = \prod_{i=1}^{N} \int P(R_i | \bar{Y}_i, Y_i^-, X_i; \delta)\ P(\bar{Y}_i, Y_i^- | X_i; \alpha)\ dY_i^- . \tag{2.13}$$

When the missing data mechanism is MAR, $P(R_i | \bar{Y}_i, Y_i^-, X_i) = P(R_i | \bar{Y}_i, X_i)$ and (2.13) becomes

$$\begin{aligned} L(\delta, \alpha; \bar{Y}, R) &= \prod_{i=1}^{N} \left\{ P(R_i | \bar{Y}_i, X_i; \delta) \int P(\bar{Y}_i, Y_i^- | X_i; \alpha)\ dY_i^- \right\} \qquad (2.14) \\ &= \prod_{i=1}^{N} \{ P(R_i | \bar{Y}_i, X_i; \delta)\ P(\bar{Y}_i | X_i; \alpha) \}\ . \end{aligned}$$

If the parameters δ and α are functionally independent, then likelihood inference for α from (2.14) is the same as a likelihood inference for α from the observed "partial" likelihood simply using the available data

$$L(\alpha; \bar{Y}) = \prod_{i=1}^{N} P(\bar{Y}_i | X_i; \alpha)\ . \tag{2.15}$$

Thus likelihood functions are unaffected by MAR mechanisms and this has contributed in part to the popularity of mixed effects models for the analysis of longitudinal data. If data are MNAR, then the simplification in (2.14) is not possible and we must use (2.13). This likelihood may lead to identifiability problems and so sensitivity analyses are often advocated for this case (Robins et al., 2000).

We remark that, as in the univariate case, one can sometimes identify an auxiliary covariate V_i which renders $R_i \perp Y_i^- | \bar{Y}_i, X_i, V_i$, so that inclusion of V_i in the analysis causes the missing data mechanism to be MAR. In this case, consider

$$\begin{aligned} P(R_i, \bar{Y}_i | X_i, V_i) &= \int P(R_i | \bar{Y}_i, Y_i^-, X_i, V_i)\ P(\bar{Y}_i, Y_i^- | X_i, V_i)\ dY_i^- \\ &= P(R_i | \bar{Y}_i, X_i, V_i)\ P(\bar{Y}_i | X_i, V_i)\ . \end{aligned}$$

This is only useful if we aim to estimate the effect of both X_i and V_i on the distribution of Y_i. Again, however, V_i may be useful for multiple imputation (as in Section 2.2.2) or for inverse weighting as we discuss in the next section.

2.3.3 Generalised Estimating Equations

Using standard notation for generalised linear models of binary data, we let $E(Y_{ik}|x_i) = P(Y_{ik} = 1|x_i) = \mu_{ik}$ and $\text{var}(Y_{ik}|x_i) = \mu_{ik}(1-\mu_{ik})$, $k = 1,\ldots,K$. Furthermore, we let $\Sigma_i(\alpha,\rho) = \text{cov}(Y_i|x_i) = \mathbb{A}_i^{\frac{1}{2}} Q(\rho) \mathbb{A}_i^{\frac{1}{2}}$ where $\mathbb{A}_i = \text{diag}\{\mu_{ik}(1-\mu_{ik}), k = 1,\ldots,K\}$ and $Q(\rho)$ is a $K \times K$ working correlation matrix with (k,k') entry, $Q_{kk'}(\rho)$, parameterised in terms of a vector of association parameters ρ. A marginal generalised linear model is formed by letting $g(\mu_{ik}) = x'_{ik}\alpha$ where $g(\cdot)$ is a known link function and $\alpha = (\alpha_0,\ldots,\alpha_p)'$ is a $(p+1) \times 1$ vector of regression coefficients.

Generalised estimating equations for α take the form

$$U(\alpha,\rho) = \sum_{i=1}^{N} U_i(\alpha,\rho) = 0 \tag{2.16}$$

where $U_i(\alpha,\rho) = G_i'(\alpha)\Sigma_i^{-1}(\alpha,\rho)(Y_i - \mu_i)$, with $\mu_i = (\mu_{i1},\ldots,\mu_{iK})'$ and $G_i(\alpha) = \partial\mu_i(\alpha)/\partial\alpha'$ a $K \times (p+1)$ matrix of derivatives (Liang and Zeger, 1986). If $\widehat{\alpha}$ is the solution for fixed $\rho = \rho_o$, then asymptotically $\sqrt{N}(\widehat{\alpha} - \alpha) \sim N(0, \text{var}(\sqrt{N}(\widehat{\alpha} - \alpha)))$ with

$$\text{var}(\sqrt{N}(\widehat{\alpha} - \alpha)) = [A^{-1}(\alpha,\rho_o)][B(\alpha,\rho_o)][A^{-1}(\alpha,\rho_o)]', \tag{2.17}$$

where $A(\alpha,\rho) = E(\partial U_i(\alpha,\rho)/\partial\alpha')$ and $B(\alpha,\rho) = E(U_i(\alpha,\rho)U_i'(\alpha,\rho))$. When ρ is not specified, estimation of α is facilitated by iteratively replacing ρ with a $\sqrt{N}$-consistent moment-type estimate based on estimates of α at successive iterations of a scoring algorithm (Liang and Zeger, 1986).

The functional form of $Q_{kk'}(\rho)$, $k \neq k'$, $k,k' = 1,\ldots,K$, is typically unknown, but even if the correlation structure is misspecified, consistent estimators of α arise from solving (2.16), and (2.17) will still hold. However, misspecification of the correlation structure in (2.16) can lead to inefficient estimators of α and, in more extreme cases, problematic asymptotic properties arise for the solution (Crowder, 1995). In many cases, the working independence assumption can yield quite efficient estimators (Sutradhar and Das, 1999), so we set $Q_{kk'}(\rho) = \rho_o = 0$ for $k \neq k'$ in what follows. An estimate of (2.17) is obtained in this case by computing

$$\widehat{\text{var}}(\sqrt{N}(\widehat{\alpha} - \alpha)) = [\widehat{A}^{-1}(\widehat{\alpha},\rho_o)][\widehat{B}(\widehat{\alpha},\rho_o)][\widehat{A}^{-1}(\widehat{\alpha},\rho_o)]', \tag{2.18}$$

where

$$\widehat{A}(\widehat{\alpha},\rho_o) = -N^{-1}\sum_{i=1}^{N} G_i'(\widehat{\alpha})\mathbb{A}_i^{-1}(\widehat{\alpha},\rho_o)G_i(\widehat{\alpha}),$$

and

$$\widehat{B}(\widehat{\alpha},\rho_o) = N^{-1}\sum_{i=1}^{N} G_i'(\widehat{\alpha})\mathbb{A}_i^{-1}(\widehat{\alpha},\rho_o)(y_i-\widehat{\mu}_i)(y_i-\widehat{\mu}_i)'\mathbb{A}_i^{-1}(\widehat{\alpha},\rho_o)G_i(\widehat{\alpha})\ .$$

As in the univariate case, however, this estimating equation approach is not appropriate when data are incomplete and not missing completely at random.

Selection models provide a natural framework for characterising factors which affect the risk of attrition in longitudinal studies. Let $R_{ik} = I(k \leq K_i)$. Selection models involve modelling the conditional probability of drop-out at each visit, which we denote here as $\lambda_{ik} = P(R_{ik} = 0 | R_{i1} = \cdots = R_{i,k-1} = 1, y_i, x_i)$. As mentioned in Section 2.3.1, the nature of the relation between this conditional probability of drop-out, covariates, and (possibly missing) responses determines the impact that drop-outs have on inferences regarding the regression coefficients in the response model. We restrict attention here to settings in which data are MAR, with any covariate dependence based only on previously observed covariates or responses. In this case, λ_{ik} may be a function of $\bar{Y}_i$ and X_i, but not of Y_i^-. Let $H_{ik}^y = \{y_{i1},\ldots,y_{i,k-1}\}$ be the history of the response up to time k. In practice, we typically let λ_{ik} depend on H_{ik}^y and X_i.

Since R_{ik} is a binary variable it is convenient to formulate logistic regression models for the conditional probability of drop-out given by

$$\log(\lambda_{ik}/(1-\lambda_{ik})) = w_{ik}'\delta^{(k)}\ , \tag{2.19}$$

where $\delta^{(k)} = (\delta_0^{(k)},\ldots,\delta_{q_k}^{(k)})'$ is a $(q_k+1)\times 1$ vector of regression coefficients characterising the nature of the relationship between w_{ik} and λ_{ik}, and w_{ik} is a covariate vector containing relevant observed information in H_{ik}^y and X_i.

The inverse-weighted estimating equations under the working independence assumption take the form

$$U(\alpha,\delta) = \sum_{i=1}^{N} U_i(\alpha,\delta) = 0 \tag{2.20}$$

where under cluster-specific weights as discussed by Fitzmaurice et al. (1995),

$$U_i(\alpha,\delta) = G_i'(\alpha)\Sigma_i^{-1}(\alpha)\Delta_i(\delta)(Y_i-\mu(X_i;\alpha)\ ,$$

$\Sigma_i(\alpha) = \mathrm{diag}\{\mu_{ik}(1-\mu_{ik}), k=1,\ldots,K\}$, $\Delta_i(\delta) = I(K_i=k_i)/\pi_i(\delta)$, and $\pi_i(\delta) = P(K_i=k_i|\bar{Y}_i,x_i;\delta)$. We often assume all subjects are available for the first assessment, so $\pi_i(\delta) = \lambda_{i2}(\delta)$ if $k_i = 1$, $\pi_i(\delta) = (1-\lambda_{i2}(\delta))\lambda_{i3}(\delta)$ if $k_i = 2$, $\pi_i(\delta) = (1-\lambda_{i2}(\delta))(1-\lambda_{i3}(\delta))$ if $k_i = 3$, etc. In practice, an estimate of δ can be obtained by fitting ordinary logistic regression models to the missing

data indicators as appropriate. Inserting $\widehat{\delta}$ into (2.20) gives estimating equations which can be solved for α in the usual fashion (Robins et al., 1995).

2.3.4 Naive Methods of Imputation for Incomplete Longitudinal Data

The "last observation carried forward" (LOCF) imputation approach for dealing with missing values due to drop-outs operates as follows: if $k_i < K$, missing observations at visits $k = k_i + 1, \ldots, K$ are replaced with the value of the most recently observed response (i.e. y_{ik_i}). To distinguish the actual (possibly latent) responses from the pseudo-responses used under this imputation scheme, we use Y_i^* to denote the response vector under LOCF imputation. Therefore $Y_{ik}^* = Y_{ik}$ for $k \leq k_i$ and $Y_{ik}^* = Y_{ik_i}$ for $k > k_i$, $k = 1, 2, \ldots, K$. We let y_{ik}^* denote the realised value of Y_{ik}^*. Assumptions made for the response Y_i are adopted for the pseudo-response Y_i^* since analyses are typically carried out under the assumption that they are in some sense equivalent. In fact, in most situations for which the assumptions regarding Y_i are true, they will not be true for Y_i^*, implying that the estimating equation (2.16) is misspecified for the pseudo response. As with the other naive imputation approaches discussed earlier, LOCF leads to inconsistent estimators in a wide variety of settings and can result in either conservative or anti-conservative estimates of treatment effect.

The frequency properties of estimators of α based on Y_i^* have been investigated under a wide range of settings by several authors (Cook et al., 2004; Prakash et al., 2008) based on the theory of misspecified models (White, 1982a; Rotnitzky and Wypij, 1994).

2.4 Methods for Dealing with Incomplete Covariate Data in Observational Studies

2.4.1 Likelihood Analyses

Now consider a similar estimation problem to the one considered in Section 2.2, but suppose that it is the covariate X which is subject to incomplete measurement. We again assume that Y is generated through (2.1) so $P(Y = 1|X, V; \eta) = \text{expit}(\eta_0 + \eta_x X + \eta_v V + \eta_{xv} XV)$, and we wish to estimate α in $P(Y = 1|X; \alpha)$ as defined in (2.2).

In this observational study setting, we let $R = I(X \text{ observed})$ indicate whether the covariate value of interest was recorded. The observed data likelihood can then be written as

$$L \propto P(Y, X, R = 1)^R \, P(Y, R = 0)^{1-R} \,, \tag{2.21}$$

where we can marginalise over X with $\sum_x P(Y,X=x,R=0)$ to obtain $P(Y,R=0)$, the contribution from individuals for whom X is unobserved.

As in the case of incomplete responses, the tendency is to focus on simple analyses such as those restricted to individuals with complete covariate data. In this case the adopted likelihood would be based on the response model with the implicit condition $R=1$ and so is proportional to

$$\begin{aligned} P(Y|X,R=1) &= \frac{P(R=1|Y,X)\ P(Y|X)}{\sum_y P(R=1|Y=y,X)\ P(Y=y|X)} \\ &= \frac{P(R=1|Y,X)}{P(R=1|X)}\ P(Y|X)\ . \end{aligned} \tag{2.22}$$

If $R \perp Y|X$, then (2.22) reduces to $P(Y|X)$ and a complete-case analysis will yield consistent estimators of α, but otherwise inconsistent estimators are obtained; we show this by example in the simulation studies that follow and in Section 2.5 we give a more in-depth examination of what is being estimated by such an approach. Note that with incomplete covariate data, missingness can depend on the potentially missing variable (X) and a complete-case analysis remains valid because it involves conditioning on this covariate; this is in contrast to the setting of missing responses where the missing data must be modelled. However even when valid, this complete-case analysis ignores the information contained in the responses from individuals with incomplete data, and therefore may result in less than optimal efficiency.

2.4.2 An EM Algorithm

If one makes assumptions regarding the distribution of the incomplete covariate in likelihood analyses based on (2.21), one can exploit information from individuals with $R=0$ and improve efficiency. To see this note that the second term in (2.21),

$$P(Y,R=0) = \sum_{x=0}^{1} P(Y|X=x)\ P(X=x)\ P(R=0|Y,X=x)\ ,$$

is indexed by α (as well as the parameters in $P(X)$ and those of the missing data process). If $P(R|Y,X) = P(R|Y)$ or $P(R)$, then the missing data process can be modelled using observed data. If $P(R|Y,X) = P(R|X)$, then while this is a desirable missing data process for complete-case analysis (see (2.22)), in this setting there is a need to make uncheckable assumptions about the missing data process, since the modelled dependence between R and X cannot be verified in general. Progress can be made here if an auxiliary variable V can be found which satisfies $R \perp X|V,Y$.

The assumptions that are needed to exploit information from individuals with $R=0$ could include the fully specified conditional covariate distribution, or simply its parametric form. In the latter

case, the EM algorithm offers a convenient method for estimation (Dempster et al., 1977). The complete data likelihood L_C corresponding to (2.21) is proportional to

$$[P(R|Y,X)\ P(Y|X)\ P(X)]^R\ \ [P(R|Y,X)\ P(Y|X)\ P(X)]^{1-R}\ .$$

We typically would work with the "partial" complete data likelihood

$$L_C \propto [P(Y|X)\ P(X)]^R\ \ [P(Y|X)\ P(X)]^{1-R} \tag{2.23}$$

under the assumption that the information regarding α in the missing-data model is negligible. Working with (2.23) then requires an expression for

$$P(X|R=0,Y) = \frac{P(R=0|Y,X)\ P(Y|X)\ P(X)}{\sum_x P(R=0|Y,X=x)\ P(Y|X=x)\ P(X=x)} \tag{2.24}$$

for the expectation step of the EM algorithm, which if $R \perp X|Y$ gives simply

$$\frac{P(Y|X)\ P(X)}{\sum_x P(Y|X=x)\ P(X=x)}\ . \tag{2.25}$$

It is clear from (2.25) that the partial complete data likelihood (2.23) can be used only if $P(R|Y,X) = P(R|Y)$.

Suppose now that $R \not\perp X|Y$ but there exists a completely observed covariate V which renders $R \perp X|Y,V$. Again for simplicity we assume V is binary with $P(V=1)=p$ and $P(V=0)=1-p$. Then the partial complete data likelihood can be expressed as

$$L_C \propto [P(Y|X)\ P(X)\ P(V|Y,X)]^R\ \ [P(Y|X)\ P(X)\ P(V|Y,X)]^{1-R} \tag{2.26}$$

(Horton and Laird, 2001), and the requirement for consistent estimation using the associated EM algorithm is then that $P(R|Y,X,V) = P(R|Y,V)$. We note however, that this approach also requires correct specification of models for $P(V|Y,X)$ and $P(X)$.

2.4.3 Multiple Imputation with Missing Covariates

Multiple imputation can be carried out using a model for $P(X|Y,V,R) = P(X|Y,V)$ and because $X \perp R|Y,V$, the model for $X|Y,V$ can be fitted based on individuals with complete data. Here this can be easily fitted using a saturated logistic regression model

$$P(X=1|Y,V) = \text{expit}(\zeta_0 + \zeta_y Y + \zeta_v V + \zeta_{yv} YV)\ . \tag{2.27}$$

Following the same arguments given in Section 2.2.2, for any given dataset we may carry out

multiple imputation of X based on the model $P(X|Y,V)$. We may also carry out multiple imputation in a less parametric manner by using a technique based on an implicit model, such as approximate Bayesian bootstrap imputation, which imputes missing values by repeatedly sampling with replacement from observations of individuals deemed to be within the same adjustment class as the individuals with a missing covariate (Rubin and Schenker, 1986; Little and Rubin, 2002). This approach is similar to hot-deck imputation (Reilly and Pepe, 1995), but is proper in the sense that variance estimates discussed earlier still apply (Little and Rubin, 2002).

2.4.4 Inverse Probability Weighted Estimating Functions

Inverse probability weighting can be used to obtain unbiased estimating functions for a complete-case analysis. If $P(R|Y,X,V) = P(R|Y,V)$, then we can write the inverse weighted estimating function as

$$U(\alpha) = \sum_{i=1}^{N} \frac{R_i}{P(R_i = 1|Y_i, V_i; \delta)} (Y_i - E(Y_i|X_i;\alpha)) \ [1, \ X_i]' \ , \tag{2.28}$$

and this can be shown to have expectation zero. Since the model in the weight indicates a dependence on (Y_i, V_i) which are always observed, this model can be fit and a $\sqrt{N}$-consistent estimator of δ inserted; a consistent estimator of α will then be obtained by setting (2.28) equal to zero and solving for α.

2.4.5 An Illustrative Simulation Study

Here we report on a simulation study designed to demonstrate the performance for several methods of dealing with missing covariates when the covariate X is incomplete and missingness is governed by the model

$$P(R = 1|Y,X,V;\delta) = \text{expit}(\delta_0 + \delta_v V). \tag{2.29}$$

We consider the response model (2.1) with $\eta_x = \log(1.5)$ and $\eta_v = 0$ in (2.2). We then consider two incomplete covariate settings – one where $\eta_{xv} = 0$ so $Y \perp V|X$ and, in turn, $Y \perp R|X$; and one setting where $\eta_{xv} = \log(4)$, so the complete dataset is response biased. We generate the binary X so that $P(X = 1|V) = \text{expit}(\beta_0 + \beta_v V)$, where $\beta_v = \log(1.5)$ and β_0 is chosen so that $P(X = 1) = 0.5$. We also set $P(V = 1) = 0.5$, $P(Y = 1) = 0.5$, $\delta_v = \log(1.2)$ and find δ_0 in (2.29) so that $P(R = 1) = 0.5$; so for 50% of subjects we would expect the covariate to be missing. We generated data for samples of 500 and 2000 individuals in each of 2000 simulated datasets. The analyses conducted included a complete-case analysis, an inverse probability weighted analysis with known weights, an inverse probability weighted analysis with weights estimated using

the true observation model $\pi_i = \text{expit}(\delta_0 + \delta_v V)$, an inverse probability weighted analysis with weights estimated using the model $\pi_i = \text{expit}(\delta_0^* + \delta_y^* Y + \delta_v^* V + \delta_{yv}^* YV)$, an EM algorithm which used saturated binomial models for modelling $P(V|Y,X)$ and $P(X)$, and multiple imputation. The imputation model adopted was a saturated logistic regression model for X given (Y,V), involving four parameters: the intercept, two main effects, and the two-way interaction. Simulations and analyses were carried out in R version 2.15.1 and SAS 9.2.

The empirical biases, empirical standard errors, average asymptotic standard errors, and empirical coverage probabilities are reported in Tables 2.2 and 2.3 for sample sizes of 500 (left column) and 2000 (right column). Table 2.2 corresponds to the case where $R \perp Y|X$; in Table 2.3, $R \not\perp Y|X$ but $R \perp Y|X,V$. For the purpose of comparison, we also include the results that could have been achieved if all covariates were available for analysis (that is if there were no missing values).

The results for the setting with $Y \perp R|X$ indicate that all methods yielded approximately unbiased estimates, close agreement between the empirical and average asymptotic standard errors, and empirical coverage that was compatible with the nominal 95% level; see Table 2.2. The efficiency gains realised by modelling the covariate distribution are apparent by comparing the standard errors from the complete-case analysis with those of the EM algorithm. Similar gains in efficiency were achieved through multiple imputation and through the modelling of the selection probabilities in the inverse probability weighted analysis using estimated post-stratified weights (i.e. using $\pi_i = \text{expit}(\delta_0^* + \delta_y^* Y + \delta_v^* V + \delta_{yv}^* YV)$). It is interesting to note that similar gains in efficiency over a complete-case analysis were not observed when weights were estimated using only V.

For Table 2.3, the empirical biases from the complete-case analyses expected due to (2.22) are apparent. The other methods yielded estimators with much smaller empirical biases and better coverage probability; the difference is especially noticeable with the smaller standard errors provided by the larger sample size. The empirical coverage probabilities for all valid methods were compatible with the nominal 95% level. The EM algorithm and MI approaches resulted in small biases and standard errors. It is interesting to note that the post-stratified inverse probability weighted analysis incorporating Y in the estimation of weights had empirical efficiency that was very similar to that achieved by the EM and MI approaches here, while the IPW and IPW_2 approaches were less efficient. The EM, MI, and IPW_{ps} approaches had relatively small standard errors and good coverage probabilities. However, a comparison with the efficiency of the analysis using full data shows that the loss in efficiency due to the missingness was significant even when these methods of analysis were employed.

Table 2.2: Simulation results of naive and adjusted analyses using inverse weighting, EM, and multiple imputation to account for an ignorably incomplete covariate X; $P(X = 1) = 0.5$; $P(V = 1) = 0.5$; $P(Y = 1) = 0.5$; $\eta_x = \log(1.5)$; $\eta_v = 0$; $P(R = 1) = 0.5$; $\delta_v = \log(2)$; Number of simulations = 2000

		Sample Size: 500				Sample Size: 2000			
Method	Parameter	Bias	ESE	ASE	ECP	Bias	ESE	ASE	ECP
					$Y \perp R\|X$ ($\eta_{xv} = 0$)				
CCA	α_0	-0.007	0.182	0.183	95.0	-0.002	0.091	0.091	95.0
	α_x	0.016	0.259	0.256	94.4	0.001	0.124	0.127	95.7
IPW	α_0	-0.006	0.185	0.185	94.9	-0.002	0.092	0.092	95.2
	α_x	0.014	0.263	0.260	94.5	0.001	0.126	0.129	95.6
IPW_2	α_0	-0.006	0.185	0.186	94.7	-0.002	0.092	0.092	95.3
	α_x	0.013	0.263	0.261	94.3	0.001	0.126	0.129	95.7
IPW_{ps}	α_0	-0.005	0.156	0.159	94.7	-0.001	0.077	0.079	96.2
	α_x	0.013	0.263	0.261	94.3	0.001	0.126	0.129	95.5
EM	α_0	-0.005	0.156	0.159	95.9	-0.001	0.077	0.079	96.2
	α_x	0.013	0.262	0.259	94.3	0.001	0.126	0.129	95.5
MI	α_0	-0.005	0.158	0.159	95.8	-0.001	0.079	0.079	95.5
	α_x	0.013	0.262	0.259	94.3	0.000	0.130	0.129	95.0
Full Data	α_0	-0.003	0.124	0.128	95.0	-0.002	0.063	0.064	95.4
	α_x	0.009	0.178	0.180	96.0	0.003	0.091	0.090	94.8

CCA denotes a Complete-Case Analysis; IPW denotes a Inverse Probability Weighted analysis with known weights; IPW_2 denotes a Inverse Probability Weighted analysis with estimated weights based on the correct model; IPW_{ps} denotes a Inverse Probability Weighted analysis with estimated weights based on the saturated model; EM denotes an EM algorithm; MI denotes Multiple Imputation, which here involved the creation of $m = 20$ pseudo-complete datasets; Full Data demonstrates the efficiency that would be possible if no covariates were missing

Table 2.3: Simulation results of naive and adjusted analyses using inverse weighting, EM, and multiple imputation to account for an ignorably incomplete covariate X; $P(X = 1) = 0.5$; $P(V = 1) = 0.5$; $P(Y = 1) = 0.5$; $\eta_x = \log(1.5)$; $\eta_v = 0$; $P(R = 1) = 0.5$; $\delta_v = \log(2)$; Number of simulations = 2000

		Sample Size: 500				Sample Size: 2000			
Method	Parameter	Bias	ESE	ASE	ECP	Bias	ESE	ASE	ECP
		$Y \not\perp R \mid X$ ($\eta_{xv} = \log(4)$)							
CCA	α_0	-0.009	0.189	0.188	95.1	-0.003	0.095	0.093	95.0
	α_x	0.136	0.270	0.268	92.4	0.117	0.132	0.133	86.5
IPW	α_0	-0.008	0.193	0.191	95.1	-0.003	0.097	0.095	94.8
	α_x	0.021	0.274	0.271	95.2	0.003	0.133	0.135	95.7
IPW_2	α_0	-0.008	0.193	0.191	95.2	-0.003	0.097	0.095	94.7
	α_x	0.020	0.271	0.271	95.5	0.004	0.133	0.134	95.7
IPW_{ps}	α_0	-0.008	0.164	0.165	95.5	-0.004	0.083	0.081	94.7
	α_x	0.020	0.272	0.271	95.5	0.004	0.132	0.134	95.7
EM	α_0	-0.008	0.164	0.164	95.5	-0.004	0.083	0.081	94.6
	α_x	0.020	0.272	0.269	95.5	0.004	0.132	0.134	95.6
MI	α_0	-0.010	0.167	0.164	95.0	-0.004	0.085	0.082	93.7
	α_x	0.025	0.281	0.267	94.4	0.004	0.137	0.133	94.6
Full Data	α_0	-0.004	0.128	0.131	96.0	-0.002	0.066	0.065	94.7
	α_x	0.012	0.186	0.187	95.0	0.001	0.091	0.093	95.7

CCA denotes a Complete-Case Analysis; IPW denotes a Inverse Probability Weighted analysis with known weights; IPW_2 denotes a Inverse Probability Weighted analysis with estimated weights based on the correct model; IPW_{ps} denotes a Inverse Probability Weighted analysis with estimated weights based on the saturated model; EM denotes an EM algorithm; MI denotes Multiple Imputation, which here involved the creation of $m = 20$ pseudo-complete datasets; Full Data demonstrates the efficiency that would be possible if no covariates were missing

2.4.6 A Note on Generalisability

When the response Y or covariate of interest X is not binary, specification of models for $E[Y|X,V;\eta]$ and $E[Y|X;\alpha]$ may be more complicated. In particular, when X is continuous, the models (2.1) and (2.2) may be incompatible in the sense that they cannot be simultaneously correct. In practice, it is generally only necessary to specify the response model of interest – here (2.2); the specification of (2.1) is presented only for ease of discussion. It would be necessary to specify the (potentially incompatible) expanded response model (2.1) when utilising a parametric imputation approach for the analysis of incomplete response data, however many non-parametric imputation approaches can be implemented which do not require this explicit model specification. Robins and Wang (2000) derived the asymptotic properties of estimators based on a fully-parametric imputation model that may be incompatible with the response model.

This approach of specifying both (2.1) and (2.2) in the setting where they are compatible is also utilised in the next section to allow for a precise discussion of the effects of model misspecification in the analysis of incomplete data.

2.5 Issues of Model Misspecification with Incomplete Data

In the discussion and simulations thus far, it has been assumed that the models were correctly specified. It is important to recognise however, that misspecification is routinely a problem and here we explore this important issue in the context of the methods we have discussed. Again suppose that interest lies in estimation of the parameter α in the mean model

$$\mu(X;\alpha) = E[Y|X;\alpha],$$

where the response model is in the exponential family (McCullagh and Nelder, 1989).

When response values are MAR, data can be analysed through the use of imputation or inverse weighting as demonstrated in Section 2.2. In different ways, these two approaches create rectangular datasets which can be analysed with available software: the inverse weighting approach restricts attention to individuals with complete responses and achieves consistent estimators by weighting contributions by the inverse of the probability of an individual being complete, while imputation creates a pseudo-complete dataset involving all individuals by replacing missing responses with imputed values.

With correct specification of the model assumptions associated with the handling of incomplete data (the model for the probability of being observed in the case of inverse weighting and the imputation model in the case of imputation), then the resultant estimators are consistent. The relative efficiencies of these methods have been compared in a variety of settings when the model assumptions are correct (Rubin, 1987; Reilly and Pepe, 1997; Seaman et al., 1999; Carpenter

et al., 2006). However, when the necessary model assumptions for handling incomplete data are incorrectly specified, it is less clear how these methods compare.

We explore here the limiting values of estimators resulting from inverse probability weighting and multiple imputation when model assumptions are violated. We derive the limiting values of estimators in the simple case where variables are binary and show the asymptotic bias of these estimators when model assumptions are incorrect. This comparison of asymptotic bias provides a greater understanding of the implicit effects of the assumptions in these analyses and allows for a comparison of robustness of the methods of analysis to model misspecification.

2.5.1 Asymptotic Behaviour of Estimators from Estimating Functions

Let $\widehat{\alpha}$ be the solution to the estimating equation $U(\widehat{\alpha}) = \sum_{i=1}^{N} U_i(\widehat{\alpha}) = 0$. The Taylor series expansion of this function is a first step towards deriving the limiting behaviour of the estimator under the mild regularity conditions discussed by Wild (1991). Since

$$U(\widehat{\alpha}) = U(\alpha) + I(\alpha)(\widehat{\alpha} - \alpha) + o_p(N^{-1/2}),$$

where $I(\alpha) = -\partial U(\alpha)/\partial \alpha'$ then with $U(\widehat{\alpha}) = 0$

$$(\widehat{\alpha} - \alpha) = I^{-1}U(\alpha) + o_p(N^{-1/2}).$$

If $E[U_i(\alpha)] = 0$, then by the Central Limit Theorem (Grimmett and Stirzaker, 2007)

$$N^{-1/2}U(\alpha) = N^{-1/2}\sum_{i=1}^{N} U_i(\alpha) \xrightarrow{D} MVN(0, B),$$

where $B = cov(U_i(\alpha))$, and by Slutsky's theorem (Resnick, 1999)

$$N^{-1/2}(\widehat{\alpha} - \alpha) \xrightarrow{D} MVN(0, A^{-1}BA^{-\prime}),$$

where $A = plim\, N^{-1}I(\alpha)$.

Therefore, $\widehat{\alpha}$ consistently estimates the parameter α satisfying $E[U_i(\alpha)] = 0$. These results are useful for deriving the limiting distribution of estimators obtained by solving estimating equations. This can be used to discover exactly what is being estimated when models for imputation or for inverse probability weighting are misspecified; one can derive the limiting value of parameter estimates for the weights when the selection model is misspecified, and given the resultant incorrect weights, the limiting values of the estimators arising from the inverse weighted estimating equations can be obtained. We pursue this and related issues in subsequent sections.

2.5.2 Misspecified Models for Handling Incomplete Responses

As before, we assume that there exists an auxiliary covariate V which is known for all individuals and which renders the response and missingness indicator conditionally independent; i.e. $Y \not\perp R|X$, but $Y \perp R|X,V$. Thus, observation of V results in a *missing at random (MAR)* mechanism, but if V is not available then response data are MNAR; conditioning on V in the response model renders the missing data mechanism *ignorable* (Little and Rubin, 2002).

We consider simple violations to the model assumptions necessary for achieving consistent estimators through inverse probability weighting and multiple imputation. These violations of model assumptions arise from ignoring interaction effects between V and X. We derive the limiting distributions of estimators arising from these methods of analysis and relate them to the true α (i.e. we will compare (2.31), (2.32), and (2.33) to (2.30)). Writing the limiting values of these estimators explicitly facilitates an understanding of how the estimators are affected by model violations and enables an assessment and comparison of asymptotic biases (that is, a comparison between what is being consistently estimated and what we want to be consistently estimating).

Suppose that Y,X,V are univariate and binary and arise according to the following models:

$$E[Y|X,V;\eta] = \text{expit}(\eta_0 + \eta_x X + \eta_v V + \eta_{xv} XV),$$
$$P(R=1|Y,X,V;\delta) = \text{expit}(\delta_0 + \delta_x X + \delta_v V + \delta_{xv} XV),$$

and suppose that X and V are independent.

The mean model of interest is $E[Y|X;\alpha] = \text{expit}(\alpha_0 + \alpha_x X)$, and the true value of α can be recovered from η and $P(V)$ by noting that

$$E[Y|X;\alpha] = E_V\{E[Y|X,V;\eta]\}. \tag{2.30}$$

Complete-Case Analysis

A natural estimator in the presence of incomplete data is one which restricts attention to only those individuals who provide complete information. That is, we could estimate α with the complete-case estimator $\widehat{\alpha}^{\text{cc}}$ which solves the estimating equation

$$0 = \sum_{i=1}^{N} R_i U_i(\alpha) = \sum_{i=1}^{N} R_i \left[Y_i - E[Y_i|X_i;\alpha]\right] [1,\ X_i]'.$$

The resulting estimator will consistently estimate the parameter of interest, α, whenever $Y \perp R|X$. Therefore complete-case (CC) analyses will yield consistent estimators of α for missing data mechanisms which are MCAR and for some MAR missing data mechanisms, since in these

special cases, the subsets of individuals that are completely observed are representative of the original sample.

Here, however, $\widehat{\alpha}^{\text{cc}}$ will not consistently estimate α since $(Y \not\perp R)|X$. The limiting value of $\widehat{\alpha}^{\text{cc}}$ is the solution to $0 = E[R_i U_i(\alpha)]$ where

$$E[Y|X;\alpha^{cc}] = E[Y|X,R=1;\eta,\delta] = \frac{E_V\left\{E[Y|X,V;\eta]P(R=1|X,V;\delta)\right\}}{E_V\left\{P(R=1|X,V;\delta)\right\}}, \tag{2.31}$$

since $X \perp V$ and $Y \perp R|X,V$. Note that if $Y \perp R|X$, which occurs if $Y \perp V|X$ or $R \perp V|X$ (that is, $\eta_v = \eta_{xv} = 0$ or $\delta_v = \delta_{xv} = 0$), then (2.31) reduces to (2.30) and the complete case estimator is consistent for the true parameter of interest (i.e. $\alpha^{cc} = \alpha$).

Inverse Probability Weighting

Analyses which restrict attention to completely observed individuals can be made more generally useful by reweighting the observations so they are representative of the original sample. This is possible for a MAR incomplete dataset by exploiting the observed variates V that render $Y \perp R|X,V$ to find the inclusion probability $\pi(X_i,V_i) = P(R_i = 1|X_i,V_i)$. The inverse probability weighted (IPW) estimator is the solution to the Horvitz-Thompson-style estimating equation

$$0 = \sum_{i=1}^{N} \bar{U}_i(\alpha) = \sum_{i=1}^{N} \frac{R_i\, U_i(\alpha)}{\pi(X_i,V_i)} = \sum_{i=1}^{N} \frac{R_i}{\pi(X_i,V_i)} \left[Y_i - E[Y_i|X_i;\alpha]\right] [1,\ X_i]'.$$

This IPW estimator will be consistent provided $Y \perp R|X,V$ and that $\pi(X_i,V_i)$ is correctly modelled and bounded away from 0 (Robins et al., 1994; Lawless et al., 1999). In this case, data will be MCAR within classes defined by (X,V); this missing data mechanism is sometimes called *quasi-randomisation* (Little and Rubin, 2002).

Suppose inverse probability weighting is used, but that the covariate V is not used properly to render the missing data mechanism MAR. We consider the situations where the model for missingness $\pi(X_i,V_i;\delta)$ is incorrectly modelled as

$$\pi(X_i,V_i;\delta^*) = \text{expit}(\delta_0^* + \delta_x^* X + \delta_v^* V)$$

where the interaction between X and V is omitted. A root of $E[\bar{U}_i(\alpha)]$ is then α^{ipw}, where

$$E[Y|X;\alpha^{\text{ipw}}] = \frac{E_V\left\{E[Y|X,V;\eta]P(R=1|X,V;\delta)/\pi(X,V;\delta^*)\right\}}{E_V\left\{P(R=1|X,V;\delta)/\pi(X,V;\delta^*)\right\}}, \tag{2.32}$$

since

$$
\begin{aligned}
0 &= E_{RYXV}\left\{\frac{R_i}{\pi(X_i,V_i)}(Y_i - E[Y_i|X_i;\alpha])\,[1,\ X_i]'\right\} \\
&= E_{YXV}\left\{\frac{P(R_i=1|X_i,V_i)}{\pi(X_i,V_i)}(Y_i - E[Y_i|X_i;\alpha])\,[1,\ X_i]'\right\}, \qquad \text{as } (R\perp Y)|X,V \\
&= E_{XV}\left\{\frac{P(R_i=1|X_i,V_i)}{\pi(X_i,V_i)}(E[Y_i|X_i,V_i] - E[Y_i|X_i;\alpha])\,[1,\ X_i]'\right\} \\
&= E_X\left\{\left[E_V\left\{\frac{P(R_i=1|X_i,V_i)}{\pi(X_i,V_i)}E[Y_i|X_i,V_i]\right\} - E_V\left\{\frac{P(R_i=1|X_i,V_i)}{\pi(X_i,V_i)}\right\}E[Y_i|X_i;\alpha]\right]\,[1,\ X_i]'\right\}.
\end{aligned}
$$

Thus if the missingness model is correctly specified so $\pi(X,V) = P(R=1|X,V;\delta)$ (i.e. if $\delta_{xv} = 0$, so $\delta^* = \delta$), then (2.32) reduced to (2.30) and the IPW estimator consistently estimates the true parameter of interest (i.e. $\alpha^{\text{ipw}} = \alpha$). Note that if V had been ignored entirely in the model for missingness (i.e. in the case that $\pi(X_i,V_i)$ is misspecified as $\pi(X_i,V_i;\delta^{**}) = \text{expit}(\delta_0^{**} + \delta_x^{**}X)$), then (2.32) would reduce to (2.31) (i.e. $\alpha^{\text{ipw}} = \alpha^{cc}$).

Multiple Imputation

As discussed in Section 2.2.2, analyses need not be restricted to individuals with complete data if it is possible to impute appropriate values for the missing data and then analyse the completed dataset. If we create a pseudo-complete set of responses Y^{imp} by imputing values for incomplete responses, α can be estimated by solving the pseudo-complete data estimating equation

$$
0 = \sum_{i=1}^{N} U_i^{imp}(\alpha) = \sum_{i=1}^{N}\left[Y_i^{imp} - E[Y_i|X_i;\alpha]\right]\,[1,\ X_i]'.
$$

There has been much research into how to properly impute these values and much care must be taken to ensure that variance estimates remain valid (Rubin, 1987; Xie and Paik, 1997; Schafer, 1999).

We consider here analogous misspecification to that defined for IPW where the covariate V is not fully exploited. Here, missing responses might be incorrectly imputed so that

$$
E[Y_i^{imp}|X_i,V_i] = E[Y_i|X_i,V_i;\eta^*] = \text{expit}(\eta_0^* + \eta_x^*X + \eta_v^*V)
$$

which ignores the interaction between X and V. This represents the use of regression imputation based on a misspecified model, or hot-deck or approximate Bayesian bootstrap imputation where adjustment classes are incorrectly specified (Little and Rubin, 2002).

In practice, this imputation estimator is found by solving

$$0 = \sum_{i=1}^{N} \left\{ R_i \left[Y_i - E[Y_i|X_i;\alpha] \right] + (1-R_i) \left[Y_i^{imp} - E[Y_i|X_i;\alpha] \right] \right\} [1,\ X_i]',$$

and a root of $E[U_i^{imp}(\alpha)]$ is α^{imp} where

$$E[Y|X;\alpha^{imp}] = E_V \left\{ P(R=1|X,V) E[Y|X,V] + [1-P(R=1|X,V)] E[Y^{imp}|X,V] \right\}.$$

We are supposing that the imputation model is misspecified so that $E[Y^{imp}|X_i,V_i] = E[Y_i|X_i,V_i;\eta^*]$, where the parameter η^* is consistently estimated using the misspecified estimating equation

$$0 = \sum_{i=1}^{N} R_i\ S_i(\eta) = \sum_{i=1}^{N} R_i\ [Y_i - E(Y_i|X_i,V_i;\eta)]\ [1,\ X_i,\ V_i]'.$$

So, η^* solves

$$\begin{aligned} 0 &= E[R_i S_i(\eta^*)] \\ &= E_{YXV} \left\{ P(R_i=1|X_i,V_i;\delta) \left[Y_i - E[Y_i|X_i,V_i;\eta^*] \right] [1,\ X_i,\ V_i]' \right\} \\ &= E_{XV} \left\{ P(R_i=1|X_i,V_i;\delta) \left[E[Y_i|X_i,V_i;\eta] - E[Y_i|X_i,V_i;\eta^*] \right] [1,\ X_i,\ V_i]' \right\}. \end{aligned}$$

Since X is binary, this gives $0 = E_V \left\{ P(R_i=1|1,V_i;\delta) \left[E[Y_i|1,V_i;\eta] - E[Y_i|1,V_i;\eta^*] \right] \right\}$ and $0 = E_V \left\{ \sum_{x=0}^{1} P(R_i=1|X_i=x,V_i;\delta) \left[E[Y_i|X_i=x,V_i;\eta] - E[Y_i|X_i=x,V_i;\eta^*] \right] P(X_i=x) \right\}$ so $P(R_i=1|0,V_i;\delta) \left[E[Y_i|0,V_i;\eta] - E[Y_i|0,V_i;\eta^*] \right] = 0$. Therefore $P(R_i=1|X_i,V_i;\delta) \left[E[Y_i|X_i,V_i;\eta] - E[Y_i|X_i,V_i;\eta^*] \right] = 0$ and so α^{imp} satisfies

$$\begin{aligned} E[Y|X;\alpha^{imp}] &= E_V \left\{ E[Y|X,V;\eta^*] + P(R=1|X,V) \left[E[Y|X,V;\eta] - E[Y|X,V;\eta^*] \right] \right\} \\ &= E_V \{ E[Y|X,V;\eta^*] \} \end{aligned} \tag{2.33}$$

when X is binary.

Note that the expected conditional mean (ECM) estimator discussed earlier estimates $E[Y|X;\alpha]$ through estimation of $E_V\{E[Y|X,V;\eta]\}$, so the misspecified imputation approach described here has the same asymptotic bias as the ECM estimator using η^* for η (i.e. when the estimating equation for η ignores the interaction term between X and V).

If the imputation model was correctly specified so that $E[Y|X,V;\eta] = E[Y|X,V;\eta^*]$ (i.e. $\eta_{xv} = 0$),

then (2.33) reduces to (2.30) and the imputation estimator consistently estimates the true parameter α generating the data. If missing responses are imputed using adjustment classes defined by X alone, then (2.33) reduces to (2.31) (i.e. $\alpha^{imp} = \alpha^{cc}$).

Augmented Inverse Probability Weighting

Robins et al. (1994) showed that the IPW estimating equation could be augmented to exploit the partial information available on individuals with incomplete data. As a result, an augmented inverse probability weighted (AIPW) estimator of the form

$$0 = \sum_{i=1}^{N} \bar{U}_i(\alpha) = \sum_{i=1}^{N} \frac{R_i}{\pi(X_i, V_i)} U_i(\alpha) - \frac{R_i - \pi(X_i, V_i)}{\pi(X_i, V_i)} \phi(X_i, V_i)$$

can be asymptotically more efficient than the IPW estimator. In the absence of further auxiliary covariates, the optimal choice for the augmentation function $\phi(\cdot)$ is $\phi_U^{\text{opt}} = E[U(\alpha)|X,V]$ (Robins et al., 1994; Tsiatis, 2006; Yu and Nan, 2006). In practice, the term ϕ_U^{opt} can be approximated by specifying an appropriate model for Y|X,V, as in the imputation approach. So, the AIPW estimator requires specification of both an 'imputation' model and a model for the inclusion probability, $\pi(X_i, V_i) = P(R_i|X_i, V_i)$. The AIPW estimator is 'double robust' in the sense that the estimator will remain consistent if either model is correctly specified. To see this note

$$E_{Y,X,V}\left\{ \frac{E_{R|Y,X,V}\{R_i\}}{\pi(X_i, V_i; \delta)} U_i(\alpha) - \frac{E_{R|Y,X,V}\{R_i\} - \pi(X_i, V_i; \delta)}{\pi(X_i, V_i; \delta)} \phi(X_i, V_i) \right\}$$

equals zero whenever $\pi(\cdot; \delta)$ is correctly modelled. Moreover, note that if the 'imputation' model is correct, so that $\phi_U^{opt}(X_i, V_i) = E_{Y|X,V}[U_i(\alpha)]$, then the AIPW estimating function is unbiased when data are MAR since

$$\begin{aligned} &E_{R,Y,X,V}\left\{ U_i(\alpha) + \frac{R_i - \pi(X_i, V_i; \delta)}{\pi(X_i, V_i; \delta)} U_i(\alpha) - \frac{R_i - \pi(X_i, V_i; \delta)}{\pi(X_i, V_i; \delta)} E_{Y|X,V}[U_i(\alpha)] \right\} \\ &\quad = E_{R,Y,V}\left\{ \frac{R_i - \pi(X_i, V_i; \delta)}{\pi(X_i, V_i; \delta)} \left(E_{Y|R,X,V}[U_i(\alpha)] - E_{Y|X,V}[U_i(\alpha)] \right) \right\} = 0. \end{aligned} \tag{2.34}$$

In practice, Robins et al. (1994) recommend estimating ϕ_U^{opt} with an empirical estimate of $E_{Y|X,V}[U_i(\alpha^{\text{ipw}})]$ based on the completely observed data. There has been some discussion recently about the utility of iteratively updating the estimate of ϕ_U^{opt} as α is estimated (Lumley et al., 2011; Scott and Wild, 2011a). In our simulations in Chapter 3, we demonstrate that this iteration results in perceivable small-sample efficiency gains for the AIPW estimator. Furthermore we note here that without this iterative updating, the expectation of the AIPW estimating

function in (2.34) is actually

$$E_{R,Y,V}\left\{\frac{R_i-\pi(X_i,V_i;\delta)}{\pi(X_i,V_i;\delta)}\left(E_{Y|R,X,V}[U_i(\alpha^{\text{aipw}})]-E_{Y|X,V}[U_i(\alpha^{\text{ipw}})]\right)\right\};$$

so the AIPW estimator is not unbiased if $\alpha^{\text{aipw}} \neq \alpha^{\text{ipw}}$, which will occur whenever the selection model $\pi(X_i,V_i;\delta)$ is incorrectly specified. Therefore, this iterative estimation of ϕ_U^{opt} is required to ensure that the AIPW estimator is double robust.

It remains of interest, however, to explore what is being estimated when both models are incorrectly specified. This question has seen some exploration through the use of simulation studies (see Bang and Robins, 2005; Kang and Schafer, 2007), but we attempt to provide a more comprehensive examination here by deriving the limiting value of the estimator under misspecified models.

We again consider the misspecified models considered above with the IPW and MI estimators. We suppose that $\pi(X_i,V_i;\delta)$ is incorrectly modelled as

$$\pi(X_i,V_i;\delta^*)=\text{expit}(\delta_0^*+\delta_x^*X+\delta_v^*V)$$

where the interaction between X and V is omitted; we further suppose that the model for the expectation of Y|X,V ignores the interaction between X and V and is modelled using

$$E[Y_i|X_i,V_i;\eta^*]=\text{expit}(\eta_0^*+\eta_x^*X+\eta_v^*V).$$

Then, the AIPW estimator is consistently estimating α^{aipw} which satisfies

$$\begin{aligned}
0=&E[\bar{\tilde{U}}_i(\alpha)]\\
=&E_{RYXV}\left\{\frac{R_i}{\pi(X_i,V_i;\delta^*)}(Y_i-E[Y_i|X_i;\alpha])\,[1,\ X_i]'-\right.\\
&\left.\frac{R_i-\pi(X_i,V_i;\delta^*)}{\pi(X_i,V_i;\delta^*)}E\left[(Y_i-E[Y_i|X_i;\alpha])\,[1,\ X_i]'\,|X_i,V_i;\eta^*\right]\right\}\\
=&E_{YXV}\left\{\frac{P(R_i=1|X_i,V_i;\delta)}{\pi(X_i,V_i;\delta^*)}(Y_i-E[Y_i|X_i;\alpha])\,[1,\ X_i]'-\right.\\
&\left.\frac{P(R_i=1|X_i,V_i;\delta)-\pi(X_i,V_i;\delta^*)}{\pi(X_i,V_i;\delta^*)}(E[Y_i|X_i,V_i;\eta^*]-E[Y_i|X_i;\alpha])\,[1,\ X_i]'\right\}
\end{aligned}$$

$$
\begin{aligned}
=&E_{XV}\left\{\frac{P(R_i=1|X_i,V_i;\delta)}{\pi(X_i,V_i;\delta^*)}(E[Y_i|X_i,V_i;\eta]-E[Y_i|X_i;\alpha]-E[Y_i|X_i,V_i;\eta^*]+E[Y_i|X_i;\alpha])\,[1,\ X_i]'+\right.\\
&\left.(E[Y_i|X_i,V_i;\eta^*]-E[Y_i|X_i;\alpha])\,[1,\ X_i]'\right\}\\
=&E_X\left\{E_V\{\frac{P(R_i=1|X_i,V_i;\delta)}{\pi(X_i,V_i;\delta^*)}(E[Y_i|X_i,V_i;\eta]-E[Y_i|X_i,V_i;\eta^*])+\right.\\
&\left.(E[Y_i|X_i,V_i;\eta^*]-E[Y_i|X_i;\alpha])\}\,[1,\ X_i]'\right\},
\end{aligned}
$$

so

$$
E[Y|X;\alpha^{aipw}]=E_V\left\{\frac{P(R=1|X,V;\delta)}{\pi(X,V;\delta^*)}\big(E[Y|X,V;\eta]-E[Y|X,V;\eta^*]\big)\right\}+E_V\left\{E[Y|X,V;\eta^*]\right\}. \tag{2.35}
$$

Note that if $\delta^*=\delta$ or $\eta^*=\eta$, then (2.35) would reduce to (2.30) and the AIPW estimator would consistently estimate the true parameter α generating the data (i.e. $\alpha^{aipw}=\alpha$).

Exploration of Asymptotic Biases under Model Misspecification

For further insight to the biases resulting from misspecification of the models in imputation (IMP) and inverse probability weighted (IPW) methods for analysing incomplete data, consider the asymptotic biases that result from a specified parameter set where $\eta_x=\delta_x=0$, $\eta_v=\delta_v=\log 1.2$, $P(Y=1)=.5, P(X=1)=.5$, and unless otherwise noted, $\eta_{xv}=\delta_{xv}=2$, $P(R=1)=.5$, and $P(V=1)=.5$; we also consider separately the effect of a range of values for η_{xv}, δ_{xv}, $P(R=1)$, and $P(V=1)$. The limiting values of estimators of relative risk ($P(Y=1|X=1;\alpha)/P(Y=1|X=0;\alpha)=\text{expit}(\alpha_0+\alpha_x)/\text{expit}(\alpha_0)$) using the above methods under misspecification by omission of the respective interaction terms can be found in Figure 2.3; the corresponding limiting values of estimators of odds ratios ($\exp(\alpha_x)$) are in Figure 2.4. For the purpose of comparison, we have also included the limiting values of the complete case (CC) estimator based on (2.31), as well as the true values and the double-robust augmented inverse probability weighted (AIPW) estimator.

Not surprisingly, all estimators were consistent when the required model assumptions (noted above) were satisfied (i.e. $\eta_{xv}=0$ for IMP; $\delta_{xv}=0$ for IPW; either $\eta_{xv}=0$ or $\delta_{xv}=0$ for AIPW). However, there are parameter combinations for which the estimators are consistent even when these assumptions are not satisfied; these conditions are sufficient, but not necessary for consistent estimation of the parameters. The necessary conditions for consistent estimation were presented in the previous section and are less restrictive than the usual assumptions associated with consistency of the methods of analysis; for example, the complete-case estimator of α_x is consistent when $\delta_{xv}=-0.166$ (see Panel 2 of Figure 2.4), despite the fact that these data are not

MAR. Furthermore, estimators of relative risks may be consistent even if neither α_0 nor α_x is consistently estimated; for example, Panel 2 of Figure 2.3 shows that the complete-case estimator of relative risk is consistent when $\delta_{xv} = -0.155$, but neither α_0 nor α_x are being consistently estimated here.

In general, there appears to be less asymptotic bias in the imputation estimator than the IPW estimator. It appears that the imputation approach here is more robust to misspecification of the imputation model than IPW is to misspecification of the missingness model when responses are incomplete. The asymptotic bias of the AIPW estimator is generally quite small when either of the model violations are small. However, there is no universally best method when models are misspecified; it is interesting to note that there are situations where the complete-case estimator has the smallest asymptotic bias. The asymptotic biases resulting from these misspecified models can be quite large when the violations to the model assumptions are large (i.e. when δ_{xv} and η_{xv} are far from 0) and when there is a large fraction of missing data. As is to be expected, the asymptotic bias of all methods decreases as the fraction of missing data decreases.

Examining the limiting values of estimators for odds ratios shows a very similar story (Figure 2.4). The complete-case estimator of the odds ratio generally demonstrates the largest asymptotic bias, while the limiting value of the imputation estimator is generally closer to the true value of α than the IPW estimator. The AIPW estimator generally suffers from the smallest amount of asymptotic bias, and the bias is quite small when either of the model assumptions are nearly correct. Again, however, no method of analysis is universally best. In all cases, the biases decrease as the percentage of missing data decreases and as the respective model violations diminish. It is also interesting to note that the direction of the asymptotic bias of the AIPW estimator is often different than the direction of the bias from the IPW and IMP approaches.

We have seen that for the considered set of parameters the imputation estimator appears to be more robust to model misspecification than the IPW estimator. However with non-binary data, the IPW estimator may become more appealing as the complexity of the imputation model necessarily increases – even with complex datasets, the IPW approach requires only modelling for binary response indicators, while imputation requires a model for the missing data itself. In such a situation, the AIPW estimator may also be quite appealing because it will be asymptotically more efficient than the IPW estimator while remaining consistent if the selection model is correctly specified. Bang and Robins (2005) suggest that this double robustness property will be advantageous even when both models are slightly misspecified, however, Kang and Schafer (2007) show that this is not always the case. We have shown here that the AIPW estimator can often result in less asymptotic bias than either imputation or an inverse probability weighting when both models are slightly misspecified. Seaman et al. (1999) propose methods combining IPW and MI to increase robustness to model specification while maintaining a high degree of efficiency.

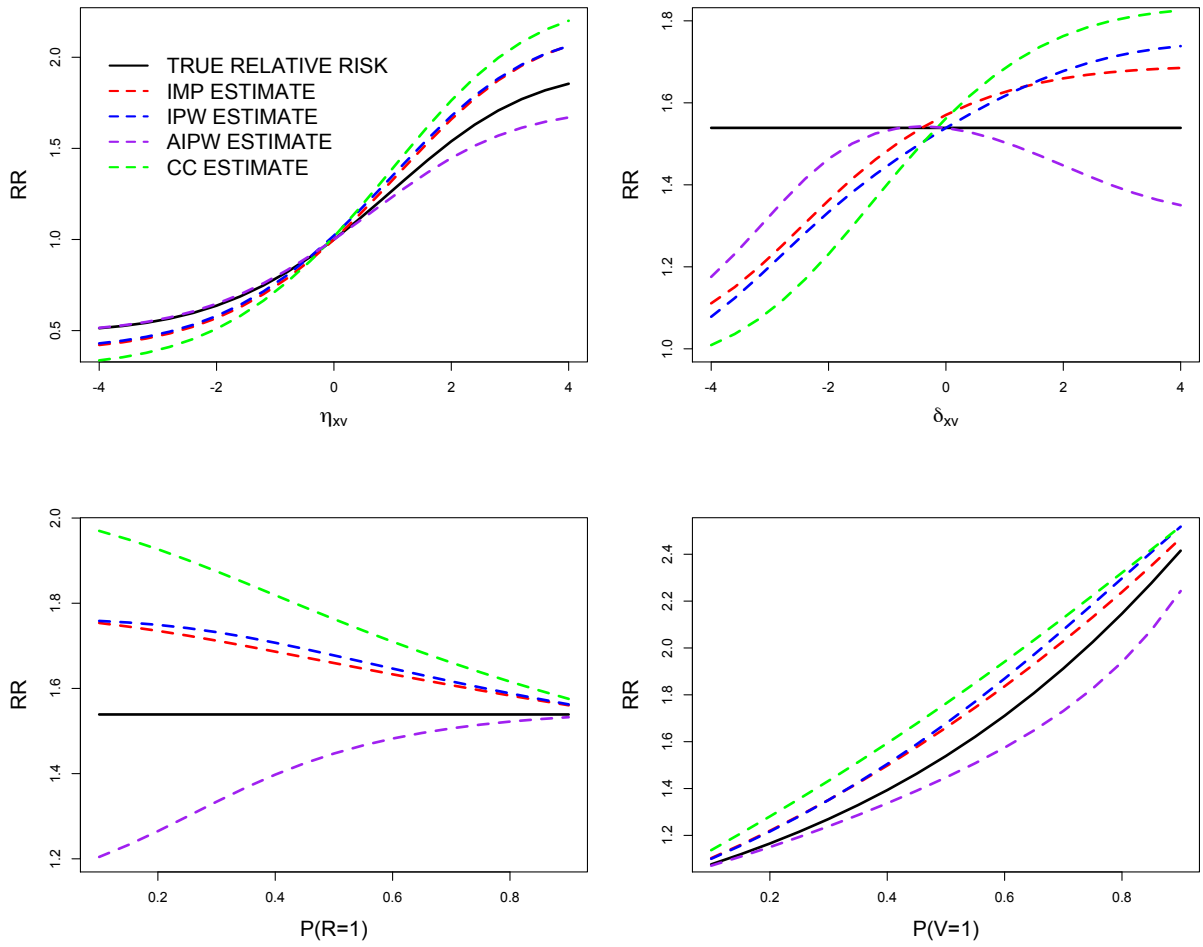

Figure 2.3: Limiting values of estimators of relative risk from analyses with misspecified models when responses are incomplete

2.5.3 Misspecified Models for Incomplete Covariate Data

Consider now the case where it is the covariate X which is incomplete. We let R_i be the indicator that X_i is observed and again consider analysis of these data through complete-case analysis, inverse probability weighting, multiple imputation, and augmented inverse probability weighting. We again suppose that there exists a covariate V which render the data MAR; that is, $X \not\perp R|Y$, but $X \perp R|Y,V$.

As before, the fully observed data are response biased within classes defined by X (i.e. $Y \not\perp R|X$), however here these data are also response biased within classes defined by both X and V (i.e.

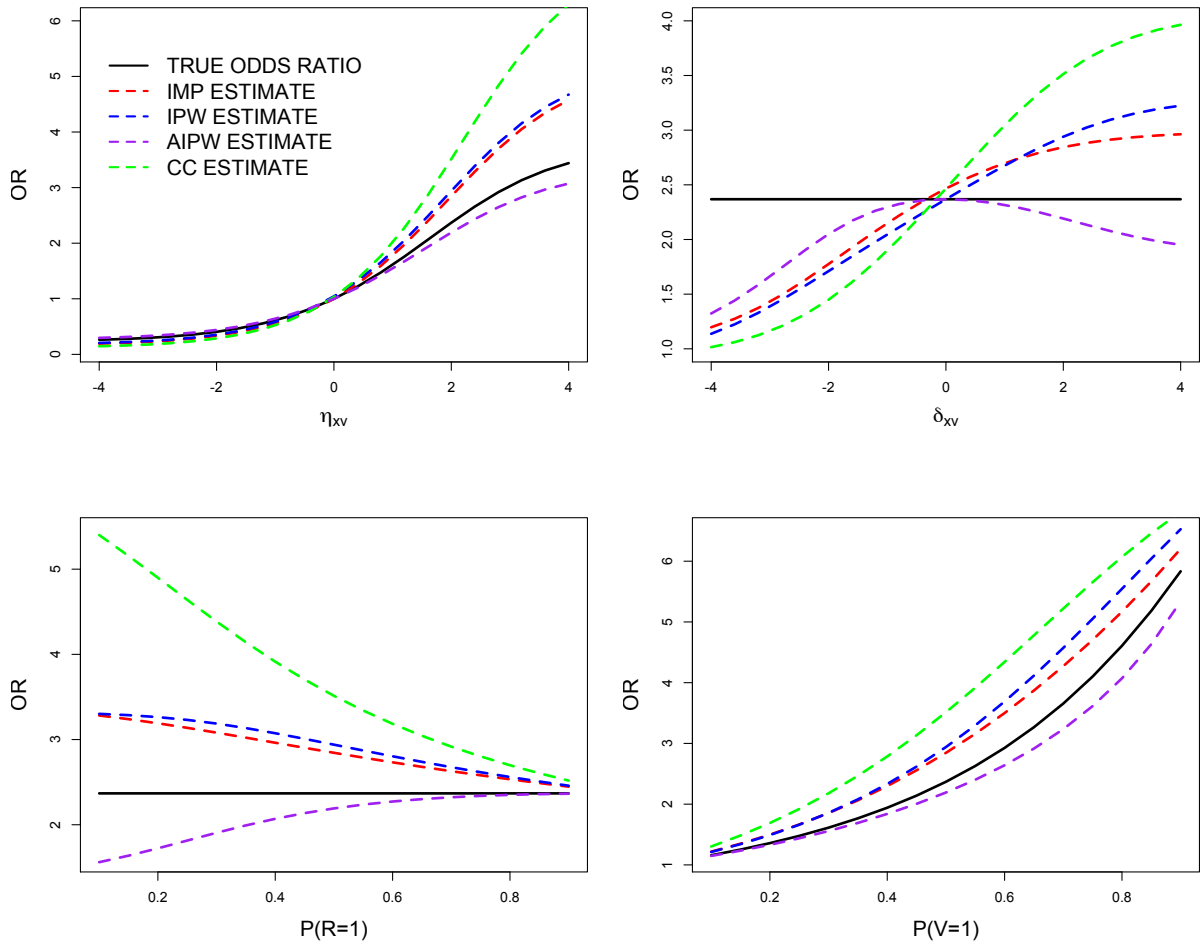

Figure 2.4: Limiting values of estimators of odds ratios from analyses with misspecified models when responses are incomplete

$Y \not\perp R|X,V$); that is, the missing data mechanism is not ignorable for estimation of α or η. Therefore, neither CC analyses nor ECM analyses will consistently estimate α, even if the models used in the ECM estimator are saturated. The presence of V does, however, allow for consistent estimation of α through inverse weighting or imputation.

IPW estimators will be consistent for α provided that the model for the probability of being observed is correctly specified as

$$\pi(Y_i,X_i,V_i) = P(R_i|Y_i,V_i;\delta).$$

Similarly, estimators of α based on imputation will be consistent provided that the imputation

model is correctly specified so that $E[X_i^{imp}|Y,V] = E[X|Y,V]$.

In Figures 2.5 and 2.6 we present the relative risk and odds ratio which are being estimated by an IPW analysis which omits the interaction term between Y and V in the model for π, an imputation analysis which omits the interaction term between Y and V in the imputation model for the missing X, and an AIPW estimator which incorporates both misspecified models. Note that the limiting value of this imputation estimator would now differ from that of the ECM approach. The estimating equation used to find the CC estimator is again

$$0 = \sum_{i=1}^{N} R_i U_i = \sum_{i=1}^{N} R_i \left[Y_i - E[Y_i|X_i;\alpha]\right] [1,\ X_i]'.$$

IPW analysis is carried out by solving the estimating equation

$$0 = \sum_{i=1}^{N} \bar{U}_i(\alpha) = \sum_{i=1}^{N} \frac{R_i}{\pi(Y_i,V_i;\delta^*)} \left[Y_i - E[Y_i|X_i;\alpha]\right] [1,\ X_i]',$$

where δ^* is found by solving

$$0 = \sum_{i=1}^{N} \left[R_i - \pi(Y_i,V_i;\delta)\right] [1,\ Y_i,\ V_i]'.$$

Note that if the missingness model was correctly specified so that the IPW estimator used weights found through $0 = \sum_{i=1}^{N} \left[R_i - \pi(Y_i,V_i;\delta)\right] [1,\ Y_i,\ V_i,\ Y_iV_i]'$, then the IPW estimator would be consistent for α.

The imputation approach involves solving

$$0 = \sum_{i=1}^{N} R_i \left[Y_i - E[Y_i|X_i;\alpha]\right] [1,\ X_i]' + (1-R_i) \left[Y_i - E[Y_i|X_i^{imp};\alpha]\right] \left[1,\ X_i^{imp}\right]',$$

to find α after imputing missing covariates based on a misspecified model where $P[X_i^{imp}|Y_i,V_i;\lambda^*]$ is found through

$$0 = \sum_{i=1}^{N} R_i \left[X_i - \text{expit}(\lambda_0^* + \lambda_y^* Y_i + \lambda_v^* V_i)\right] [1,\ Y_i,\ V_i]',$$

instead of the correctly specified saturated imputation model $0 = \sum_{i=1}^{N} R_i \left[X_i - \text{expit}(\lambda_0^* + \lambda_y Y_i + \lambda_v V_i + \lambda_{yv} Y_i V_i)\right] [1,\ Y_i,\ V_i,\ Y_iV_i]'$ that would lead to consistent estimators for α.

The AIPW estimator is found here by solving the estimating equation

$$0=\sum_{i=1}^{N}\frac{R_i}{\pi(Y_i,V_i;\delta^*)}\left[Y_i-E[Y_i|X_i;\alpha]\right][1,\ X_i]'-\frac{R_i-\pi(X_i,V_i;\delta^*)}{\pi(X_i,V_i;\delta^*)}E\left[\left[Y_i-E[Y_i|X_i;\alpha]\right][1,\ X_i]'\middle|Y_i,V_i;\lambda^*\right].$$

As in the missing response case, we consider the asymptotic biases that result from a specified parameter set. Here we consider $\eta_x=\delta_y=0,\ \eta_v=\delta_v=\log 1.2,\ \beta_v=\log(1.5),\ P(Y=1)=.5, P(X=1)=.5$, and separately consider a range of values for $\eta_{xv},\ \delta_{yv},\ P(R=1)$, and $P(V=1)$; unless otherwise noted $\eta_{xv}=\delta_{yv}=2$, $P(R=1)=.5$, and $P(V=1)=.5$, where as before we use $P(X=1|V)=\text{expit}(\beta_0+\beta_vV)$. The asymptotic biases for estimating the relative risk ($P(Y=1|X=1;\alpha)/P(Y=1|X=0;\alpha)=\text{expit}(\alpha_0+\alpha_x)/\text{expit}(\alpha_0)$) using the above methods can be found in Figure 2.5; the corresponding limiting values of estimators of odds ratios ($\exp(\alpha_x)$) are in Figure 2.6.

Here the limiting value of the AIPW estimator was generally the closest to the true value when the models were only slightly misspecified (i.e. when δ_{yv} was small in absolute value, and when λ_{xv} – or equivalently η_{xv} – was near zero). In the presence of greater model assumption violations, however, the IPW estimator often had less asymptotic bias than any of the other analyses.

2.6 Discussion

Incomplete data can arise in a number of settings for a variety of different reasons. Key factors influencing the extent of the impact on standard analyses are the proportion of missing data, and as demonstrated in this chapter, the nature of the stochastic mechanism which causes the data to be incomplete. Even when analyses are valid, loss of efficiency and decreased power are always issues. When possible, the extent of missing data should always be minimised.

When data are ignorably missing, then there is no need for imputation unless it is meant to recover power by exploiting information about the missing response from the available data. Naive imputation strategies (non-responder imputation, last observation carried forward, etc.) can introduce biases that can make interpretation of treatment effects difficult. Likelihood methods which have been developed and applied to minimise the effect of incomplete data are often directed at retrieving information about parameters of interest and improving power, but these come at the cost of making modelling assumptions beyond those typically made in analyses with complete data. These additional model assumptions are explicit, for example, when a parametric multiple imputation approach is adopted for incomplete response data. When covariates are missing and the EM algorithm is applied, one must make assumptions regarding the covariate distribution, which is not customary in routine analyses. When inverse probability weights are used, a model for the missing data process must be specified, which again is not something that is routinely

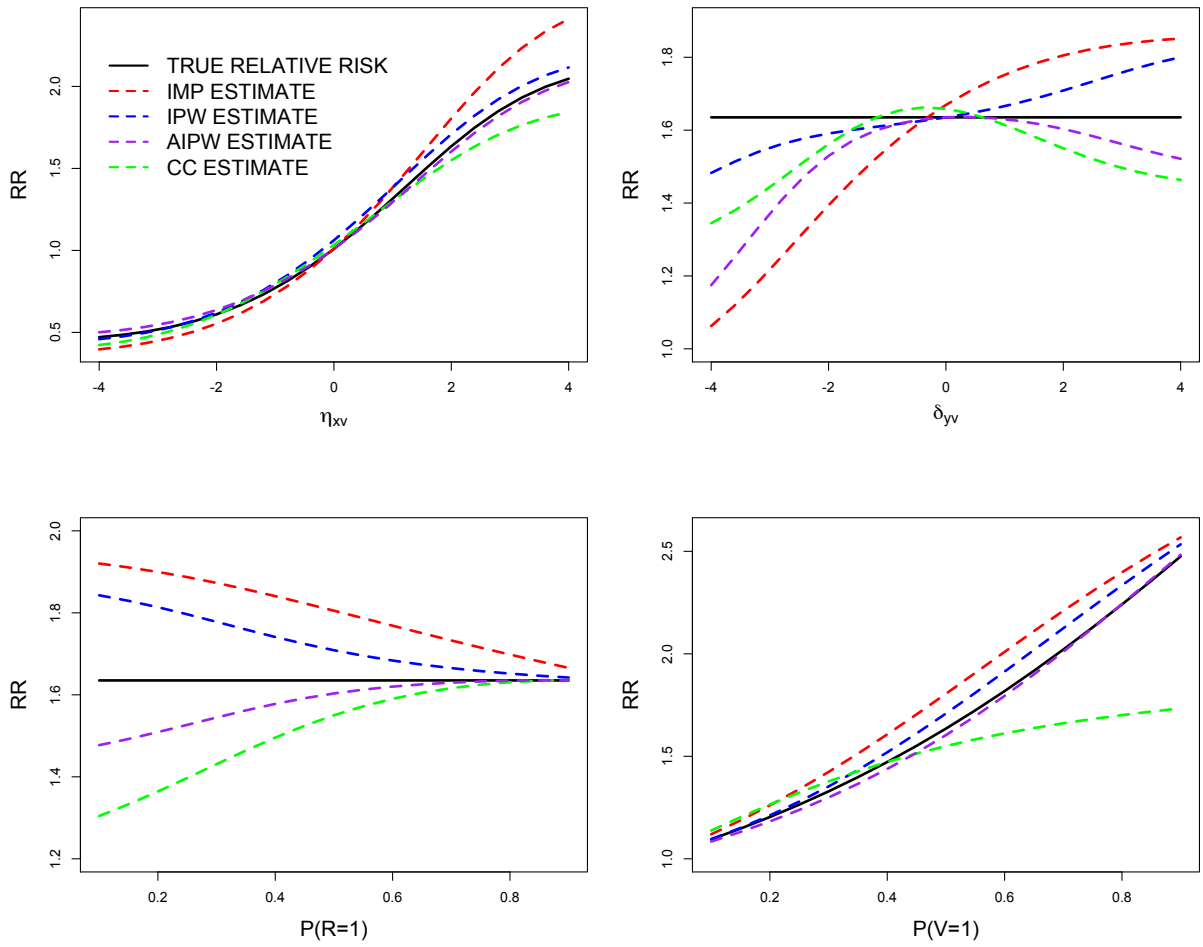

Figure 2.5: Limiting values of estimators of relative risk from analyses with misspecified models when covariates are incomplete

done in standard analyses. The specified models should be checked carefully since consistent estimators only result if these are correct.

When these models are incorrectly specified, it is difficult to characterise exactly what is being estimated. What is clear is that MI, IPW, and AIPW all worked reasonable well when violations to the relevant specified model were minimal and when the amount of missingness was small. In contrast to the findings of Kang and Schafer (2007), the double-robustness property of the AIPW estimator was beneficial here when neither model was correctly specified, but violations to either of the model assumptions were small. In our explorations, MI estimators were generally more robust to model misspecification than IPW estimators when responses were missing, but IPW

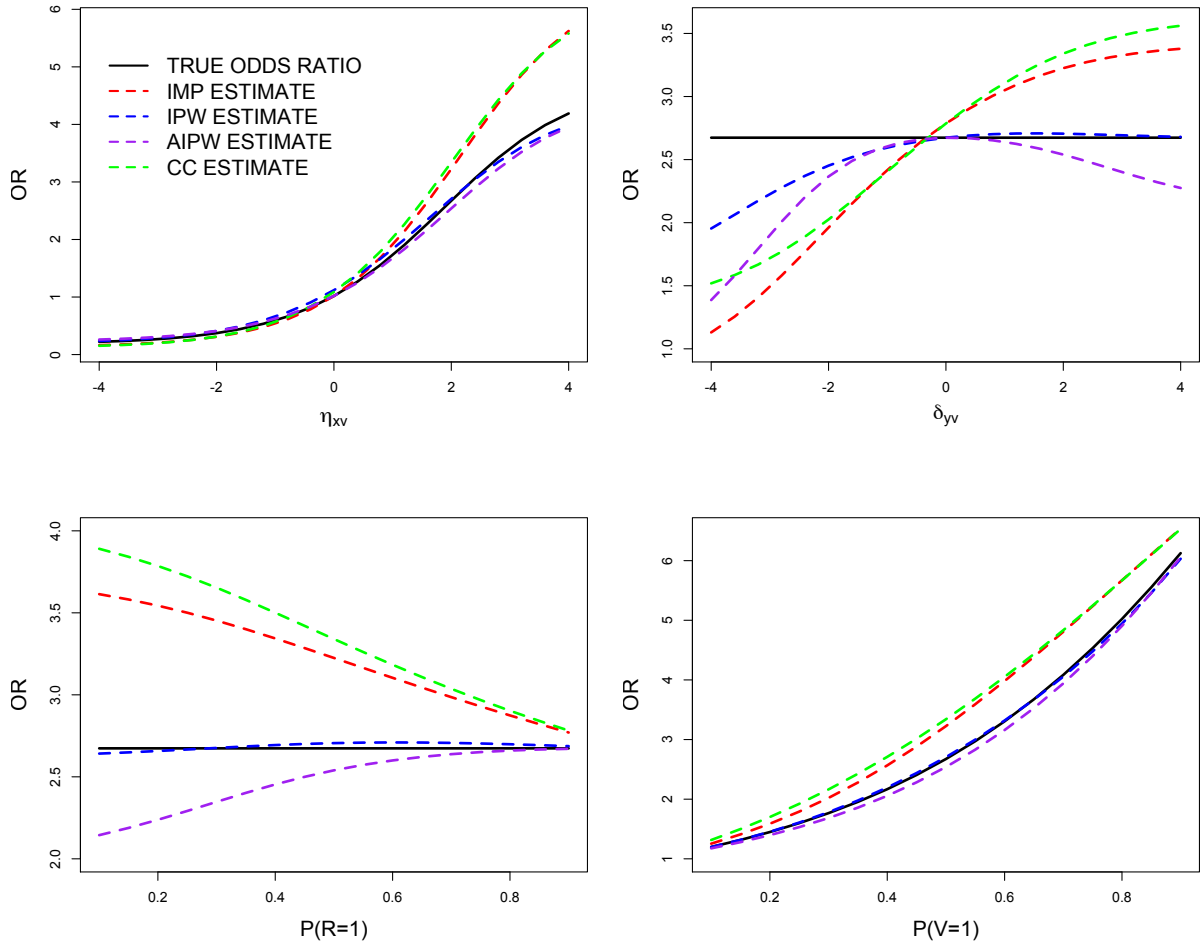

Figure 2.6: Limiting values of estimators of odds ratios from analyses with misspecified models when covariates are incomplete

estimators were generally more robust when covariates were incomplete. Both methods were generally preferable to the CC estimator, but the attempt to adjust for the incomplete nature of the data was far from universally beneficial when the nature of the missing data mechanism was misconstrued. However, these results support the assertions of Carpenter et al. (2006) that MI is more useful in the presence of incomplete responses, while IPW techniques are better suited to studies involving missing covariates.

We have emphasised simple models with binary data, primarily for transparency and so that explicit results would be easy to obtain. When incomplete variables are continuous, inverse probability weighting changes very little; this approach requires modelling the missing data in-

dicator which remains binary. Multiple imputation can be carried out in this case based on a linear regression model. The methods for longitudinal data can be similarly adapted. When incompletely observed variables are continuous or categorical, the necessary model assumptions for the EM algorithm or multiple imputation may become more involved and robustness of inferences becomes more of a concern. When multiple covariates are missing, high-dimensional joint models for the covariates are required and these can be challenging to specify and check. These challenges, in part, are reasons for the appeal of inverse probability weighted analyses of individuals with complete data.

We have considered the cases of a missing response or a single missing covariate separately. Frequently both responses and covariates can be missing in a given dataset and hybrid methods can be employed (Chen et al., 2010).

We have emphasised the setting in which interest lies in a regression model for a marginal mean parameter. In some settings, association parameters (e.g. correlations or odds ratios) are viewed as of comparable importance. This occurs when scientific interest lies in the nature of the association structure, or if concerns lie in optimising efficiency. In this case, regression models can be formulated for the association parameters and appropriate likelihood functions can be formed (Heagerty and Zeger, 2000; Heagerty, 2002). Zhao and Prentice (1990) describe how to do this using second order estimating equations. In the likelihood setting, the EM algorithm can be adopted and the idea of using inverse weighting for estimating association parameters can be adapted (Yi and Cook, 2002).

Chapter 3

Two-Phase Sampling Designs

3.1 Introduction

3.1.1 Study Design under Two-Phase Sampling Schemes

In two-phase studies involving response-dependent sampling, information is collected in the first phase on the outcome of interest and inexpensive covariates for a large number of individuals. A subset of individuals are then selected in a second phase of sampling, and the expensive exposure variables are measured for this sub-sample. The appeal of two-phase sampling designs lies in their ability to guide the efficient selection of individuals for the collection of expensive covariate data in the second phase through careful exploitation of inexpensive information available from the first phase of sampling.

Two-phase designs have received considerable attention in the statistical literature, and are being used increasingly often in public health research. Breslow and Chatterjee (1999) discuss two-phase designs in the context of the Wilms' tumour study, where interest lies in examining the association between outcome and tumour histology (classified as "unfavourable histology" or "favourable histology"). They demonstrate that by using the histological diagnoses of clinicians at patients' cancer centres as a surrogate for a more definitive diagnosis made at the National Wilms' Tumour Study Group Pathology Center, sampling schemes can be developed to ensure regression estimators remain very efficient while drastically reducing the need for central pathology assessments. In another setting, Clayton et al. (1998) show that estimation of the incidence of dementia in the elderly can also be made more efficient as a two-phase study by using Mini-Mental State Examination scores as a surrogate for definitive diagnoses of dementia (see also Zhao et al., 2009) . Robins et al. (1994) consider a two-phase design for a regression problem involving the Nurses Health Study (Stampfer et al., 1985). The response of interest was a binary indicator of myocardial infarction during a five-year period of followup. This response and a

surrogate for the expensive covariate were known for all individuals in a large (phase-I) sample. The expensive covariates, vitamin A and E measurements at study entry, were measured for a small (phase-II) sub-sample through the processing of stored sera. They show that the incomplete information available on phase-I individuals can often be exploited to achieve more efficient regression estimators than would be possible using only the phase-II sample in a complete-case analysis.

The efficiency of two-phase designs depends on how phase-I data are exploited at the second phase of sampling (Reilly and Pepe, 1995; Reilly, 1996; Schaubel et al., 1997). If available phase-I data are ignored during the design of the second phase, for example, then the phase-II sampling scheme is equivalent to simple random sampling. If the phase-II sample is chosen by stratification on a binary response variable indicating disease status, a case-control sampling scheme is obtained. Auxiliary covariate data are often also available in phase one. When these covariates are used along with response data, a more general two-phase sampling approach is used, which can lead to greater efficiency (White, 1982b; Breslow and Holubkov, 1997). The extent of the efficiency gain in any setting depends on the framework for the analysis and on how the phase-I data are exploited in deriving phase-II sampling probabilities.

Two-phase sampling designs have proven useful in epidemiology for ensuring efficient use of resources when estimating the effect of expensive or otherwise difficult to measure exposure variables on a response. Under such designs, a regression model is often of interest with a binary response indicating disease status and a covariate vector recording the exposure variable of interest along with possible auxiliary covariates. The first phase of sampling is used to collect data on the response and auxiliary covariates while a subset of these individuals is chosen at the second phase for measurement of the expensive exposure variable. The full sample ultimately contains a sub-sample with complete information on the outcome and all explanatory variables, while for the other individuals only the outcome and auxiliary covariates are available. Viewed as a whole, the full sample therefore features missing exposure data with the missing data mechanism determined by the nature of the phase-II sampling probabilities.

There is a wide range of statistical methods for dealing with regression based on incomplete covariate data including those based on maximum likelihood (Sprott, 2000), semiparametric maximum likelihood (Lawless et al., 1999; Zhang and Rockette, 2005), multiple imputation (Little and Rubin, 2002), mean score equations (Reilly and Pepe, 1995; Reilly, 1996) inverse probability weighted estimating functions, and augmented inverse probability weighted estimating functions (Robins et al., 1994; Tsiatis, 2006). These approaches differ in the nature of the assumptions required and the extent to which data from individuals with incomplete exposure data are used. Maximum likelihood, while potentially optimally efficient, requires one to model the distribution of the exposure variable given any auxiliary variables and misspecification of this model can lead to inconsistent estimates. The mean score method involves specifying unbiased estimating functions through approximating the conditional distribution of the exposure variable given the response and auxiliary variables using the phase-II sample. In their simplest form,

inverse probability weighted estimating equations restrict attention to individuals in the phase-II sample and hence do not require modelling of the covariate distribution. The resulting estimates are consistent provided the weights are correctly specified, but are typically less efficient than maximum likelihood estimates. Augmented inverse probability weighted estimating equations aim to improve efficiency by exploiting information in the individuals who only provide information in the phase-I sample. We explore asymptotic and empirical properties of each of these methods of analysis and derive phase-II sampling designs that optimally exploit available phase-I data.

When planning studies and attempting to optimise efficiency, the challenge is to specify the phase-II sampling model which will lead to the greatest precision of estimators for the parameters of interest; this is typically the coefficient of the exposure variable.

The remainder of the chapter is organised as follows. In Section 3.2, we introduce notation, specify the models, and formalise the problem of interest. In Section 3.3 we review several methods for fitting regression models with incomplete covariate data and give large sample properties of associated estimators. Section 3.4 contains guidelines for derivation of optimal designs for various methods of analysis, and simulation studies are presented to demonstrate the empirical efficiencies of the designs based on minimising the asymptotic variance of the parameter of interest. In Section 3.5, we explore the sensitivity of optimal designs to misspecification of design parameters, and consider the robustness of an optimal design to misspecification of the nuisance covariate model. Further, we examine the utility of optimal designs when necessary parameters are not known *a priori*, but are estimated using external pilot studies of varying sizes. Section 3.6 summarises these results and offers recommendations for the study of expensive biomarkers for progression of psoriatic arthritis.

Much of the work to date on two-phase designs involves univariate outcomes reflecting disease status. The purpose of the latter part of this chapter is to consider statistical issues in two-phase designs with more complex disease outcomes, motivated by our involvement in two studies discussed in Section 1.3. In Section 3.7 we consider the setting of clustered responses with cluster-level exposure and auxiliary variables. Marginal models (Liang and Zeger, 1986) are adopted in this setting with analysis frameworks based on maximum likelihood, and inverse probability weighted pseudo-likelihood. In Section 3.8 we give a framework for two-phase designs in longitudinal studies where interest lies in modelling the effect of an exposure variable on the onset of disease under a first-order Markov model. Asymptotic theory and optimal designs are provided for each setting. Concluding remarks and topics for further research are given in Section 3.9.

3.2 Design Of Studies With Two-Phase Sampling

Consider the setting where scientific interest lies in detecting and quantifying the effect of a new biomarker X on the mean of a categorical response Y while adjusting for a known categorical

prognostic variable V. This response model of interest is denoted

$$\mu(X,V;\alpha) = E[Y|X,V;\alpha] \tag{3.1}$$

and the conditional density of X given V is

$$g(X|V;\beta), \tag{3.2}$$

$P(V;\gamma)$ denotes the marginal probability mass function of V, and we let $\theta = (\alpha',\beta')'$ and $\Psi = (\alpha',\beta',\gamma')'$. We suppose that Y and V are known for all individuals in a phase-I sample giving $\{(Y_i,V_i), i = 1,\ldots,N\}$, but due to budgetary constraints the covariate X can only be observed for a subset of individuals. Let $R_i = 1$ if individual i is selected for inclusion in the phase-II sample (and hence for measurement of X_i), and let $R_i = 0$ otherwise. Thus, the data ultimately consists of N individuals: $n = \sum_{i=1}^{N} R_i$ of whom provide complete data (Y_i,X_i,V_i), and $(N-n)$ of whom provide information only on (Y_i,V_i).

Note that within this two-phase sampling framework, incomplete data arise by design and the researcher can control the sampling probabilities at the second phase through specification of the selection model

$$\pi(Y,V;\delta) = P(R = 1|Y,V;\delta); \tag{3.3}$$

we let $\Omega = (\alpha',\beta',\gamma',\delta')'$. We consider optimal two-phase designs, which involve the selection of individuals in phase-II who minimise the asymptotic variance of the estimator of α_x. For now we consider Bernoulli sampling (Lawless et al., 1999) or basic variable probability sampling (Lawless, 1997), wherein all sampling decisions are independent. The resulting data will then be *missing at random* (MAR) in the terminology of Little and Rubin (2002) since $P(R = 1|Y,X,V) = P(R = 1|Y,V)$.

As is the case with most design problems, determining optimal phase-II selection probabilities requires a specification of values for the parameters to be estimated, which are often chosen based on pilot data (Pepe et al., 1994; Reilly, 1996). This problem is sometimes ignored under the assumption that the optimal designs will not be sensitive to parameter misspecification (Whittemore and Halpern, 1997).

3.3 Frameworks For Analysis

3.3.1 Maximum Likelihood Estimation with Complete Data

If the data $\{(Y_i, X_i, V_i), i = 1, 2, \ldots, N\}$, were available for a random sample of size N from a large population, the corresponding complete-data conditional likelihood would be

$$L = \prod_{i=1}^{N} P(Y_i, X_i | V_i) = \prod_{i=1}^{N} P(Y_i | X_i, V_i; \alpha) \cdot g(X_i | V_i; \beta). \tag{3.4}$$

Provided β is functionally independent of α, the partial log-likelihood

$$\ell(\alpha) = \sum_{i=1}^{N} \ell_i(\alpha) = \sum_{i=1}^{N} \log P(Y_i | X_i, V_i; \alpha) \tag{3.5}$$

yields the maximum likelihood estimator $\widehat{\alpha}$, as the solution to the score equation

$$\sum_{i=1}^{N} S_i(\alpha) = \sum_{i=1}^{N} \partial \ell_i(\alpha) / \partial \alpha = 0,$$

and modelling of $X|V$ is not necessary if β is viewed as a nuisance parameter.

3.3.2 Maximum Likelihood With Covariates Missing At Random

With X_i known only for the subset of individuals with $R_i = 1$, efficient estimation for α can be based on the observed-data conditional likelihood

$$\prod_{i=1}^{N} \left\{ P(Y_i | X_i, V_i; \alpha) g(X_i | V_i; \beta) \pi(Y_i, V_i; \delta) \right\}^{R_i} \left\{ E_{X|V_i} \big[P(Y_i | X, V_i; \alpha) \big] (1 - \pi(Y_i, V_i; \delta)) \right\}^{1-R_i}$$

(Robins et al., 1994). Since R is completely observed, and δ and $\theta = (\alpha', \beta')'$ are functionally independent, we need only consider the observed-data partial likelihood

$$L(\theta) = \prod_{i=1}^{N} \left\{ P(Y_i | X_i, V_i; \alpha) g(X_i | V_i; \beta) \right\}^{R_i} \left\{ E_{X|V_i} \big[P(Y_i | X, V_i; \alpha) \big] \right\}^{1-R_i}, \tag{3.6}$$

which requires specification of both the response model (3.1) and the nuisance covariate model (3.2) (Lawless et al., 1999). The maximum likelihood estimate $\widehat{\theta}^{\text{ml}}$ is the solution to the score

equations

$$\sum_{i=1}^{N}\left\{R_i S_i(\alpha)+(1-R_i)E_{X|Y,V}[S_i(\alpha)]\right\}=0 \tag{3.7}$$

$$\sum_{i=1}^{N}\left\{R_i S_{i2}(\beta)+(1-R_i)E_{X|Y,V}[S_{i2}(\beta)]\right\}=0 \tag{3.8}$$

where $S_{i2}(\beta)=\partial \log g(X_i|V_i;\beta)/\partial\beta$. Equations (3.7) and (3.8) may be solved directly or via an EM algorithm (Dempster et al., 1977).

The limiting distribution of $\widehat{\theta}^{\text{ml}}$ depends on (3.1)-(3.3) such that asymptotically

$$\sqrt{N}(\widehat{\theta}^{\text{ml}}-\theta)\sim N(0,\mathbf{A}(\Omega)^{-1}),$$

where $\mathbf{A}(\Omega)=E\left[\mathscr{S}_i(\theta)\mathscr{S}_i'(\theta)\right]=E\left[-\partial\mathscr{S}_i(\theta)/\partial\theta'\right]$, and

$$\begin{aligned}\mathscr{S}_i(\theta)=&\partial\log\{[f(Y_i,X_i|V_i;\theta)]^{R_i}[E_{X|V_i}[P(Y_i|X,V_i;\theta)]]^{1-R_i}\}/\partial\theta\\ =&R_i\left[\frac{\partial\log P(Y_i|X_i,V_i;\alpha)}{\partial\theta}+\frac{\partial\log g(X_i|V_i;\beta)}{\partial\theta}\right]+(1-R_i)\frac{\partial\log E_{X|V_i}[P(Y_i|X,V_i;\theta)]}{\partial\theta}\end{aligned}$$

is the score function corresponding to the observed-data likelihood in (3.6). Note that $\mathbf{A}(\Omega)$ is a function of the full parameter set Ω since the expectation is taken with respect to (R_i,Y_i,X_i,V_i).

When the covariate X is continuous, the ML approach based on (3.6) is less appealing as consistent estimators of α require correct specification of a model for the nuisance distribution of $X|V$ (Pepe, 1992; Reilly and Pepe, 1995; Robins et al., 1995). This problem can be overcome through the use of the semiparametric restricted maximum likelihood (SPML) estimator, which involves maximisation of the likelihood

$$L(\alpha,G)=\prod_{i=1}^{N}\left\{P(Y_i|X_i,V_i;\alpha)G(X_i|V_i)\right\}^{R_i}\left\{P(Y_i|V_i;G,\alpha)\right\}^{1-R_i},$$

over the set of all discrete distributions G supported by the observed values of X (Lawless et al., 1999; Zhang and Rockette, 2005). This likelihood can be maximised using an EM algorithm (Zhao et al., 2009) or a profile likelihood approach (Breslow and Holubkov, 1997; Scott and Wild, 1997). Variance estimates can be obtained by inverting the observed information matrix produced by the profile likelihood approach (Lawless et al., 1999) or through numerical differentiation of the log-likelihood (Zhao et al., 2009).

Multiple Imputation (MI) can also be useful in the presence of an incomplete continuous covariate X as discussed in Chapter 2. With multiple imputation, we create complete pseudo-datasets that can be analysed to get "ML estimates" of α by maximising the complete-data partial log

likelihood (3.5) (Little and Rubin 2002). The imputation process is repeated m times creating m pseudo-datasets that can be used to find "ML estimates" $\widehat{\alpha}^1, \ldots, \widehat{\alpha}^m$, which can be averaged to obtain one final estimate $\bar{\widehat{\alpha}}$. The variance of this MI estimator can be estimated as

$$\bar{\omega} + (1 + m^{-1})\omega^*,$$

where $\bar{\omega}$ is the average within-imputation variance arising from the ML estimations and $\omega^* = (m-1)^{-1}\sum_{r=1}^{m}(\widehat{\alpha}^r - \bar{\widehat{\alpha}})'(\widehat{\alpha}_r - \bar{\widehat{\alpha}})$ is the between-imputation variance. In the simulations that follow, we use approximate Bayesian bootstrap imputation (Rubin and Schenker, 1986; Little and Rubin, 2002).

3.3.3 Inverse Probability Weighted Estimating Equations

Greater robustness can be achieved if we avoid specification of the covariate distribution by restricting attention to individuals with complete data. Suppose $U_i(\alpha) = h(X_i, V_i)[Y_i - \mu(X_i, V_i; \alpha)]$ is an unbiased estimating function for α with complete data. With incomplete data, the complete-case estimating equation $\sum_{i=1}^{N} R_i U_i(\alpha) = 0$ yields an inconsistent estimator of α if $R_i \not\perp Y_i | V_i$. A consistent estimator can be obtained, however, if the estimating function is weighted according to the selection model (Robins et al., 1994). This leads to the inverse probability weighted (IPW) estimating equations,

$$\sum_{i=1}^{N} \bar{U}_i(\alpha, \delta) = \sum_{i=1}^{N} \frac{R_i}{\pi(Y_i, V_i; \delta)} U_i(\alpha) = 0, \tag{3.9}$$

where the selection probabilities for phase-II, $\pi(Y_i, V_i; \delta)$, are assumed to be bounded away from zero. For the remainder of this chapter we assume that the estimating function is taken to be the score function from the complete-data likelihood; for example, with a binary response arising according to the logistic model $E[Y|X,V;\alpha] = \text{expit}(\alpha_0 + \alpha_x X + \alpha_v V)$ we assume that $U_i(\alpha) = S_i(\alpha) = (1, X_i, V_i)'[Y_i - \mu(X_i, V_i; \alpha)]$, which is the optimal estimating equation for these logistic regression parameters in the presence of complete data (Tsiatis, 2006).

The estimating function in (3.9) is unbiased whenever the $\pi(Y_i, V_i; \delta)$ are correctly specified since

$$E\left\{ \frac{R_i}{\pi(Y_i, V_i; \delta)} U_i(\alpha) \right\} = E_{Y,X,V}\left\{ \frac{E_{R|Y,X,V}\{R_i\}}{\pi(Y_i, V_i; \delta)} U_i(\alpha) \right\} = E_{Y,X,V}\left\{ U_i(\alpha) \right\} = 0.$$

Note that it is not necessary to model the nuisance covariate model (3.2) here, so the IPW approach is potentially more robust than the ML approach. A drawback of the IPW approach, however, is that none of the partial information available from the incomplete observations is exploited, and so there results a loss of efficiency. Furthermore, when some phase-II selection

probabilities are close to zero, the IPW estimator can perform poorly as estimates are greatly influenced by observations taken from these less frequently sampled strata (Tsiatis, 2006).

If $\widehat{\alpha}^{\text{ipw}}$ is the IPW estimator of α that uses known stratum-specific selection probabilities defined by δ, then asymptotically

$$\sqrt{N}(\widetilde{\alpha}-\alpha) \sim N(0, \bar{\mathbf{A}}(\Psi)^{-1}\, \bar{\mathbf{B}}(\Omega)\, \bar{\mathbf{A}}(\Psi)^{-1}),$$

where

$$\begin{aligned}
\bar{\mathbf{A}}(\Psi) &= E\big[-\partial \bar{U}_i(\alpha,\delta)/\partial\alpha'\big] \\
&= E_{Y,X,V}\left\{ -E_{R|Y,X,V}\left[\frac{R_i}{\pi(Y,V;\delta)}\frac{\partial U_i(\alpha)}{\partial\alpha'}\right]\right\} \\
&= E_{Y,X,V}\left\{\frac{-\partial U_i(\alpha)}{\partial\alpha'}\right\},
\end{aligned}$$

and

$$\begin{aligned}
\bar{\mathbf{B}}(\Omega) &= E\big[\bar{U}_i(\alpha,\delta)\bar{U}_i'(\alpha,\delta)\big] \\
&= E_{Y,X,V}\left\{E_{R|Y,X,V}\big[\frac{R_i}{\pi(Y,V;\delta)^2}U_i(\alpha)U_i'(\alpha)\big]\right\} \\
&= E_{Y,X,V}\{\pi(Y,V;\delta)^{-1}U_i(\alpha)U_i'(\alpha)\} \\
&= \sum_{Y,V}\pi(Y,V;\delta)^{-1}\cdot P(Y,V;\Psi)\cdot E_{X|Y,V}\big[U_i(\alpha)U_i'(\alpha)\big].
\end{aligned}$$

The asymptotic variance of the IPW estimator can then be written as

$$\bar{\mathbf{A}}(\Psi)^{-1}\,\Big[\sum_{Y,V}\frac{P(Y,V;\Psi)}{\pi(Y,V;\delta)}\cdot E_{X|Y,V}\big[U_i(\alpha)U_i'(\alpha)\big]\Big]\,\bar{\mathbf{A}}(\Psi)^{-1},$$

where $\bar{\mathbf{A}}(\Psi)$ is functionally independent of δ.

3.3.4 Inverse Probability Weighted Estimating Equations With Estimated Weights

It is also possible to proceed using an IPW analysis where the weights are estimated, so that the IPW_2 estimator, $\widehat{\alpha}^{\text{ipw}_2}$, is found by solving

$$\sum_{i=1}^{N}\bar{U}_i(\alpha;\widehat{\delta}) = \sum_{i=1}^{N}\frac{R_i}{\pi(Y_i,V_i;\widehat{\delta})}U_i(\alpha) = 0. \tag{3.10}$$

The estimator arising from (3.10) remains consistent provided the sampling probabilities $\pi(Y,V;\delta)$ are replaced with a $\sqrt{N}$-consistent estimator $\pi(Y,V;\widehat{\delta})$, which in turn hinges on correct specification of a parametric selection model. Interestingly, the estimating function using $\pi(Y,V;\widehat{\delta})$ will be at least as efficient as (3.9) asymptotically (Robins et al., 1994; Whittemore and Halpern, 1997; Lawless et al., 1999). When Y and V are discrete, equivalent approaches have alternatively been proposed as weighted likelihood (Breslow and Holubkov, 1997; Lawless et al., 1999), the mean score method (Reilly and Pepe, 1995), and the Horvitz-Thompson estimating function (Whittemore and Halpern, 1997). Reilly and Pepe (1997) have shown that in this case this approach is also asymptotically equivalent to hot-deck imputation.

Reilly and Pepe (1995) show that the asymptotic variance of this IPW_2 estimator can be written as

$$\bar{\mathbf{A}}(\Psi)^{-1} + \bar{\mathbf{A}}(\Psi)^{-1}\bar{\mathbf{B}}^*(\Omega)\ \bar{\mathbf{A}}(\Psi)^{-1},$$

where

$$\bar{\mathbf{B}}^*(\Omega) = \sum_{Y,V} P(Y,V)[\pi(Y,V;\delta)^{-1} - 1] \cdot \text{var}_{X|Y,V}[U_i(\alpha)],$$

with $\text{var}_{X|Y,V}[U_i(\alpha)] = E_{X|Y,V}\left[U_i(\alpha)U_i'(\alpha)\right] - E_{X|Y,V}\left[U_i(\alpha)\right]E_{X|Y,V}\left[U_i'(\alpha)\right]$.

3.3.5 Augmented Inverse Probability Weighted Estimating Equations

Robins et al. (1994) introduced augmented inverse probability weighted estimating equations (AIPW) of the form

$$\sum_{i=1}^{N} \bar{\bar{U}}_i(\alpha,\delta) = \sum_{i=1}^{N} \frac{R_i}{\pi(Y_i,V_i;\delta)} \cdot U_i(\alpha) - \frac{R_i - \pi(Y_i,V_i;\delta)}{\pi(Y_i,V_i;\delta)} \cdot \phi(Y_i,V_i) = 0, \tag{3.11}$$

which can incorporate partial information from individuals only in the phase-I sample and lead to more efficient estimators than those arising from (3.9) or (3.10).

Robins et al. (1995) and Tsiatis (2006) show that AIPW estimators are consistent and asymptotically normal with asymptotic variance equal to the variance of the corresponding influence function. Therefore, if $\widehat{\alpha}^{\text{aipw}}$ is the AIPW estimator found by solving (3.11), then $\sqrt{N}(\widehat{\alpha}^{\text{aipw}} - \alpha)$ is asymptotically normal with mean 0 and variance

$$\bar{\mathbf{A}}(\Psi)^{-1}\ \bar{\bar{\mathbf{B}}}(\Omega)\ \bar{\mathbf{A}}(\Psi)^{-1}, \tag{3.12}$$

where $\bar{\bar{\mathbf{B}}}(\Omega) = E\left\{\bar{\bar{U}}_i(\alpha,\delta)\bar{\bar{U}}_i'(\alpha,\delta)\right\}$

As discussed in the previous chapter, the optimal choice for the augmentation function $\phi(\cdot)$ in (3.11) here is $\phi_U^{\text{opt}} = E[U(\alpha)|Y,V]$ (Robins et al., 1994; Tsiatis, 2006; Yu and Nan, 2006). In the

simulation studies in Section 3.5, we approximate $\phi_U^{\text{opt}}(y,v)$ with the phase-II sample average

$$\widehat{E}[U(\widehat{\alpha})|y,v] = \frac{\sum_{i=1}^{N} R_i\, I(Y_i = y, V_i = v)\, U_i(\widehat{\alpha}^{\text{ipw}})}{\sum_{i=1}^{N} R_i\, I(Y_i = y, V_i = v)}, \tag{3.13}$$

where $\widehat{\alpha}^{\text{ipw}}$ is the IPW estimator of α. This estimation of ϕ_U^{opt} does not affect the asymptotic properties of the AIPW estimator (Robins et al., 1994). When selection probabilities are known, the estimating equation in (3.11) is particularly appealing as the AIPW estimator will, unlike the ML estimator, necessarily be consistent; the specification of $\phi(\cdot)$ will determine the efficiency of the estimators arising from (3.11).

Lumley et al. (2011) elucidate the relationship between this estimator and calibration estimators which have been developed in the context of survey sampling. In the discussion of Lumley et al. (2011), alternative methods for estimating ϕ_U^{opt} are offered (Lawless and Kalbfleisch, 2011; Scott and Wild, 2011a; Tsiatis and Davidian, 2011).

Note that while ϕ_U^{opt} allows for the most efficient possible estimation of α amongst estimating equations of the form (3.11) (that is, amongst estimating equations based on the optimal full-data estimating function $U_i = S_i$), we remark that it is possible in principle to achieve greater efficiency by deriving the optimal incomplete-data estimating function U^{eff} and its corresponding optimal augmentation term $\phi_{U^{\text{eff}}}^{\text{eff}}$ (Robins et al., 1994). This process, however, can be computationally intensive and require iterative methods (Robins et al., 1994; Tsiatis, 2006). Tsiatis (2006) recommends finding estimating equations that are optimal within a restricted class, and his optimal restricted class 1 estimators promise efficiency gains without requiring complex, iterative calculations as long as the selection probabilities are known (unlike other possible augmented inverse probability weighted estimating equations, however, these are not double robust). Implementation of these optimal restricted class 1 estimators requires the calculation of the matrix $\widehat{U}_{22}^{-1}$, where $\widehat{U}_{22} = N^{-1}\sum_{i=1}^{N}[R_i - \pi(Y_i,V_i;\delta)](1,Y,V,YV)'(1,Y,V,YV)$ if phase-I data are binary. This method implicitly requires the undesirable additional restriction that certain strata must not be sampled completely. For example, if we wish to sample all individuals from the stratum with $(Y=1,V=1)$, then we would set $\pi(1,1) = P(R=1|Y=1,V=1) = 1$. However, in this case $\widehat{U}_{22}$ (and the corresponding U_{22}) will necessarily be singular.

For the remainder of this paper, we restrict attention to the efficient, but potentially sub-optimal, augmented estimating equations of the form (3.11) that utilise $U_i = S_i$ and ϕ_U^{opt}. This estimating function is called the *efficient augmented estimator* by Kulich and Lin (2004).

3.4 Relative Efficiency Of Phase-Two Designs

3.4.1 Derivation of Phase-Two Sampling Designs

We consider first the case in which both Y and V are binary, yielding discrete phase-I data $\{(Y_i, V_i), i = 1, 2, \ldots, N\}$ which can be partitioned into four strata defined by the values of (Y, V), with respective sample sizes $\{N_{YV} : Y, V \in \{0, 1\}\}$. These strata sample sizes can be exploited in determining stratum-specific phase-II sampling probabilities, $\pi(Y_i, V_i; \delta)$. The response model of interest is

$$\mu(X, V; \alpha) = E[Y|X, V; \alpha] = \text{expit}(\alpha_0 + \alpha_x X + \alpha_v V).$$

We consider six sampling designs which exploit these phase-I data in different ways: simple random sampling, balanced sampling, optimal ML sampling, optimal IPW sampling, optimal IPW_2 sampling, and optimal AIPW sampling. In each optimal design, the selection models are derived to minimise the asymptotic variance of the estimator of α_x. These optimal designs require information on the phase-I stratum sizes and require specification of the (unknown) parameter values. In contrast, simple random sampling does not exploit the phase-I information, and balanced sampling only requires knowledge of the sizes of the phase-I strata.

Budgetary constraints limit the number of individuals that can be sampled in the second phase. We reflect these budgetary constraints by specifying some $0 < P_R \leq 1$ so that

$$P(R = 1; \delta) = \sum_{Y,V} \pi(Y, V; \delta) \cdot N_{YV}/N = P_R, \tag{3.14}$$

where it is assumed that the N_{YV} are known at the design stage.

Simple Random Sampling

Under simple random sampling, phase-II selection probabilities are the same for all individuals and

$$\pi(Y_i, V_i; \delta^{\text{srs}}) = P_R.$$

This design is easy to implement and renders covariate data *missing completely at random* (MCAR) (Little and Rubin, 2002). Analysis of the resulting data is therefore straight-forward as MCAR mechanisms are ignorable for analyses based on both fully-parametric likelihoods and semi-parametric estimating equations. This naive sampling scheme does not exploit information available in the phase-I data and so it will be used as a baseline to assess the efficiency gains of more sophisticated designs.

Balanced Sampling

Breslow and Cain (1988) and Breslow and Chatterjee (1999) advocate a balanced sampling design which samples equally from each of the strata defined by the phase-I data. That is, the phase-II sampling probabilities are inversely proportional to the size of the stratum, so that each stratum is equally represented in the phase-II sample. The stratum-specific sampling probabilities for this design are

$$\pi(y,v;\delta^{\text{bal}}) = \frac{P_R/\sum_{Y,V} 1}{N_{y,v}/N}, \tag{3.15}$$

where $\sum_{Y,V} 1$ is the number of strata defined at phase I.

While this design is not necessarily efficient, it is thought to offer a "reasonable compromise between the competing demands of efficiency and the need to check model assumptions" (Breslow and Chatterjee, 1999). Breslow and Chatterjee (1999) also point out that optimal designs will differ depending on the method of estimation that is used and advocate the balanced design as a more generally applicable approach.

Optimal Sampling Under Maximum Likelihood

The asymptotic variance of the ML estimator, obtained from $\mathbf{A}(\Omega)^{-1}$ (see Section 3.3.2), is a function of the choice of phase-II selection probabilities, so optimal designs under likelihood analyses can be obtained for any specified set of parameters Ψ. Specifically, phase-II selection probabilities can be found that minimise the asymptotic variance of α_x, subject to the budgetary constraints in (3.14); this sampling design, $\pi(Y_i,V_i;\delta^{\text{ml}})$, can be found using numerical minimisation procedures and will be optimally efficient for ML estimation of α_x whenever the covariate model and the parameter values are correctly specified at the design stage and N is large.

Optimal Sampling Using Inverse Probability Weighted Estimating Equations

Optimal phase-II sampling in this context requires finding the selection probabilities $\pi(Y,V;\delta)$ that result in the smallest asymptotic variance of our estimator. Primary interest lies in estimation of α_x, so we focus on minimising the $[2,2]$ entry of the asymptotic variance matrix, where $\{\mathbb{A}\}_{[k,k]}$ denotes the $[k,k]$ entry of matrix $\mathbb{A}$. This minimisation is subject to the budgetary constraint (3.14) so, similarly to Reilly and Pepe (1995), we wish to minimise

$$\Lambda = \left\{ \bar{\mathbf{A}}(\Psi)^{-1} \Big[\sum_{Y,V} \frac{P(Y,V;\Psi)}{\pi(Y,V;\delta)} \cdot E_{X|Y,V}\big[U_i(\alpha)U_i'(\alpha)\big] \Big] \bar{\mathbf{A}}(\Psi)^{-1} \right\}_{[2,2]} + \lambda \Big(\sum_{Y,V} \pi(Y,V;\delta) \frac{N_{YV}}{N} - P_R \Big).$$

The optimal $\pi(Y,V;\delta)$ is the root of

$$\frac{\partial\Lambda}{\partial\pi(Y,V;\delta)} = \left\{\bar{\mathbf{A}}(\Psi)^{-1}\big[-\frac{P(Y,V;\Psi)}{\pi(Y,V;\delta)^2}\cdot E_{X|Y,V}\big[U_i(\alpha)U_i'(\alpha)\big]\big]\bar{\mathbf{A}}(\Psi)^{-1}\right\}_{[2,2]} + \lambda\frac{N_{YV}}{N} = 0.$$

At the true Ψ, the optimal stratum-specific sampling probabilities for IPW estimation of α_x, $\pi(y,v;\delta^{\text{ipw}})$, can be written explicitly as

$$\frac{P_R\big[P(y,v;\Psi)/(N_{yv}/N)\big]^{1/2}\left\{\bar{\mathbf{A}}(\Psi)^{-1}E_{X|y,v}\big[U_i(\alpha)U_i'(\alpha)\big]\bar{\mathbf{A}}(\Psi)^{-1}\right\}_{[2,2]}^{1/2}}{\sum_{Y,V}\big[P(Y,V;\Psi)N_{YV}/N\big]^{1/2}\left\{\bar{\mathbf{A}}(\Psi)^{-1}E_{X|Y,V}\big[U_i(\alpha)U_i'(\alpha)\big]\bar{\mathbf{A}}(\Psi)^{-1}\right\}_{[2,2]}^{1/2}}. \tag{3.16}$$

Inverse Probability Weighted Estimating Equations With Estimated Weights

Following a similar argument, the optimal phase-II sampling probabilities $\pi(y,v;\delta^{\text{ipw}_2})$ for efficient estimation of α_x using IPW_2 can be shown to be

$$\frac{P_R\big[P(y,v;\Psi)/(N_{yv}/N)\big]^{1/2}\left\{\bar{\mathbf{A}}(\Psi)^{-1}\text{var}_{X|y,v}[U_i(\alpha)]\bar{\mathbf{A}}(\Psi)^{-1}\right\}_{[2,2]}^{1/2}}{\sum_{Y,V}\big[P(Y,V;\Psi)\cdot N_{YV}/N\big]^{1/2}\left\{\bar{\mathbf{A}}(\Psi)^{-1}\text{var}_{X|Y,V}[U_i(\alpha)]\bar{\mathbf{A}}(\Psi)^{-1}\right\}_{[2,2]}^{1/2}}. \tag{3.17}$$

Augmented Inverse Probability Weighted Estimating Equations

As with maximum likelihood, if Ψ could be known, the optimal choice of the selection probabilities for AIPW analysis could be found numerically by minimising (3.12) subject to the budgetary constraint in (3.14). However with discrete phase-I data, this efficient augmented AIPW estimator is asymptotically equivalent to the IPW_2 estimator (Robins et al., 1994) and therefore the optimal design for AIPW analysis can also be found explicitly through (3.17).

Note that in the consideration of optimal designs in the presence of discrete data that follows, tedious but straightforward algebra can show that IPW_2 and AIPW estimators of α_x are also asymptotically equivalent to the IPW estimator of α_x. Therefore, the optimal design for the IPW_2, AIPW, and IPW approaches in this context are all the same. However, the IPW estimator of α_0 and α_v are not asymptotically equivalent to the IPW_2 and AIPW estimators (the IPW_2 and AIPW approaches are asymptotically more efficient), so if optimality were not defined solely in terms of minimising the variance of the estimator of α_x, then the optimal design for IPW would

differ from the optimal design for IPW_2 or AIPW analysis. This phenomenon appears to be true for all settings involving discrete phase-I data considered in this thesis. Therefore, in what follows we have only two unique optimal designs: one for ML analysis, and one for the IPW, IPW_2, and AIPW analyses. Furthermore, this optimal design for ML would also be optimal for the AIPW estimator using U^{eff} in the consideration of binary data that follows since these estimators can be shown to have the same asymptotic variance in this context.

Optimal designs for the mean score method (which is equivalent to our IPW_2 method) are discussed by Reilly and Pepe (1995) and Whittemore and Halpern (1997), among others. However, optimal designs discussed here differ in that our budgetary constraint (3.14) is based on the observed phase-II stratum sizes, whereas Reilly and Pepe (1995) and Whittemore and Halpern (1997) base their constraint on the expected stratum sizes, as in

$$P(R=1;\delta)=\sum_{Y,V}\pi(Y,V;\delta)\cdot P(Y,V;\Psi)=P_R. \tag{3.18}$$

We have focussed on the budgetary constraint in (3.14) which uses observed stratum sizes for two reasons: i) it has practical appeal when phase-I data are known, and ii) this facilitates an exploration of the sensitivity of designs to changes in the parameters specified at the design stage. To illustrate this second point, note that our budgetary constraint in (3.14) does not depend on Ψ, but (3.18) is a function of the parameters used at the design stage. Therefore, changes to design-stage parameter values of Ψ would affect not only the optimal design for a given budgetary constraint, but also the constraint itself. In fact, the optimal designs derived by Whittemore and Halpern (1997) (their Table IV) would not sample the expected number of individuals from their observed phase-I data (their Table I) due to differences between the expected and observed phase-I stratum sizes. In the next section, we consider the loss of efficiency that will result when optimal designs are derived using incorrect parameter estimates. As discussed above, use of the budgetary constraint in (3.18) would yield incompatible results since these changes to the design parameters would also lead to changes in the budgetary constraint.

3.4.2 Empirical Properties of Phase-Two Designs

Here we develop results in the context of the two-phase study of biomarkers for joint damage in psoriatic arthritis and consider optimal two-phase designs for selecting patients for measurement of the biomarker MMP-3 in the University of Toronto Psoriatic Arthritis Clinic. The data consist of a binary response Y which indicates disease progression over the follow-up period, an inexpensive binary covariate V which indicates an elevated level of ESR at baseline, and an expensive covariate X related to baseline MMP-3 levels. Phase-I data (Y and V) have been gathered for 504 patients who had a clinic visit between 2003 and 2008 during which information was recorded on the patient's ESR level and on the extent of damage to the patient's joints. These patients also provided blood samples during these visits which enables assessment of the level of the

biomarker MMP-3. Followup visits that occurred roughly two years later were used to determine whether individuals had experienced an increase in the number of damaged joints constituting disease progression. The phase-I stratum sizes were $(N_{00}, N_{01}, N_{10}, N_{11}) = (196, 143, 61, 104)$, where N_{yv} is the number of individuals in the phase-I sample with $(Y = y, V = v)$.

The initial parameter estimates used in the design stage were obtained from two pilot studies which gathered complete data, including MMP-3 level, on 53 additional psoriatic arthritis patients. Of these 53 individuals, 9 individuals had elevated ESR and experienced an increase in their number of damaged joints during the follow-up period, and 17 had elevated ESR but did not experience disease progression. Of the 27 individuals without elevated ESR, 7 experienced disease progression during the follow-up period.

Optimal designs have been selected based on the limiting distributions given in Section 3.4, so it is important to explore how closely these asymptotic results relate to empirical standard errors in finite samples. Simulation studies were conducted with small samples ($N = 500$) and large samples ($N = 3000$). In both cases, optimal designs were derived for each method of analysis with budgetary constraints reflected through the specification of $P(R=1)=0.25$.

We consider empirical performance for the case of a binary (Section 3.4.2) and continuous (Section 3.4.2) covariate X; these two scenarios respectively reflect interest in the effect of dichotomised and raw values of the MMP-3 marker.

Empirical Properties of Phase-Two Designs with Binary X

For the simulation study, 1000 complete datasets of size N were generated according to response model (3.1) and covariate distributions

$$P(X = 1|V;\beta) = \text{expit}(\beta_0 + \beta_v V),$$

and

$$P(V = 1;\gamma) = \text{expit}(\gamma_0). \tag{3.19}$$

The parameter values $\Psi_0 = (\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_v, \gamma_0)' = (-1.95, 1.00, 0.90, 1.05, -0.41, -0.04)'$ were obtained from analysis of the PsA pilot data where MMP-3 levels were dichotomised based on whether they exceeded two standard deviations above the mean of controls (as specified by researchers at the PsA clinic).

Phase-I data were considered in each generated dataset, and phase-II samples were selected according to each of the four unique designs: simple random sampling (SRS), balanced sampling (Bal), the optimal design for ML analysis (Opt_{ml}), and the optimal design for IPW, IPW_2, and AIPW analyses (Opt_{ipw}). Each of the 1000 simulated incomplete datasets were analysed via

ML, IPW, IPW_2, and AIPW analyses. The empirical biases (Bias) and empirical standard errors (ESE) are presented in Table 3.1.

Note that in order to avoid undesirable, degenerate designs with near-zero selection probabilities in some strata (Breslow and Cain, 1988), we restricted the stratum-specific selection probabilities to be at least 0.05 (in our simulations, this restriction only affected the Opt_{ml} design). This restriction additionally ensured that samples could be analysed with IPW, where near-zero selection probabilities are especially problematic. Selection probabilities were also constrained to be less than or equal to 1. These boundary conditions can easily be incorporated into the numerical minimisation used for finding Opt_{ml}. We derived the balanced and the Opt_{ipw} designs by proceeding along the boundaries of the parameter constraints, as in Reilly and Pepe (1995). That is, when the solution to (3.15) (or to (3.16)) did not satisfy the boundaries conditions (i.e. $0.05 \leq \pi(y,v) \leq 1$ for some y and v), the offending stratum-specific selection probabilities were fixed at the corresponding boundary condition (i.e. $\pi(y,v)$ was fixed at 0.05 or 1), and the design was balanced (or optimised) over the remaining strata by updating the budgetary constraint in (3.14) to

$$\sum_{(Y,V)\neq(y,v)} \pi(Y,V;\delta) \cdot N_{YV}/N = P_R - \pi(y,v)N_{yv}/N.$$

The stratum-specific selection probabilities $\pi = [\pi(0,0), \pi(0,1), \pi(1,0), \pi(1,1)]$ employed by the SRS design were $\pi^{\text{srs}} = [0.25, 0.25, 0.25, 0.25]$. The balanced and optimal designs were based (at least in part) on phase-I data, so these designs depended on each simulated dataset. The average selection probabilities over the 1000 simulated datasets for the balanced, Opt_{ipw}, and Opt_{ml} designs were $\bar{\pi}^{\text{bal}} = [0.16, 0.22, 0.52, 0.32]$, $\bar{\pi}^{\text{ml}} = [0.05, 0.33, 0.16, 0.58]$, and $\bar{\pi}^{\text{ipw}} = [0.10, 0.25, 0.45, 0.43]$, respectively, so there was considerable variation between designs.

None of the methods of analysis showed appreciable empirical bias (see Table 3.1). The fully parametric ML analysis was the most efficient; it yielded estimates with standard errors at least as small as those arising from the estimating equation approaches under all designs. The asymptotic variances of α_x from the IPW, IPW_2, and AIPW are equal, however, the asymptotic variance of the estimators of α_0 and α_v are much smaller for the IPW_2 and AIPW approaches. These relationships were reflected in the empirical properties with $N = 3000$. However with $N = 500$, the IPW_2 approach resulted in much smaller empirical standard errors and slightly less empirical bias than AIPW; the IPW_2 estimator appears to have better small sample-sample properties than the asymptotically equivalent AIPW. We also explored the small-sample properties of an AIPW estimator where the approximation of ϕ_U^{opt} used in (3.13) was updated iteratively with the estimator of α, as suggested in Scott and Wild (2011a). This iterative AIPW estimator did have better small sample properties (not shown) than the AIPW estimator. The small-sample properties of analysis using this iterative AIPW estimator were identical to the small-sample results for the IPW_2 analysis; in fact, the estimates resulting from these methods of analysis were nearly identical for all simulations. A similar improvement in the small sample properties was achieved by using $\widehat{\alpha}^{\text{ipw}_2}$ in (3.13) instead of $\widehat{\alpha}^{\text{ipw}}$; in this case, the IPW_2 and AIPW estimates were gen-

Table 3.1: Empirical biases and standard errors resulting from analysing 1000 simulated datasets consisting of $N = 500$ and $N = 3000$ individuals with binary covariates and response, and employing four different phase-II sampling designs[†] with an expected phase-II sample size of $P(R = 1) = 0.25$. The data were generated using Ψ_0.

		Method of Analysis							
		ML		IPW		IPW_2		AIPW	
Design		Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE
					Phase-One Sample Size $N = 500$				
SRS	α_0	-0.075	0.469	-0.103	0.556	-0.080	0.478	-0.089	0.482
	α_x	0.068	0.519	0.069	0.520	0.073	0.527	0.084	0.533
	α_v	0.013	0.219	0.020	0.413	0.014	0.220	0.016	0.222
Bal	α_0	-0.038	0.418	-0.040	0.486	-0.040	0.435	-0.068	0.442
	α_x	0.037	0.473	0.036	0.485	0.036	0.490	0.047	0.496
	α_v	0.005	0.217	-0.006	0.390	0.007	0.221	0.023	0.232
Opt_{ml}	α_0	**-0.030**	**0.391**	-0.013	0.676	-0.023	0.486	-0.110	0.633
	α_x	**0.044**	**0.426**	0.029	0.556	0.029	0.577	0.045	0.608
	α_v	**-0.008**	**0.227**	-0.012	0.565	-0.010	0.241	0.062	0.414
Opt_{ipw}	α_0	-0.046	0.372	**-0.033**	**0.459**	**-0.041**	**0.376**	**-0.085**	**0.388**
	α_x	0.055	0.419	**0.049**	**0.423**	**0.049**	**0.424**	**0.059**	**0.428**
	α_v	0.005	0.220	**-0.002**	**0.390**	**0.004**	**0.223**	**0.033**	**0.240**
					Phase-One Sample Size $N = 3000$				
SRS	α_0	-0.008	0.173	-0.013	0.201	-0.011	0.173	-0.012	0.174
	α_x	0.007	0.194	0.010	0.194	0.010	0.194	0.012	0.194
	α_v	0.003	0.086	0.009	0.165	0.004	0.086	0.004	0.087
Bal	α_0	-0.008	0.157	-0.003	0.178	-0.006	0.160	-0.009	0.161
	α_x	0.008	0.178	0.004	0.179	0.004	0.179	0.006	0.180
	α_v	0.003	0.086	-0.001	0.146	0.003	0.087	0.005	0.087
Opt_{ml}	α_0	**0.005**	**0.153**	-0.006	0.245	-0.005	0.175	-0.015	0.179
	α_x	**0.010**	**0.169**	0.006	0.207	0.006	0.207	0.008	0.208
	α_v	**-0.012**	**0.092**	0.001	0.208	0.000	0.093	0.007	0.097
Opt_{ipw}	α_0	-0.008	0.155	**-0.001**	**0.187**	**-0.005**	**0.156**	**-0.011**	**0.156**
	α_x	0.009	0.178	**0.002**	**0.178**	**0.002**	**0.178**	**0.004**	**0.178**
	α_v	0.003	0.088	**0.001**	**0.155**	**0.004**	**0.090**	**0.007**	**0.090**

[†]The bolded entries indicate that sampling designs were optimal for estimation of α_x using the chosen method of analysis.

erally equivalent to 7 decimal places. Interestingly, the IPW estimators of α_x was actually more efficient here than either the IPW_2 or AIPW estimators.

As expected, the optimal designs (bolded in Table 3.1) resulted in the smallest empirical standard errors in the estimates of α_x for the respective method of analysis when N was large. However, the Opt_{ipw} design actually demonstrated the greatest efficiency for estimation of α_x when the sample size was small ($N = 500$). In all cases, the Opt_{ipw} design resulted in more efficient estimators of α_x than either SRS or the balanced design; this design also generally resulted in more efficient estimators for α_0, but the balanced design was often more efficient for estimation of α_v. Of course, if efficient estimation of α_v were also of primary importance, the definition of optimality could be modified and the optimal designs could be updated accordingly. The form of the optimal design in (3.17) can be easily extended for any specified linear function, h, of elements of the asymptotic variance matrix by replacing $\{\mathbb{A}\}_{[2,2]}$ with $h(\mathbb{A})$; in particular, analogs of A-optimality and C-optimality (Walter and Pronzato, 1990; Emery and Nenarokomov, 1999) can be achieved by taking $h = trace(H\mathbb{A}H')$ with $H = I$ or $H = diag\{\alpha\}^{-1}$, respectively.

It is interesting to note that the empirical efficiency of IPW_2 analysis under its optimal design was similar to ML analysis under its optimal design when the sample size was small. While the Opt_{ml} design resulted in inefficient estimating equation analyses, the Opt_{ipw} design was more efficient than either the SRS or balanced designs for estimation of the parameter of interest even when ML analysis was used.

Note that the asymptotic variance for the optimal class 1 restricted AIPW (see Section 3.3) are equivalent to the asymptotic variance for the ML approach here. This suggests that the use of the ϕ_U^{opt} term in (3.11) increases efficiency of the parameters not associated with the incompletely observed covariate (α_0 and α_v), while greatly increasing the efficiency of α_x requires replacing the optimal full-data estimator U_i with the corresponding optimal incomplete-data estimator, U^{eff}, as discussed in Robins et al. (1995) and Tsiatis (2006).

Empirical Properties of Phase-Two Designs with Continuous X

Here we consider the case of a continuous X, but we do not consider the Opt_{ml} design because the numerical minimisation of the asymptotic variance function has a high computational cost and because the ML method of analysis that is optimised by Opt_{ml} is less appealing as it requires correct specification of the conditional model for the continuous $X|V$; instead of analysing data using the ML approach which requires specification of the covariate distribution, here we employ the SPML and MI analyses along with the IPW, IPW_2, and AIPW analyses. The simple form of the Opt_{ipw} design makes it attractive as a basis for design in a wide range of situations (Whittemore and Halpern, 1997), so we contrast this efficient design with SRS and balanced designs for these methods of analysis. We next compare asymptotic and small-sample properties ($N = 500$) of designs in the presence of a continuous covariate X.

As before, optimal designs were derived so that $P(R=1)=0.25$. For the simulation study, 1000 complete datasets of size N were generated according to response model (3.1) and covariate distributions given by (3.19) and

$$g(X|V;\beta) = \frac{1}{\Gamma(\beta_0)(\beta_1+\beta_v V)^{\beta_0}} X^{\beta_0-1} e^{-\frac{X}{\beta_1+\beta_v V}}, \tag{3.20}$$

so that V remained binary, but now $X|V$ followed a gamma distribution with shape β_0 and scale $\beta_1+\beta_v V$. The parameters used in generating these data were $\Psi'_{c0} = (\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_1, \beta_v, \gamma_0)' =$ (-2.18, 0.03, .84, 1.40, 10, 5, -.04)$'$, where values of β were chosen to reflect the distribution of actual MMP-3 values given ESR status seen in the PsA pilot data.

Phase-II samples were chosen from each simulated set of phase-I data using three different designs described above: SRS, balanced, and the Opt_{ipw} design. The SRS design was identical to that used for a binary covariate X. The average balanced and Opt_{ipw} designs over the 1000 simulations were, respectively, $\bar{\pi}^{\text{bal}} = [0.15, 0.20, 0.81, 0.38]$ and $\bar{\pi}^{\text{ipw}} = [0.12, 0.30, 0.46, 0.41]$.

The simulated datasets were analysed using SPML, MI, IPW, IPW_2, and AIPW analyses as described in Section 3.3. The empirical biases and standard errors are presented in Table 3.2. Again the IPW_2 analysis had better small-sample properties than the asymptotically equivalent AIPW analysis, but the small-sample properties of AIPW could be improved by estimating ϕ_U^{opt} iteratively or by using $\widehat{\alpha}^{\text{ipw}_2}$ in (3.13). As before, the IPW method of analysis was much less efficient for estimation of α_0 and α_v, but slightly more efficient for estimation of α_x. The SPML estimator had the smallest empirical standard errors for estimation of all parameters under all designs. The IPW_2 analysis was generally more efficient than MI.

For all methods of analysis, the Opt_{ipw} design resulted in the most efficient estimator of α_x. As with the binary data case, this suggests that the Opt_{ipw} design is useful in a wide variety of situations, not only when the goal is finding an IPW estimate of α_x.

3.5 Misspecification in Optimal Two-Phase Designs

In the previous simulations, optimal designs were derived using the true parameter values, which in practice are unknown. In this section we explore the sensitivity of optimal phase-II sampling designs to misspecification of the parameter values at the design stage for a binary (Section 3.5.1) and a continuous (Section 3.5.2) covariate X. We further examine the effect of this misspecification on the asymptotic efficiency of estimators that use these designs.

Table 3.2: Empirical biases and standard errors resulting from analysing 1000 simulated datasets with a phase-I sample size of $N = 500$ when $X|V$ followed a gamma distribution. We consider maximum likelihood estimation through SPML and through MI with $m = 20$ imputations. Here $\Psi_{c0}' = (\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_1, \beta_v, \gamma_0)' = (-2.18, 0.031, .837, 1.40, 10.0, 5.00, -.040)'$

		Method of Analysis									
		SPML		MI		IPW		IPW_2		AIPW	
Design	Parameter	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE
SRS	α_0	-0.039	0.323	-0.051	0.333	-0.094	0.458	-0.040	0.326	-0.046	0.330
	α_x	0.001	0.015	0.002	0.016	0.001	0.015	0.001	0.015	0.002	0.016
	α_v	0.013	0.251	0.012	0.252	0.049	0.486	0.013	0.250	0.009	0.256
Bal	α_0	-0.031	0.273	-0.051	0.294	-0.018	0.350	-0.043	0.290	-0.077	0.303
	α_x	0.002	0.014	0.003	0.015	0.002	0.015	0.002	0.015	0.003	0.015
	α_v	0.007	0.250	0.006	0.257	-0.008	0.392	0.008	0.255	0.024	0.264
Opt_{ipw}	α_0	-0.026	0.277	-0.040	0.285	-0.016	0.377	-0.028	0.280	-0.062	0.320
	α_x	0.001	0.013	0.002	0.014	0.001	0.013	0.001	0.013	0.001	0.013
	α_v	0.011	0.247	0.007	0.254	-0.007	0.404	0.008	0.252	0.033	0.272

3.5.1 Sensitivity Analyses of Optimal Designs with Binary X

Misspecification of Design Parameters with Binary X

Table 3.3 contains the parameter estimates obtained by the analysis of the PsA pilot data which were used as the basis of the previous simulation studies; we now treat this set of parameter values as the true values and denote it by Ψ_0. Table 3.3 also contains parameter estimates that result from analysing the PsA pilot data under alternative "misspecified" models. Specifically, if we analyse the PsA data under the assumptions that $X \perp V$ we obtain the values in Ψ_1; assuming $Y \perp (X,V)$ gives the values in Ψ_2; assuming $Y \perp X|V$, $X \perp V$ yields Ψ_3; and assuming $Y \perp (X,V)$, $X \perp V$ gives Ψ_4. Compared to the 'true' parameter, Ψ_0, the other parameter sets in Table 3.3 represent varying degrees of misspecification: Ψ_1 represents the mistaken assumption that the covariates X and V are independent, in Ψ_2 it is mistakenly assumed that the covariates do not affect the response, in Ψ_3 it is assumed that X is independent of Y and V, and Ψ_4 assumes that all three variables are independent. These alternative parameter sets (Ψ_1, Ψ_2, Ψ_3, and Ψ_4) are used in the sensitivity analyses exploring the effect of parameter misspecification at the design stage; we examine the efficiency loss that would result in estimation of Ψ_0 if the other parameter sets were mistakenly used in the derivation of optimal designs. These sensitivity analyses are important since truly optimal designs can only be derived when the parameters of interest are known *a priori*.

The optimal designs derived using each of these different parameter sets are also in Table 3.3, where it is assumed that we are deriving designs for sampling from the 504 available PsA patients whose phase-I data resulted in observed stratum sizes of $(N_{00}, N_{01}, N_{10}, N_{11}) = (196, 143, 61, 104)$. Note that the SRS and balanced designs do not require initial parameter estimates and are therefore unaffected by misspecification of the parameters. Here again, Opt_{ipw} is the optimal design for efficiently estimating α_x through IPW, IPW_2, or AIPW analyses; Opt_{ml} is similarly defined to be optimal for ML analysis. In order to simplify the following discussion of asymptotic efficiencies of estimators of α_x, we will use "AIPW" to refer to the asymptotically equivalent IPW, IPW_2, and AIPW estimators.

Table 3.4 contains the asymptotic relative efficiencies for estimation of α_x under the supposedly-optimal designs which have been derived based on potentially misspecified parameter sets. The column labelled Ψ_0 in Table 3.4, for example, contains the asymptotic relative efficiency related to each of the designs when optimal designs are based on knowledge of the true parameter set Ψ_0, while the Ψ_1 column contains the asymptotic relative efficiencies when the parameters used in deriving optimal designs are mistakenly assumed to be Ψ_1. These asymptotic efficiencies are presented relative to the asymptotic efficiency achieved through SRS; the asymptotic relative efficiency (ARE) of the estimator $\widehat{\alpha}_x$ under the Opt_{ml} design, for example, would be calculated as

$$\frac{asvar_{\widehat{\alpha}_x}(\delta^{\text{ml}})}{asvar_{\widehat{\alpha}_x}(\delta^{\text{srs}})},$$

Table 3.3: Parameter estimates from the PsA data (Ψ_0) along with alternative parameter sets ($\Psi_1, \Psi_2, \Psi_3, \Psi_4$) that are derived from the PsA data using analyses that made incorrect assumptions about the data. Stratum-specific sampling probabilities for the optimal designs derived using different parameter sets are also presented where $P(R{=}1){=}0.25$.

Parameter	Ψ_0	Alternative Parameter Values			
		Ψ_1	Ψ_2	Ψ_3	Ψ_4
		$X \perp V$	$Y \perp (X,V)$	$Y \perp X \mid V, X \perp V$	$Y \perp (X,V), X \perp V$
α_0	-1.95	-1.95	-0.73	-1.16	-0.73
α_x	1.00	1.00	0.00†	0.00†	0.00†
α_v	0.90	0.90	0.00†	0.81	0.00†
β_0	1.05	0.84	1.05	0.84	0.84
β_v	-0.41	0.00†	-0.41	0.00†	0.00†
γ_0	-0.04	-0.04	-0.04	-0.04	-0.04
Selection probabilities		Optimal Design for Maximum Likelihood			
$\pi(0,0)$	0.05	0.05	0.21	0.12	0.32
$\pi(0,1)$	0.34	0.32	0.05	0.27	0.05
$\pi(1,0)$	0.15	0.19	0.99	0.41	0.81
$\pi(1,1)$	0.56	0.57	0.16	0.37	0.07
		Optimal Design for IPW, IPW$_2$, and AIPW			
$\pi(0,0)$	0.10	0.11	0.12	0.14	0.18
$\pi(0,1)$	0.25	0.24	0.15	0.25	0.20
$\pi(1,0)$	0.44	0.47	0.56	0.45	0.45
$\pi(1,1)$	0.42	0.40	0.37	0.34	0.34

The SRS and balanced designs, which do not require parameter values at the design stage, were respectively $\pi = (\pi(0,0), \pi(0,1), \pi(1,0), \pi(1,1)) = (0.25, 0.25, 0.25, 0.25)$, and $\pi = (0.16, 0.22, 0.52, 0.30)$.
†indicates values constrained to be zero

where $asvar_{\widehat{\alpha}_x}(\delta^{\text{ml}})$ is the asymptotic variance of $\widehat{\alpha}_x$ under the Opt_{ml} design; smaller AREs therefore correspond to more efficient designs. Here, the asymptotic variance of the ML and AIPW estimators under SRS are equivalent to 10 decimal places. Note again that these initial parameter misspecifications do not affect the SRS or balanced designs. The asymptotic relative efficiencies of the designs are also displayed graphically in Figure 3.1 for different values of the phase-II sampling fraction $P(R = 1)$.

An examination of Table 3.3 shows that the Opt_{ml} design was very sensitive to the design pa-

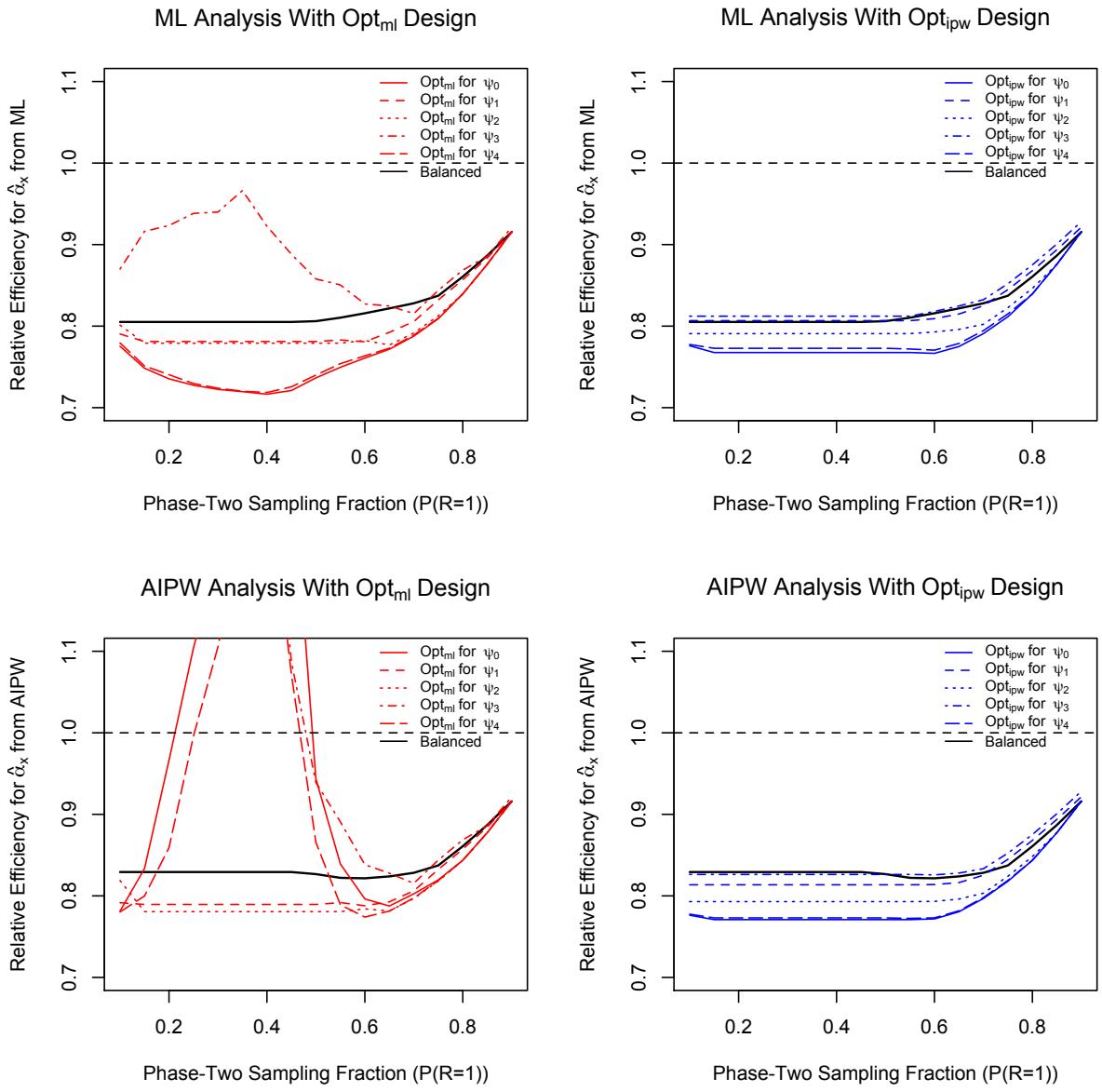

Figure 3.1: Asymptotic efficiency relative to SRS of estimators for α_x under different designs as a function of the phase-II sampling fraction

Table 3.4: Asymptotic relative efficiency of the ML, IPW, IPW_2, and AIPW estimators of α_x when $P(R{=}1){=}0.25$ and designs are based on the true parameters (Ψ_0) and on misspecified parameters (Ψ_1, Ψ_2, Ψ_3, and Ψ_4); efficiencies are relative to the asymptotic efficiency of estimators of α_x under the SRS design[†].

		Parameters used at the design stage				
Analysis	Design	Ψ_0	Ψ_1	Ψ_2	Ψ_3	Ψ_4
ML	Bal	0.805	0.805	0.805	0.805	0.805
	Opt_{ml}	0.728	0.730	0.781	0.779	0.938
	Opt_{ipw}	0.768	0.773	0.807	0.791	0.812
AIPW[‡]	Bal	0.829	0.829	0.829	0.829	0.829
	Opt_{ml}	1.100	0.996	0.789	0.781	2.550
	Opt_{ipw}	0.771	0.773	0.814	0.793	0.826

[†] ML and IPW/IPW_2/AIPW estimators of α_x under SRS were asymptotically equivalent to 10 decimal places.

[‡] Asymptotic efficiencies for IPW, IPW_2, and AIPW are equivalent here.

rameter settings. This is reflected in Table 3.4 where we see that the efficiency of the 'optimal' ML design differed greatly when different parameter values were used at the design stage. This indicates that efficient use of the Opt_{ml} design relies heavily on *a priori* knowledge of the true parameters. The Opt_{ipw} design, however, appears to be much more robust to changes in the parameter estimates used at the design stage. This design robustness is reflected in Table 3.4 where it can be seen that the Opt_{ipw} design was always more efficient than SRS for analyses based on both ML and estimating equation. The Opt_{ipw} design was always the most efficient choice for estimating equation analyses regardless of the parameter values used in the design. The Opt_{ipw} design was also generally more efficient than the balanced design for ML estimation of α_x; in the worst cases the Opt_{ipw} design is similar to the balanced design in terms of efficiency. The optimal ML design often contains selection probabilities which are small for certain strata (see Table 3.3). These designs result in very inefficient estimators of α_x based on weighted estimating equations. However, the Opt_{ipw} designs are very efficient for all considered analyses. Figure 3.1 shows that these trends are not specific to the situation where $P(R=1)=0.25$, but hold quite generally.

It is interesting to note that in using Ψ_3, it is assumed at the design stage that $X \perp (Y,V)$. In this case, the optimal designs are functionally independent of α and could potentially be derived using only phase-I data. However, the design parameters used in deriving optimal designs are often based on pilot data. In the next section, we consider the effect that this has on the efficiency of optimal designs.

Estimation of Design Parameters From External Pilot Studies (Binary X)

As mentioned above, the derivation of true optimal sampling designs requires *a priori* knowledge of parameter values which will generally be unknown. We have shown in the previous section that the Opt_{ml} design is sensitive to the misspecification of parameters at the design stage, but the Opt_{ipw} design appears to be more robust. In practice, when *a priori* knowledge of the necessary parameters is not available, it is possible to derive optimal designs by estimating parameter values from a small validation sample (Pepe et al., 1994; Reilly and Pepe, 1995; Reilly, 1996). However, if optimal designs are sensitive to the small changes in design parameters that will result from the use of pilot data, then these optimal designs will be of little practical use. Here we explore the sensitivity and efficiency of optimal designs when they are based on parameter estimates from simulated pilot studies of different sizes.

We considered external pilot studies of size m, where $m \in \{50, 200, 500, 1000\}$; note that $m = 1000$ represents an idealised pilot study which would most likely be prohibitively large in practice. For each pilot study, we simulated data (Y_i, X_i, V_i) for m individuals according to the parameter set Ψ_0 in Table 3.3 and we added one observation to each of the 8 strata defined by the binary (Y, X, V) in order to achieve greater stability in our estimates (Pepe et al., 1994). These simulated data were then used to find $\widehat{\Psi}$, a maximum likelihood estimate of the true parameters. Optimal designs were derived using $\widehat{\Psi}$ and the asymptotic variances that would result from employing these designs were recorded. This process was repeated 1000 times for each m. The results of these simulations are presented in Figures 3.2 and 3.3. It should be noted here that the Opt_{ml} design was previously seen to perform very poorly for the weighted estimating equation analyses and this situation should not be expected to improve when estimating design parameters using pilot data. However, the Opt_{ipw} was very efficient for the estimating equation analyses, and both the Opt_{ipw} and the Opt_{ml} designs performed well for ML analysis when the true parameters were used at the design stage. It is therefore important to examine how optimal designs will perform when pilot data are used for estimating design parameters, and it is interesting to compare the efficiencies of the Opt_{ipw} and Opt_{ml} designs for ML analysis when design parameters are estimated from pilot data.

Figure 3.2 shows that there is a great deal of variability in the phase-II selection probabilities employed by the Opt_{ml} design even when the pilot studies are large (m=1000), and the variability in the selection probabilities used by Opt_{ipw} is much smaller. In fact, even with small pilot studies ($m = 50$), the Opt_{ipw} design appears to only vary slightly between simulations.

The robustness of the Opt_{ipw} design is reflected in the relatively small changes in the asymptotic relative efficiency of the Opt_{ipw} seen in Figure 3.3. Here it can be seen that the Opt_{ml} design can often be inefficient, even under ML analysis, when based on relatively small pilot studies ($m = 50$). In fact, even with very large pilot studies ($m = 1000$), the Opt_{ml} was sometimes seen to be less efficient than the balanced design for ML analysis. The Opt_{ipw} design, on the other hand, was asymptotically more efficient than the balanced design for both ML and AIPW

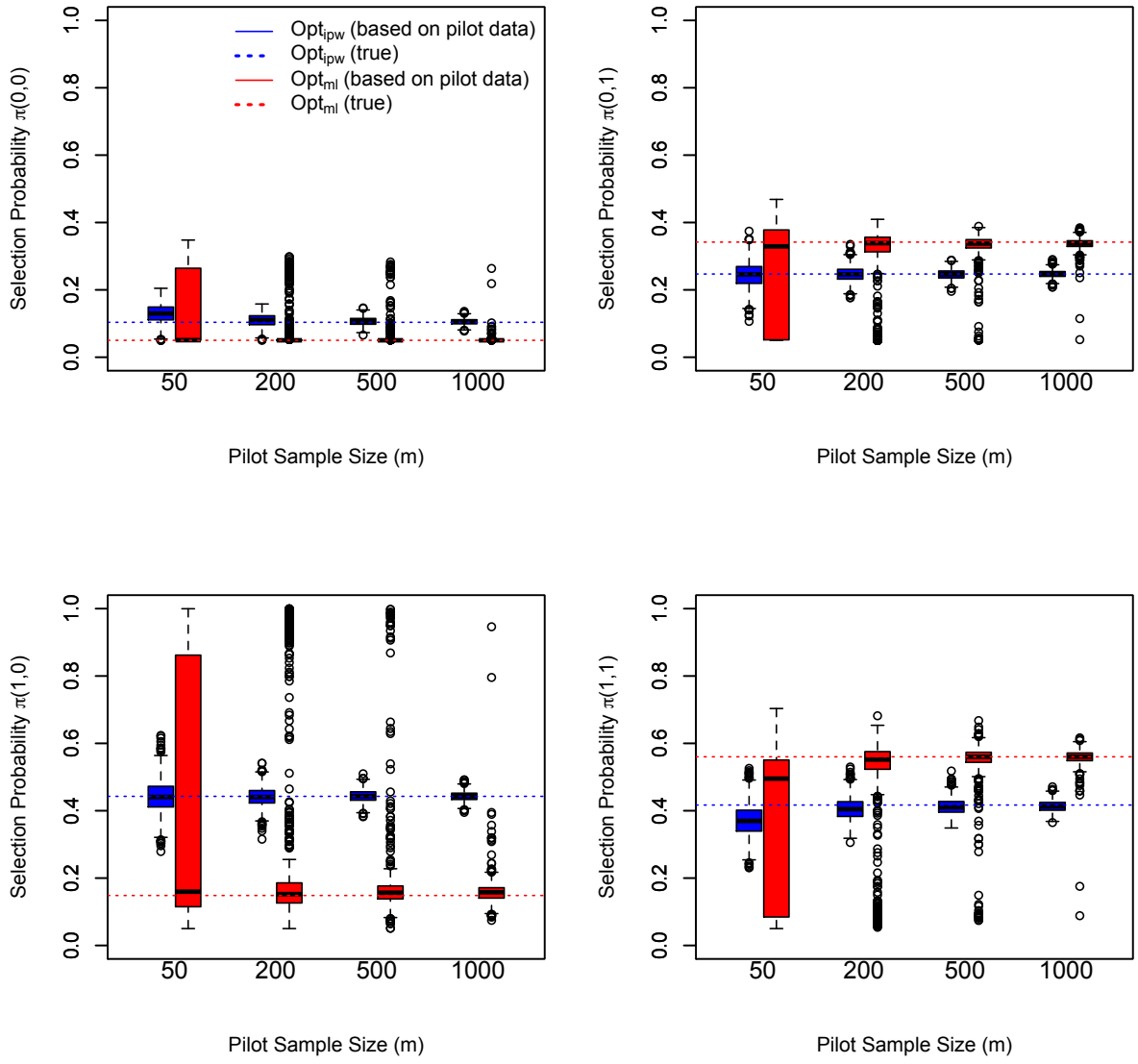

Figure 3.2: Boxplots of optimal selection probabilities derived using parameter estimates from pilot studies of size m with binary covariates and response. Here a wide range in the selection probabilities suggests that the optimal design is sensitive to the small changes in design parameter estimates that will result from the use of pilot study estimates instead of *a priori* knowledge of parameters; the dotted lines represent the optimal designs based on knowledge of the true parameters.

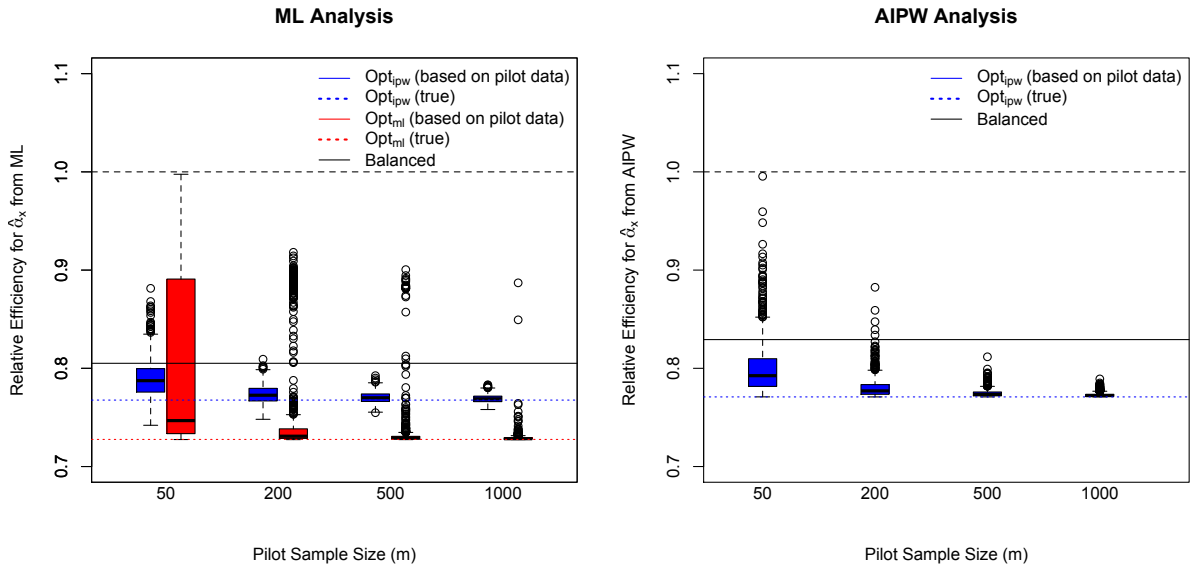

Figure 3.3: Asymptotic efficiencies of designs relative to SRS as a function of the size of the pilot studies used to estimate the design parameters with binary covariates and response. The balanced and SRS designs do not utilise the pilot data and are not affected by its size; the dotted lines here represent the relative efficiency of the optimal designs based on knowledge of the true parameters.

analyses in all but a handful of trials, even when the pilot study is moderately sized ($m = 200$). With the smallest pilot studies, the resulting Opt_{ipw} design was generally more efficient than the balanced design and was always more efficient than SRS. It should also be reiterated that, while the balanced design is shown to be quite efficient here, there is no guarantee that the balanced design will be efficient in general; in other design problems (see Sections 3.7 and 3.8) we have experienced situations where the balanced design is less efficient than the naïve SRS design. This Opt_{ipw} design, however, is guaranteed to be the most asymptotically efficient design for IPW, IPW_2, and AIPW analyses when the parameters are known *a priori*, and this design is shown here to be fairly efficient even when the design parameters are based on pilot data. In addition, the Opt_{ipw} design is also guaranteed to be fairly efficient for use with ML analysis since, asymptotically, the fully-parametric ML will be at least as efficient as estimating equation analyses for any given design.

The potential efficiency of the Opt_{ml} design is very appealing for ML analysis, but as seen in the previous section, this design is inefficient for other methods of analysis and is sensitive to small changes in the parameters used at the design stage so it may even be inefficient for ML analysis when parameters are not known *a priori*. The Opt_{ipw} design in contrast works well for ML, IPW, IPW_2, and AIPW analyses and is relatively robust to parameter misspecification.

3.5.2 Sensitivity Analyses of Optimal Designs with Continuous X

Misspecification of Covariate Model for Continuous X

When X is continuous, the derivation of the Opt_{ipw} design requires not only *a priori* knowledge of certain parameters, but also correct specification of the covariate distribution in (3.20). Here we wish to explore the efficiency of the Opt_{ipw} design if $X|V$ is modelled incorrectly at the design stage. Note that the incorrect specification of the model for $X|V$ will not effect the consistency of our estimators since this model is only used for specifying the $\pi(Y,V;\delta)$; none of the methods of analysis that we consider here – SPML, MI, IPW, IPW_2, and AIPW – require specification of a model for $X|V$.

Here we report on a simulation study to examine the effect on efficiency of misspecifying the nuisance covariate distribution. As in Section 3.4.2, the optimal design Opt_{ipw} was derived under the assumption that $X|V$ followed a gamma distribution. However, datasets of $N = 500$ individuals were simulated based on a log-normal model for $X|V$:

$$g(X|V;\beta) = \frac{1}{X\sqrt{2\pi\beta_1}} \exp\Big\{-\frac{(\log X - [\beta_0(1-V)+\beta_v V])^2}{2\beta_1}\Big\}.$$

The parameters used in generating the simulated data ($\Psi_{c1}' = (\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_1, \beta_v, \gamma_0)' =$ (-2.18, 0.03, .84, 0.73, 2.77, 2.37, -.04)$'$) were chosen so that the mean and variances of $X|V$ would be approximately equal to the mean and variances of the gamma distributed $X|V$ that was assumed in the derivation of Opt_{ipw} (in Section 3.4.2, we used $\Psi_{c0}' = (\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_1, \beta_v, \gamma_0)' =$ (-2.18, 0.03, .84, 1.40, 10, 5, -.04)$'$ to reflect the distribution of MMP-3 given ESR in the PsA pilot data). For the simulated data, phase-II samples were selected using SRS, balanced, and the (misspecified) Opt_{ipw} designs; these three potential two-phase samples were analysed using SPML, MI, IPW, IPW_2, and AIPW analyses. In order to curtail some particularly poor AIPW estimates, the AIPW estimator reported here was iterated to update the estimate of ϕ_U^{opt} a single time. The empirical biases and standard errors of the resultant estimators from 1000 simulated datasets are presented in Table 3.5. This simulation represents what would arise if the Opt_{ipw} design were derived for the PsA data under the mistaken assumption that the conditional distribution of the undichotomised MMP-3 given dichotomised ESR followed a gamma distribution, when in reality the distribution was log-normal.

Again, the IPW_2 analysis was at least as efficient as MI for estimation of all parameters under all designs (Table 3.5) and these methods of analysis were generally more efficient than IPW and less efficient than SPML in estimation of α_0 and α_v. The once-iterated AIPW estimator had very similar empirical properties as the IPW_2 estimator, but it was still slightly less efficient.

The Opt_{ipw} design was more efficient than the SRS and balanced designs for estimation of α_x for all methods of analysis, despite the fact that the Opt_{ipw} design was derived based on an incorrect

Table 3.5: Empirical biases and standard errors resulting from analysing 1000 simulated datasets with a phase-I sample size of $N = 500$ when $X|V$ followed a log-normal distribution, but Opt_{ipw} was derived using a misspecified covariate distribution.

		Method of Analysis									
		SPML		MI		IPW		IPW_2		AIPW[‡]	
Design	Parameter	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE
SRS	α_0	-0.032	0.346	-0.043	0.353	-0.028	0.458	-0.033	0.346	-0.031	0.346
	α_x	0.001	0.017	0.002	0.018	0.001	0.017	0.001	0.017	0.001	0.017
	α_v	0.005	0.259	0.005	0.264	-0.003	0.446	0.005	0.260	0.004	0.260
Bal	α_0	-0.022	0.280	-0.043	0.294	-0.031	0.348	-0.033	0.292	-0.030	0.292
	α_x	0.001	0.015	0.003	0.016	0.002	0.015	0.002	0.015	0.002	0.015
	α_v	-0.003	0.255	-0.007	0.261	0.020	0.388	-0.004	0.259	-0.005	0.260
Opt_{ipw}[†]	α_0	-0.028	0.283	-0.044	0.298	-0.046	0.393	-0.033	0.293	-0.028	0.294
	α_x	0.001	0.014	0.003	0.015	0.002	0.014	0.002	0.014	0.002	0.014
	α_v	0.001	0.258	-0.010	0.267	0.004	0.433	-0.008	0.264	-0.012	0.268

[†] The Opt_{ipw} design here was derived under the assumption that $X|V$ followed a gamma distribution, when in reality $X|V$ was log-normal.

[‡] The AIPW estimator was based on an iterative approach here

specification of the covariate distribution. The Opt_{ipw} design was also more efficient than the SRS design in estimation of α_0 for all methods of analysis, although the balanced design was generally most efficient for estimation of α_v.

Estimation of Design Parameters From External Pilot Studies (Continuous X)

Again it is important to explore the potential efficiency of the optimal design when the parameters specified at the design stage have been estimated from external pilot data. A simulation study was conducted in which data (Y,X,V) were simulated for each of $m \in \{50,200,500,1000\}$ individuals according to the parameters Ψ_{c0} with $X|V$ following a gamma distribution. As in the binary case, eight observations were added to the pilot data in order to stabilise the resulting estimates of the design parameters; since X is continuous, these eight additional observations consisted of the four combinations of Y and V along with a large and small value of X which were chosen here to be the means of $X|V=0$ and $X|V=1$. Parameter estimates found using these simulated pilot data were then used to derive the Opt_{ipw} design (using the correct covariate distribution) and the resultant selection probabilities and asymptotic efficiencies are displayed in Figures 3.4 and 3.5, respectively.

The Opt_{ipw} design is fairly robust even with a continuous covariate; the design changes little between simulations, even with fairly small simulated pilot studies (Figure 3.4). Despite the similarity between the asymptotic efficiencies under the balanced and the true Opt_{ipw} designs, the Opt_{ipw} designs based on pilot data still result in more efficient estimators in the vast majority of cases, even when the pilot studies are quite small (Figure 3.5). Therefore, the Opt_{ipw} is preferable to either the SRS or balanced designs here even though the expensive covariate is continuous and the design is based on parameter estimates from relatively small pilot studies.

3.6 Recommendations for Selection of PsA Patients

We return now to the problem of interest to the researchers at the Psoriatic Arthritis Clinic: selection of serum samples for measurement of the expensive covariate MMP-3 for use in the study of the effect of the biomarker MMP-3 on progression of PsA while accounting for the effect of the inexpensive and easy to measure ESR. We have shown that in this setting, efficiency can be gained by using an optimal IPW sampling design regardless of whether MMP-3 values are to be dichotomised and regardless of whether analysis is to be carried out using likelihood or weighted estimating equation approaches.

IPW, IPW_2, and AIPW were asymptotically equivalent in terms of their efficiency in estimating α_x and in terms of their optimal designs, but IPW_2 and AIPW were shown to be more efficient at estimating α_0 and α_v, and the IPW_2 method of analysis had consistently better small-sample

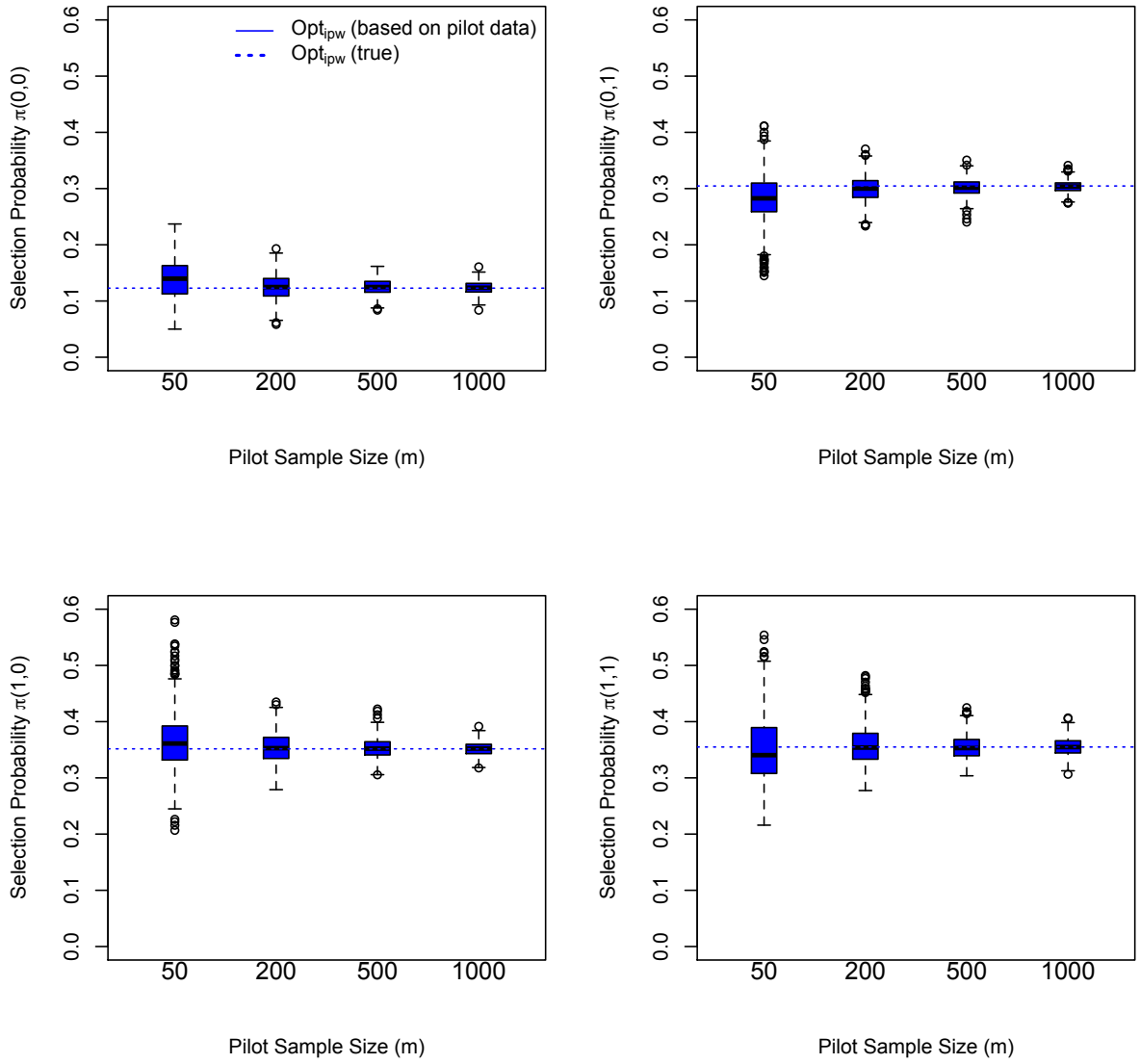

Figure 3.4: Boxplots of optimal selection probabilities derived using parameter estimates from pilot studies of size m when $X|V$ followed a gamma distribution.

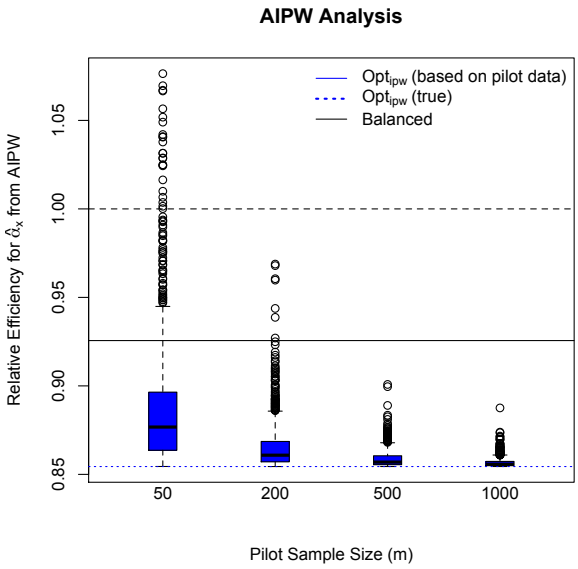

Figure 3.5: Boxplots of asymptotic efficiencies of the Opt_{ipw} designs relative to SRS as a function of the size of the pilot studies used to estimate the design parameters when $X|V$ followed a gamma distribution.

properties than the AIPW approach. The small-sample properties of our AIPW estimator were improved by iteratively updating the augmentation term, but even this iteration did not result in empirical standard errors that were smaller than those from the IPW_2 estimator. It appears that the IPW_2 estimator is preferable here, although that would change if one were to find the optimal incomplete-data estimating function U^{eff} and its corresponding optimal augmentation term $\phi^{\text{eff}}_{U^{\text{eff}}}$ in the AIPW estimator. Therefore, if a weighted estimating equation approach is to be used, we recommend using the AIPW approach only if one is willing to invest the time and resources in the derivation of U^{eff} and only if one is willing to bound stratum-specific selection probabilities away from both 0 and 1 (to avoid degenerative designs and the invertibility problems discussed earlier); otherwise the IPW_2 estimator should be used. If a likelihood-based analysis is to be implemented, we have seen that the SPML has much better small sample properties than MI. The SPML estimation procedure may be preferable in general as the IPW_2 analysis has been shown to be inefficient in some settings (Lawless et al., 1999).

Perhaps the most troubling aspect of implementation of optimal designs is the requirement for *a priori* knowledge of parameters. Reilly (1996) uses a small pilot study to estimate parameters for use in finding optimal designs. We have demonstrated that the Opt_{ml} design can be sensitive to misspecification of parameters, but the Opt_{ipw} design is efficient even when some parameters are incorrectly specified. In fact, efficient designs could be found in our scenario even in the absence of a pilot study by deriving the optimal Opt_{ipw} using only phase-I data and the assumption that

$X \perp (Y, V)$.

Based on the results of our sensitivity analyses it seems that the use of the Opt_{ml} design is only appropriate when one is very confident about the parameter estimates used at the design stage. However, the Opt_{ipw} design was shown to be efficient in a wide variety of circumstances: with a binary or continuous expensive covariate, with design parameters misspecified or estimated using pilot data, and even when derived using a misspecified covariate distribution. This efficiency of the Opt_{ipw} design was seen whether estimation was carried out with IPW, IPW_2, AIPW, ML, SPML, or MI analyses. Therefore when samples are to be selected based on discrete auxiliary information, it appears that the Opt_{ipw} design could effectively be used to efficiently select a phase-II sample regardless of the desired method of analysis.

Our recommendations for the Psoriatic Arthritis clinic are therefore that they select from the phase-I patients according to the Opt_{ipw} design based on parameter values estimated using the 53 patients in the pilot study. If budgetary constraints dictate that only 25% of the 504 available serum samples can be analysed for measurement of MMP-3, we recommend selecting according to the model

$$\pi^{\text{ipw}} = [0.10, 0.25, 0.44, 0.42]$$

if MMP-3 is to be dichotomised in the analysis, and according to the model

$$\pi^{\text{ipw}} = [0.12, 0.30, 0.35, 0.35]$$

if it is the effect of continuous values of MMP-3 that are of interest.

Our only reservations in this recommendation are two-fold. Firstly, this optimal design was based on an independent sampling procedure (called variable probability sampling in Lawless et al., 1999) which means that the phase-II sample size will be random and the budgetary constraint fixes only the expected sample size. Secondly, the pilot data are used solely to derive the design-parameter values and would be excluded from the final analysis; the money spent to get complete information on these 53 patients would represent a large portion of our available budget here and so this does not represent a good use of resources. We address both of these concerns in Chapter 4, where we consider basic stratified sampling and an adaptive design that does not require external pilot data.

In the next sections, we extend our investigation of two-phase designs to more complex settings motivated by the PsA and CLSA studies. We identify settings in which balanced designs are particularly inefficient.

3.7 Response-Dependent Sampling with Clustered Binary Data

3.7.1 The Response Model for Clustered Data

Let $Y_i = (Y_{i1}, Y_{i2})'$, denote the bivariate binary response for cluster i, and let X_i and V_i be the univariate expensive and auxiliary covariates, respectively, defined at the cluster level (i.e. all subjects in a given cluster have the same values of these covariates). In the context of the study from the University of Toronto Psoriatic Arthritis Clinic, the responses correspond to the status of the left and right sacroiliac joints. The expensive covariate could be a genetic marker and the auxiliary variable could be an inexpensive marker of inflammation measured at the time of clinic entry (e.g. baseline IL4 or ESR). For $\mu_{ij} = E[Y_{ij}|X_i, V_i] = P(Y_{ij} = 1|X_i, V_i)$, the conditional mean of Y_{ij} given covariates X_i and V_i, we consider the logistic response model

$$\text{logit } \mu_{ij} = \alpha_0 + \alpha_x X_i + \alpha_v V_i, \tag{3.21}$$

where the covariates are assumed to have a common effect on both responses. We adopt the marginal model of Lipsitz et al. (1991) to account for an association between Y_{i1} and Y_{i2} given (X_i, V_i). Let $\mu_{ikl} = P(Y_{i1} = k, Y_{i2} = l|X_i, V_i; \alpha)$, where $\alpha = (\alpha_0, \alpha_x, \alpha_v, \psi)'$. Then, specifically, we let

$$\psi_i = \frac{P(Y_{i1} = 1, Y_{i2} = 1|X_i, V_i)/P(Y_{i1} = 0, Y_{i2} = 1|X_i, V_i)}{P(Y_{i1} = 1, Y_{i2} = 0|X_i, V_i)/P(Y_{i1} = 0, Y_{i2} = 0|X_i, V_i)} = \frac{\mu_{i11}/\mu_{i01}}{\mu_{i10}/\mu_{i00}}$$

be the odds of subunit 1 in cluster i responding when subunit 2 responds, versus the respective odds when subunit 2 doesn't respond, given the cluster-level covariates X_i and V_i. We will further assume a common odds-ratio so that $\psi_i = \psi,\ i = 1, \ldots, N$. The joint probability is then

$$P(Y_{i1} = 1, Y_{i2} = 1|X_i, V_i; \alpha) = \begin{cases} \frac{c_i - [c_i^2 - 4\psi(\psi-1)\mu_{i1}\mu_{i2}]^{1/2}}{2(\psi-1)} & \text{if } \psi \neq 1 \\ \mu_{i1}\mu_{i2} & \text{if } \psi = 1 \end{cases}$$

where $c_i = 1 - (1 - \psi)(\mu_{i1} + \mu_{i2})$. The marginal means and the odds ratio completely specify the full distribution of the clustered binary response data with clusters of size 2. We consider binary covariates X and V which arise so that logit $P(X_i = 1|V_i; \beta) = \beta_0 + \beta_v V_i$ and logit $P(V_i = 1; \gamma) = \gamma_0$.

3.7.2 The Selection Model

We specify the phase-II sampling design for these bivariate data through the choice of selection parameters δ in the probabilities $\pi(Y_i, V_i; \delta) = P(R_i = 1|Y_i, V_i; \delta)$, where we consider the selection model

$$\text{logit } \pi(Y_i, V_i; \delta) = \delta_0 + \delta_1 Y_{i1} + \delta_2 Y_{i2} + \delta_3 V_i + \delta_4 Y_{i1} Y_{i2} + \delta_5 Y_{i1} V_i + \delta_6 Y_{i2} V_i + \delta_7 Y_{i1} Y_{i2} V_i.$$

Note that since the covariate V and the responses Y_1 and Y_2 are binary, the use of this saturated selection model is equivalent to specifying stratum-specific sampling probabilities which indicate the selection probabilities that should be used within each of the eight strata defined by the phase-I data (Y_1, Y_2, V). We will also concentrate here only on asymptotic calculations and therefore define the budgetary constraint in terms of expectations as in (3.18). A simple random sampling design in this context still involves phase-II selection probabilities which are equal for all strata (so $\delta_k = 0,\ k = 1, \ldots, 7$).

Optimal IPW$_2$ Sampling

When using the IPW$_2$ analysis from (3.10) where now

$$U_i(\alpha) \equiv \partial \log P(Y_{i1}, Y_{i2} | X_i, V_i; \alpha)/\partial \alpha$$

(i.e when maximising the weighted pseudo-likelihood $R_i \widehat{\pi}_i^{-1} U_i$), the optimal design for efficient estimation of α_x, Opt$_{\text{ipw}}$, can be derived in a similar manner to (3.17) so that now

$$\pi(y, v; \delta^{\text{ipw}}) = \frac{P_R \left\{ \bar{\mathbf{A}}(\Psi)^{-1} var_{X|y,v}\left[U_i(\alpha)\right] \bar{\mathbf{A}}(\Psi)^{-1} \right\}^{1/2}_{[2,2]}}{\sum_{Y,V} P(Y, V; \Psi) \left\{ \bar{\mathbf{A}}(\Psi)^{-1} var_{X|Y,V}\left[U_i(\alpha)\right] \bar{\mathbf{A}}(\Psi)^{-1} \right\}^{1/2}_{[2,2]}}, \tag{3.22}$$

where y represents the vector (y_1, y_2) and

$$\bar{\mathbf{A}}(\Psi) = E_{Y_1, Y_2, X, V}[-\partial U_i(\alpha)/\partial \alpha'].$$

Balanced Sampling

For the clustered data problem, we will consider two balanced sampling designs. In the first balanced sampling design, the phase-I sample will be divided into the eight classes defined by all possible values of (Y_1, Y_2, V). However, since we are defining efficiency in terms of the variance of the estimator of α_x, and (3.21) assumes a common effect of X on either response, it may be more in the spirit of the balanced design to sample equally from the six strata defined by $(Y_1 + Y_2, V)$; therefore we also consider this second balanced design when analysing the clustered data. It is clear here that implementation of a balanced design is not as straightforward as one might like as it is not always clear how phase-I data should be used to stratify the sample. Note that in our asymptotic calculations, these designs are based on expected phase-I stratum sizes, which come from having knowledge of the true parameters at the design stage.

3.7.3 Asymptotic Relative Efficiencies

As before, we considered analyses based on maximising the observed data likelihood where saturated models are used for modelling the nuisance covariate distributions (ML), and based on maximising the weighted pseudo-likelihood with weights estimated using a saturated model for missingness (IPW_2). For both of these methods of analysis we considered four designs: simple random sampling (SRS), balanced sampling over all eight strata defined by (Y_1, Y_2, V) (Bal 8), balanced sampling over the six strata defined by $(Y_1 + Y_2, V)$ (Bal 6), and the sampling design which is asymptotically optimal for precise estimation of α_x with IPW_2 analysis (Opt_{ipw}). The asymptotic efficiencies of the estimators of α_x resulting from these three latter designs were calculated relative to the asymptotic efficiency of a simple random sampling design where, as before, greater precision corresponds to lower asymptotic relative efficiency.

We also consider the relative efficiency of the designs for estimating α_0, α_v, and ψ; note that the 'optimal' design will not necessarily be efficient for estimation of parameters other than α_x. The asymptotic relative efficiencies of the different sampling designs is presented in Figure 3.6 for IPW_2 estimation and in Figure 3.7 for ML estimation. The relative efficiencies are presented for a range of values of the association parameter ψ while the other parameters were chosen so that $E[Y_1] = E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.5; E[R] = 0.25; \alpha_x = \log(1.25); \alpha_v = \log(1.5); \beta_v = \log(1.5)$.

It can be seen that the Opt_{ipw} design allows for a great increase in the efficiency of estimation of α_x regardless of the method of analysis (Figures 3.6 and 3.7). The Opt_{ipw} design also increased the efficiency with which α_0 was estimated. This is similar to that which was reported by Reilly (1996), where optimising for efficient estimation of one parameter led to efficiency gains for other parameters. The choice of design had little effect on the efficiency of estimators of α_v and ψ.

The Bal 6 design was consistently more efficient than SRS for estimation of α_0 and α_x, while the Bal 8 design sometimes resulted in efficiency gains and sometimes resulted in a loss of efficiency compared to SRS. This illustrates the main problem with implementing a balanced sampling design: it is difficult to determine how to best stratify the phase-I sample based on the available data for balanced sampling, and it is unclear if a balanced design will result in more efficient estimators than a SRS design would. The Opt_{ipw} design, on the other hand, is guaranteed to increase the efficiency of the IPW_2 estimators of parameters of interest and has been shown to often increase efficiency for other estimators and parameters not specified in the definition of optimality.

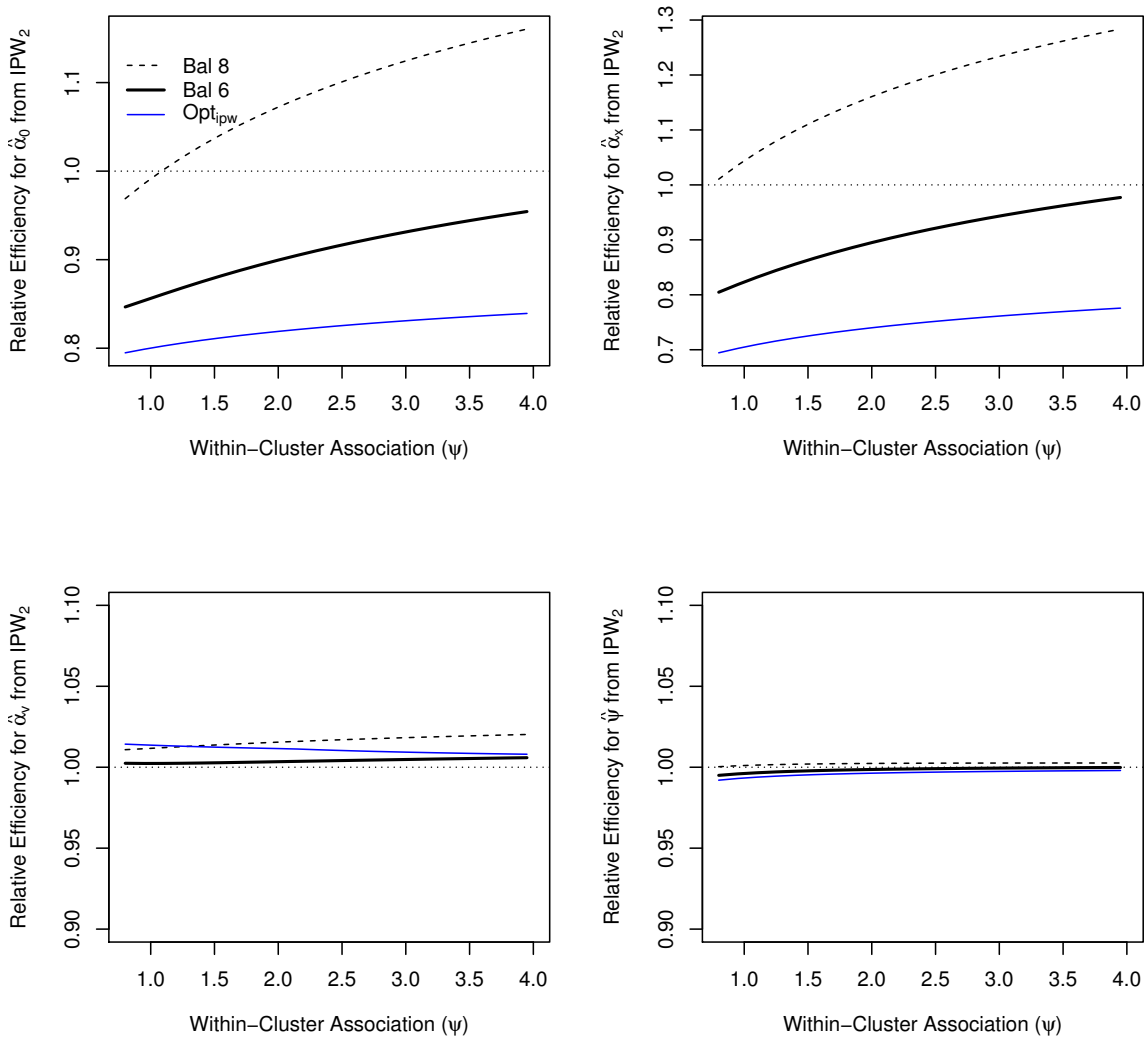

Figure 3.6: Asymptotic relative efficiencies (vs. SRS) of IPW$_2$ estimators of α_x using balanced (8 and 6 strata) and Opt$_{\text{ipw}}$ designs with clustered responses. $E[Y_1] = E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.5; E[R] = 0.25; \alpha_x = \log(1.25); \alpha_v = \log(1.5); \beta_v = \log(1.5)$

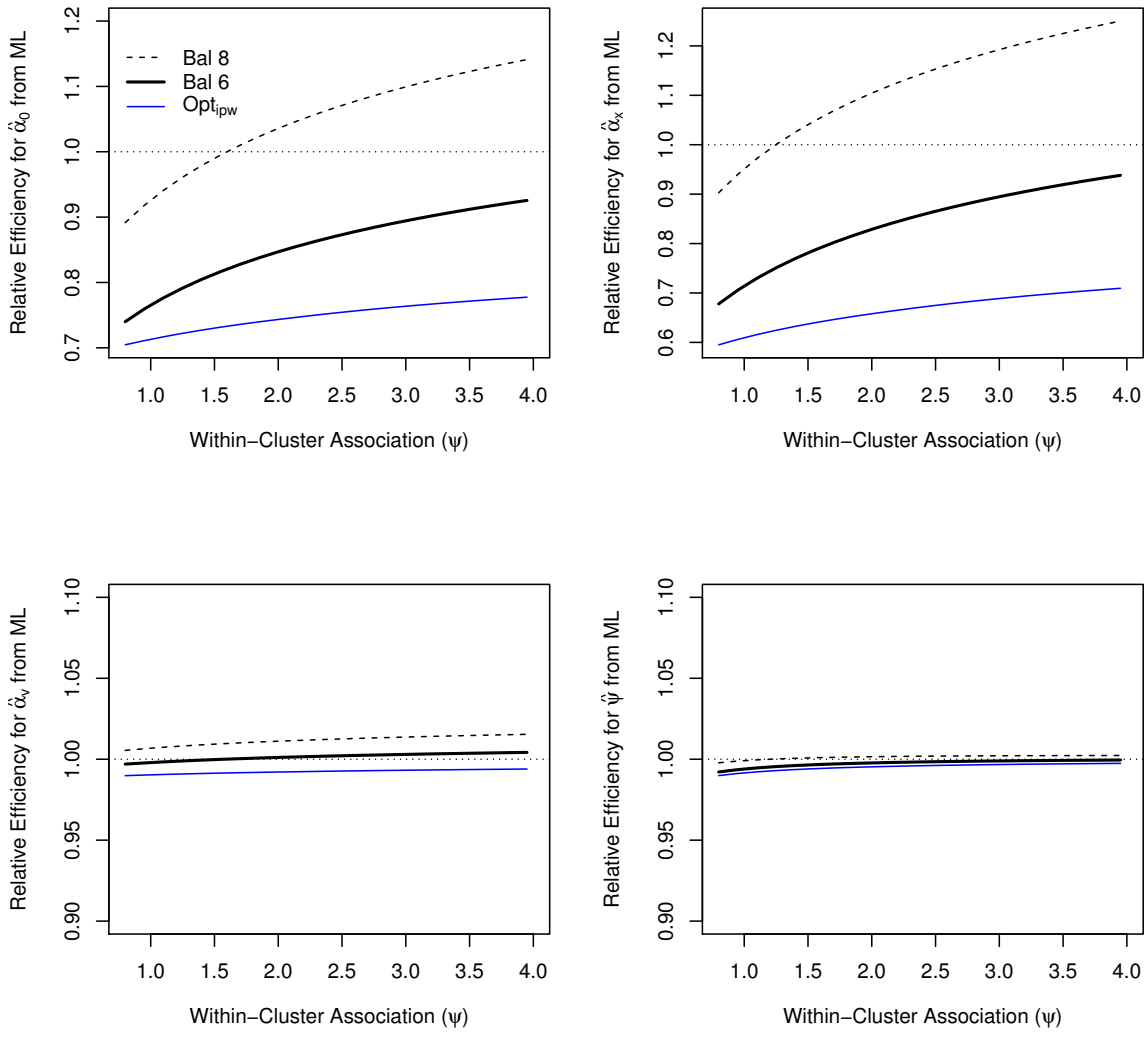

Figure 3.7: Asymptotic relative efficiencies (vs. SRS) of ML estimators of α_x using balanced (8 and 6 strata) and Opt$_{\text{ipw}}$ designs with clustered responses. $E[Y_1] = E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.5; E[R] = 0.25; \alpha_x = \log(1.25); \alpha_v = \log(1.5); \beta_v = \log(1.5)$

3.8 Response-Dependent Sampling with Longitudinal Binary Data

3.8.1 The Response Model for Longitudinal Data

Here we consider the analysis of binary data arising from a longitudinal study where the binary response variable is measured at up to K prespecified time points (one baseline assessment and up to $K-1$ follow-up assessments). We denote the response for individual i as $Y_i = (Y_{i0}, Y_{i1}, \ldots, Y_{i,K-1})'$. We again consider binary covariates X_i and V_i, where V_i is known for all individuals at time 0, but X_i will only be collected for individuals selected into a phase-II sample. We further assume that $Y_{i0} = 0,\ i = 1, \ldots, N$ and $P(Y_{ik} = 0 | Y_{i,k-1} = 1) = 0$, so that once an individual has $Y_{ik} = 1$, then $Y_{il} = 1$ with probability 1 for $l = k+1, \ldots, K$.

Thus if we consider Y_{it} to be an indicator of disease status at time point t, then these data represent a study of an irreversible disease process in which all individuals are disease-free at study entry. Specifically, it is of interest to examine how the change in disease status is affected by a time-invariant, expensive binary covariate X_i (possibly a genetic factor), after accounting for an available baseline auxiliary covariate V_i. This framework is consistent with the aims of the CLSA where the objective is to model risk factors for the onset of disease.

We again consider ML and IPW_2 analyses. For these data, we are not interested in estimating marginal parameters as in (3.21), rather we are primarily interested in the effect of the covariate X in the transitional response model

$$\text{logit}\, P(Y_{ik} = 1 | Y_{i,k-1} = 0, X_i, V_i; \alpha) = \alpha_{0k} + \alpha_x X_i + \alpha_v V_i, \qquad k = 1, 2, \ldots, K.$$

Due to the irreversible nature of the disease process, if $K = 2$ the joint response model on which the likelihood methods are based is

$$\begin{aligned} P(Y_i | X_i, V_i; \alpha) &= I(Y_{i1} = 1) I(Y_{i2} = 1) P(Y_{i1} = 1 | Y_{i0} = 0, X_i, V_i; \alpha) \\ &\quad + I(Y_{i1} = 0) I(Y_{i2} = 0) [1 - P(Y_{i1} = 1 | Y_{i0} = 0, X_i, V_i; \alpha)] [1 - P(Y_{i2} = 1 | Y_{i1} = 0, X_i, V_i; \alpha)] \\ &\quad + I(Y_{i1} = 0) I(Y_{i2} = 1) [1 - P(Y_{i1} = 1 | Y_{i0} = 0, X_i, V_i; \alpha)] P(Y_{i2} = 1 | Y_{i1} = 0, X_i, V_i; \alpha). \end{aligned}$$

3.8.2 The Selection Model

Here we set $K = 2$ and consider balanced and Opt_{ipw} designs for the selection of a phase-II sample at each of the three time points. This allows us to examine how the efficiency of designs is affected by the amount of auxiliary information available at phase-I for choosing the phase-II sample. Note that simple random sampling is not affected by the time at which the phase-II sample is chosen as this design does not exploit the data available at phase-I.

The selection model at time t can be expressed as $\pi_{it}(\overline{Y}_{it}, V_i; \delta^{(t)}) = P(R_i = 1 | Y_{i1}, \ldots, Y_{it}, V_i)$. At each progressive time point, more phase-I information is available for exploitation in deriving efficient phase-II selection probabilities. At time point 0, the phase-I sample can be divided into two strata based on the available information on V, so $\pi(\overline{Y}_0, V; \delta^{(0)}) = \pi(V; \delta^{(0)})$; at time point 1, the phase-I sample can be stratified into four classes based on the available information on V and Y_1, so $\pi(\overline{Y}_1, V; \delta^{(1)}) = \pi(Y_1, V; \delta^{(1)})$; at time point 2, the phase-I sample can be stratified into six classes based on the available information on V, Y_1, and Y_2, where $P(Y_2 = 0 | Y_1 = 1) = 0$.

Simple random sampling is the same at each time point, but the efficiency of the balanced and optimal designs will be affected by the amount of information available at phase-I. Therefore, for this study of transitional effects, we consider 7 designs for each method of analysis: simple random sampling (SRS), balanced sampling using the phase-I data available at each time point (call these Bal 0, Bal 1, and Bal 2 at time points 0, 1, and 2, respectively) and the sampling designs which are optimal for IPW_2 estimation of α_x given the data that are available at the time of selection (call these Opt_{ipw} 0, Opt_{ipw} 1, and Opt_{ipw} 2). We will again present the efficiencies of the designs relative to simple random sampling, which would be carried out in the same way at all time points as SRS does not exploit the available phase-I information.

The asymptotic variances and optimal designs can be found as before; however summations are no longer over all strata (defined here by (Y_1, Y_2, V)), but rather over strata defined by the data that are available at the time of selection. This decrease in phase-I data essentially places added constraints on the previously derived optimal sampling design; for example, at time point 0, when only V is available for phase-II sampling decisions, then $\pi(Y, V; \delta) = \pi(V; \delta)$ for all $Y = (Y_1, Y_2) \in \{(0,0), (0,1), (1,1)\}$. The Opt_{ipw} sampling design at time point 1, for example, is defined by stratum-specific sampling probabilities $\pi(y_1, v; \delta^{\text{ipw1}})$ that can be found as

$$\frac{P_R \left\{ \bar{\mathbf{A}}(\Psi)^{-1} \sum_{Y2} \left\{ P(y_1, Y_2, v; \Psi) / P(y_1, v; \Psi) E_{X|y_1, Y_2, v} \left[U_i(\alpha) U_i'(\alpha) \right] \right\} \bar{\mathbf{A}}(\Psi)^{-1} \right\}_{[2,2]}^{1/2}}{\sum_{Y_1, V} \left\{ \bar{\mathbf{A}}(\Psi)^{-1} \sum_{Y_2} \left\{ P(Y_1, Y_2, V; \Psi) P(Y_1, V; \Psi) E_{X|Y_1, Y_2, V} \left[U_i(\alpha) U_i'(\alpha) \right] \right\} \bar{\mathbf{A}}(\Psi)^{-1} \right\}_{[2,2]}^{1/2}},$$

where $U_i(\alpha) \equiv \partial \log P(Y_{i1}, Y_{i2} | X_i, V_i; \alpha) / \partial \alpha$ and $\bar{\mathbf{A}}(\Psi) = E_{Y_1, Y_2, X, V}[-\partial U_i(\alpha) / \partial \alpha']$.

3.8.3 Asymptotic Relative Efficiencies

We derived optimal designs for a range of values of P_R, which defines the budgetary constraint as in (3.18). Other parameters were chosen so that $E[Y_1] = 0.2; E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.75; \alpha_x = \log(1.5); \alpha_v = \log(1.5); \beta_v = \log(1.5)$. The relative efficiencies of the different sampling designs is presented in Figures 3.8 and 3.9 for IPW_2 and ML analyses, respectively. We

consider the relative efficiency of each of the considered designs for estimating α_0, α_1, α_x, and α_v.

As expected, the Opt_{ipw} sampling designs offered large efficiency gains over simple random and balanced designs when estimating α_x. As before, these designs also added efficiency to the estimation of α_0 (Figures 3.8 and 3.9). Having more information at the time of sampling increased the efficiency of the optimal design for the estimation of both α_0 and α_x. However Bal 2, the balanced design at time point 2, was often less efficient than Bal 1, the balanced design which was based only on the auxiliary information available at time point 1. This indicates that, as was seen in the comparison of Bal 6 and Bal 8 in the previous section, having more phase-I information does not necessarily improve the efficiency of balanced designs. For IPW_2 analysis, none of the balanced designs was consistently better than the others and all of the balanced designs were generally less efficient than SRS. For ML analysis, the Bal 1 design was the best balanced design and was preferable to SRS. The Bal 0 design, on the other hand, was generally less efficient than SRS for ML analysis and Bal 2 sometimes resulted in greater asymptotic efficiency than SRS and sometimes did not.

The asymptotic variances of the ML and IPW_2 estimators under SRS were very similar. The efficiency gain of the Opt_{ipw} designs over SRS for estimation of α_x was also very similar for both methods of analysis; the Opt_{ipw} 0 design was very similar to SRS, and the Opt_{ipw} 1 and Opt_{ipw} 2 designs resulted in increasing levels of asymptotic efficiency for estimation of α_x and α_0. The balanced designs were most efficient for estimation of α_1 and α_v, but the differences between designs was small. Furthermore, if precision of estimation of α_v was also of primary importance, the optimality criteria could easily be modified to find designs that were also efficient for estimation of α_v.

Note that as the sampling fraction increases, smaller strata are selected in their entirety by the balanced designs (the selection probabilities must be capped at 1, as discussed previously); this accounts of the lack of smoothness in the change in asymptotic efficiency of the balanced designs.

3.9 Discussion

One criticism of optimal sampling approaches offered by Breslow and Chatterjee (1999) is that optimal designs differ for different methods of analysis. Furthermore, Breslow and Cain (1988) encountered "degenerate" optimal designs which were undesirable due to the fact that selection probabilities for certain strata were near zero. We have shown that the optimal IPW_2 design is also optimal for AIPW analysis and in our situation was also optimal for IPW estimation of α_x. Additionally, the Opt_{ipw} design offered improved efficiency for ML, MI, and SPML analyses in a wide variety of settings. Furthermore, while the Opt_{ml} designs did contain very small sampling probabilities, this did not occur with the Opt_{ipw} design. In fact, a necessary condition for IPW

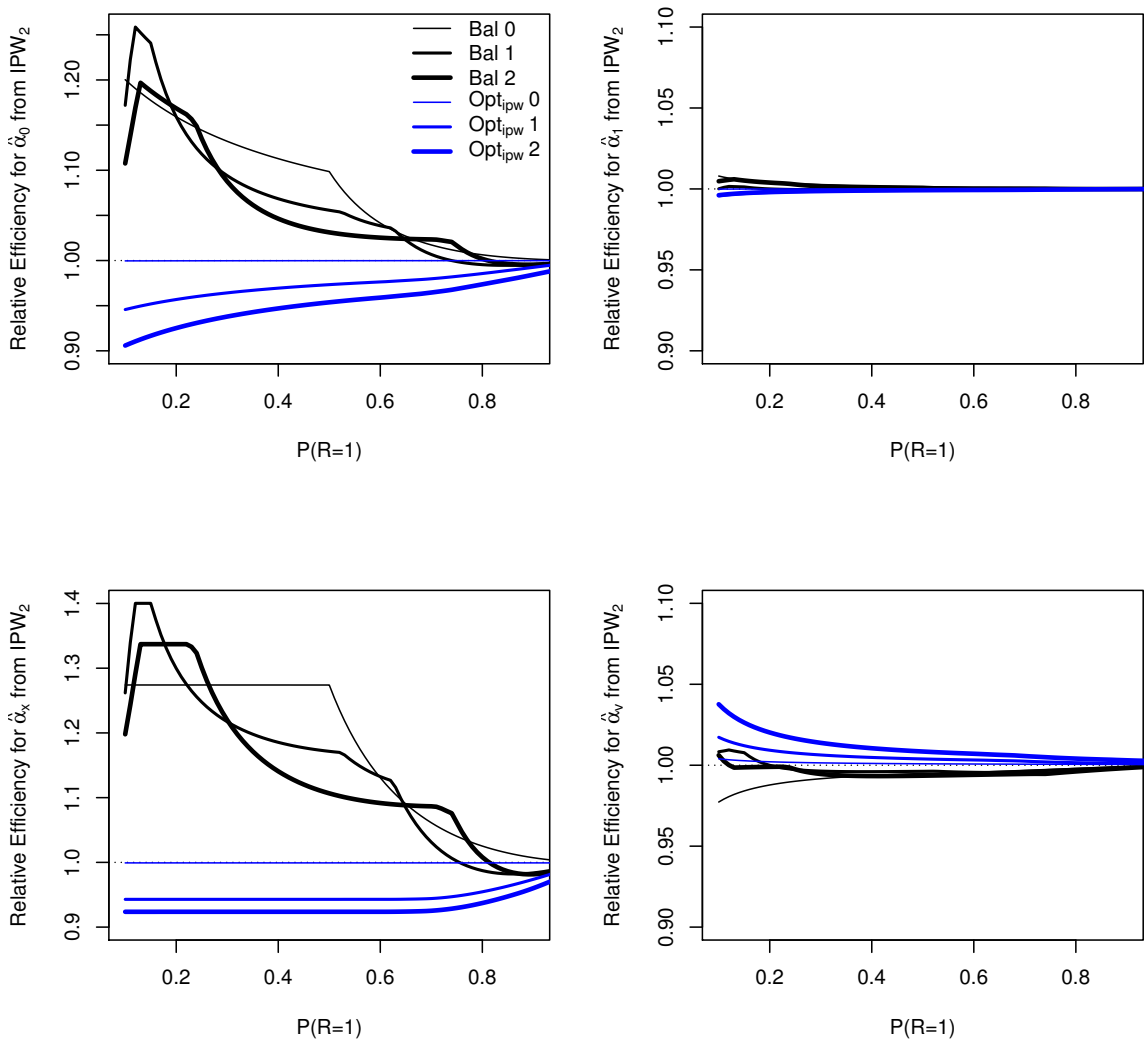

Figure 3.8: Asymptotic efficiency of estimators under balanced and optimal IPW designs relative to simple random sampling when using IPW$_2$ analysis to estimate transitional effects. $E[Y_1] = 0.2; E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.75; \alpha_x = \log(1.5); \alpha_v = \log(1.5); \beta_v = \log(1.5)$

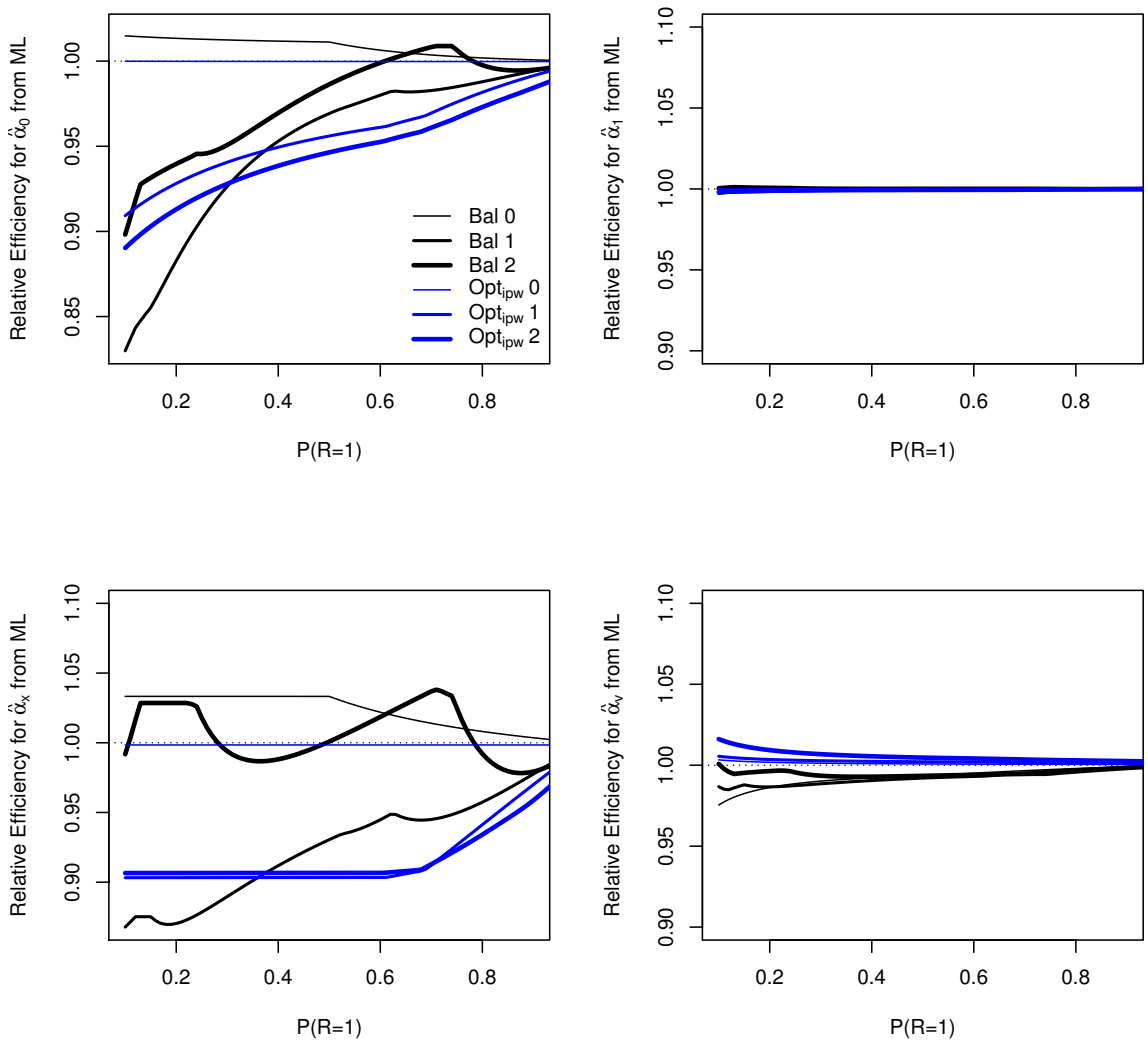

Figure 3.9: Asymptotic efficiency of estimators under balanced and optimal IPW designs relative to simple random sampling when using maximum likelihood analysis to estimate transitional effects. $E[Y_1] = 0.2; E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.75; \alpha_x = \log(1.5); \alpha_v = \log(1.5); \beta_v = \log(1.5)$

analysis is that the selection probabilities be bounded away from 0, so optimising for IPW analysis should avoid this problem entirely. In any case, this problem of degenerate designs could easily be avoided by searching for optimal designs amongst those designs that have a minimum selection probability (for example, we required that $\pi(Y,V;\delta) \geq 5\%$). Since this optimisation is over a class of designs that includes the balanced design, this approach would still be at least as efficient as using a balanced design and will avoid degenerate designs. Importantly, all methods of analysis demonstrated improved efficiency in estimation of the parameter of interest when the Opt_{ipw} design was used. This held true even when the design was derived based on pilot data, or derived using slightly misspecified parameters or covariate distributions. This design also often resulted in improved efficiency of estimators of other parameters and the optimality criteria could easily be modified to include any set of parameters whose estimation is of primary importance.

The investigations in this chapter are among the first to study the relative efficiencies of two-phase sampling designs involving clustered or longitudinal data. Given the increased interest in studies involving cross-sectionally clustered data and the recent trend towards the design of massive cohort studies of health and disease, the insights that result from this work are important.

For the setting of clustered data, the first decision to make is typically on the method of analysis and there are a variety of frameworks one can adopt. We restricted attention to bivariate response data and marginal models for characterising the effects of exposure. In this setting, maximum likelihood and the IPW_2 methods can be much more efficient than IPW for estimation of the covariate effects; maximum likelihood is asymptotically the most efficient method, but necessitates modelling the covariate distribution. When the exposure variable is continuous, a robust implementation of the inverse probability weighting methods may be more appealing since no modelling of exposure is required. We found that optimal designs based on maximum likelihood analyses may be more sensitive to small changes in the parameters used at the design stage than optimal IPW designs. So, if models for exposure variable are difficult to formulate with confidence, the robustness of the IPW approaches may be more appealing. When the auxiliary variable is continuous, discretising seems the most practical approach to addressing the curse of dimensionality and this has been recommended by several authors (Lawless et al., 1999).

When comparing the effect of different frameworks for analysis and design, it is interesting to note that the conclusions about optimality bear only on the criteria adopted for the optimal design; the intercept, effect of the auxiliary variable and association parameters do not necessarily behave in the same way.

The pragmatic approach of using balanced sampling designs as a compromise between robustness and efficiency does not yield clear and consistent recommendations; the resulting estimators sometimes perform well and sometimes perform poorly. Furthermore, it is unclear what auxiliary information should be considered when implementing a balanced design especially in more complex settings.

There are several directions of future research that are natural to consider. We focus on clusters of size two because of interest in the two sacroiliac joints among patients with psoriatic arthritis.

However, clusters can naturally be much larger as would be the case if all joints were to be modelled. Dealing with larger cluster sizes is in principle straightforward but may suggest the use of second-order generalised estimating functions rather than likelihood analyses. One may elect to retain the robustness of a first order analysis by refraining from higher order assumptions, or invoke further moment assumptions to try to optimise efficiency at the expense of robustness in the estimating equation framework.

We have also restricted attention to a first order Markov model in the longitudinal context with only three assessments. Longer term follow-up, as is planned for the Canadian Longitudinal Study in Aging (Raina et al., 2009) raises questions about the need for more elaborate response models, the need for greater collapsing of strata, and issues surrounding time-varying covariates. These and other issues are subject to further research.

Chapter 4

Adaptive Stratified Two-Phase Sampling

In Section 3.6, we identified two potential shortcomings with our recommended phase-II sampling design for studies of expensive covariates in the presence of budgetary constraints: (i) the size of the phase-II sample, and therefore the cost of the study, was random and fixed only in expectation; and (ii) the specification of parameter values at the design stage would typically require good *a priori* knowledge of the covariate effects and the other parameters or the collection of expensive external pilot data. In this chapter, we aim to address these concerns by proposing a multiphase, adaptive, basic stratified sampling design for efficient estimation through inverse probability weighted estimating equations.

4.1 Inverse Probability Weighting with Basic Stratified Sampling

With basic stratified sampling (BSS), we make sampling decisions about the number of individuals to be selected from each stratum, $\{n_{yv}\}$, rather than about the selection probabilities $\{\pi(y,v)\}$. This approach differs from the previously considered variable probability sampling (VPS) in that the phase-II sample size, $n = \sum_{yv} n_{yv}$, is fixed and the sampling decisions are no longer independent within strata.

Lawless et al. (1999) derived the asymptotic variance of a number of estimators under both VPS and BSS. In an effort to exploit the simplicity, robustness, and general utility of designs based on inverse probability weighted (IPW) analyses demonstrated in the previous chapter, we choose to focus on this framework here.

4.1.1 Asymptotic Variance of IPW under BSS

As in Lawless et al. (1999) and under mild regularity conditions discussed by Wild (1991), the estimator $\widehat{\alpha}$ which solves the weighted unbiased score equation

$$\bar{U}(\alpha) = \sum_{i=1}^{N} R_i \pi_i^{-1} U_i(\alpha) = 0$$

has the property that

$$\sqrt{N}(\widehat{\alpha} - \alpha) \xrightarrow{p} N(0, \bar{\mathbf{A}}(\Psi)^{-1} \bar{\mathbf{B}}(\Omega) \bar{\mathbf{A}}(\Psi)^{-1}),$$

where

$$\bar{\mathbf{A}}(\Psi) = plim(-N^{-1} \partial \bar{U}(\alpha)/\partial \alpha') = \lim E[-N^{-1} \partial \bar{U}(\alpha)/\partial \alpha'],$$

and $\bar{\mathbf{B}}(\Omega) = \lim \text{var}(N^{-1/2} \bar{U}(\alpha))$. As in the case of VPS sampling,

$$\bar{\mathbf{A}}(\Psi) = \lim\ E_{RYXV}[-\frac{1}{N} \sum_{i=1}^{N} \partial R_i \pi_i^{-1} U_i(\alpha)/\partial \alpha'] = E_{YXV}[-\partial U_i(\alpha)/\partial \alpha'].$$

However, now

$$\begin{aligned}
\bar{\mathbf{B}}(\Omega) &= \lim\ \text{var}(N^{-1/2} \sum_{i=1}^{N} R_i \pi_i^{-1} U_i(\alpha)) \\
&= \lim \left\{ \text{var}_{YXV}(E_{R|YXV}[N^{-1/2} \sum_{i=1}^{N} R_i \pi_i^{-1} U_i(\alpha)]) + E_{YXV}(\text{var}_{R|YXV}[N^{-1/2} \sum_{i=1}^{N} R_i \pi_i^{-1} U_i(\alpha)]) \right\} \\
&= \lim\ N^{-1} \left\{ \text{var}_{YXV}(\sum_{i=1}^{N} U_i(\alpha)) + E_{YXV}(\text{var}_{R|YXV}[\sum_{i=1}^{N} R_i \pi_i^{-1} U_i(\alpha)]) \right\}
\end{aligned}$$

We use $\pi_i = P(R_i = 1|y_i, v_i) = n_{y_i,v_i}/N_{y_i,v_i}$ to denote the sampling probability for an individual i and denote the second-order inclusion probability as $\pi_{ij} = P(R_i = 1, R_j = 1|y_i, v_i, y_j, v_j)$. Under VPS, $\pi_{ij} = \pi_i \cdot \pi_j$; under BSS, $\pi_{ij} = \pi_i \cdot \pi_j = n_{y_i,v_i}/N_{y_i,v_i} \cdot n_{y_j,v_j}/N_{y_j,v_j}$ if individuals i and j are from different strata (i.e. if $(y_i, v_i) \neq (y_j, v_j)$), while $\pi_{ij} = n_{y_i,v_i}/N_{y_i,v_i} \cdot (n_{y_i,v_i} - 1)/(N_{y_i,v_i} - 1)$ if

they are from the same stratum (i.e. $(y_i, v_i) = (y_j, v_j)$). So, under BSS

$$
\begin{aligned}
& E_{YXV}\left[\mathrm{var}_{R|YXV}[\sum_{i=1}^{N} R_i \pi_i^{-1} U_i(\alpha)]\right] \\
&= E_{YXV}\Big[\sum_{i=1}^{N} \mathrm{var}_{R|YXV}(R_i)\pi_i^{-2} U_i(\alpha)U_i(\alpha)' + \\
&\qquad \sum_{i=1}^{N}\sum_{j=1;i\neq j}^{N} cov(R_i, R_j | y_i, x_i, v_i, y_j, x_j, v_j)\pi_i^{-1}\pi_j^{-1} U_i(\alpha)U_j'(\alpha)\Big] \\
&= E_{YXV}\Big[\sum_{i=1}^{N}(\pi_i - \pi_i^2)\pi_i^{-2} U_i(\alpha)U_i'(\alpha) + \sum_{i=1}^{N}\sum_{j=1;i\neq j}^{N} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} U_i(\alpha)U_j'(\alpha)\Big] \\
&= E_{YXV}\Big[\sum_{i=1}^{N}(\pi_i^{-1} - 1)U_i(\alpha)U_i'(\alpha)\Big] + \\
&\qquad \sum_{i=1}^{N} E_{YV}\left[(N_{YV} - 1)\left(\frac{n_{YV} - 1}{N_{YV} - 1}\frac{N_{YV}}{n_{YV}} - 1\right) E_{X|YV}[U_i(\alpha)]E_{X|YV}[U_i'(\alpha)]\right] \\
&= E_{YXV}\Big[\sum_{i=1}^{N}(\pi_i^{-1} - 1)U_i(\alpha)U_i'(\alpha)\Big] - \sum_{i=1}^{N} E_{YV}\left[\left(\pi_i^{-1} - 1\right) E_{X|YV}[U_i(\alpha)]E_{X|YV}[U_i'(\alpha)]\right] \\
&= \sum_{i=1}^{N} E_{YV}\left[(\pi_i^{-1} - 1)\left(E_{X|YV}[U_i(\alpha)U_i'(\alpha)] - E_{X|YV}[U_i(\alpha)]E_{X|YV}[U_i(\alpha)']\right)\right]
\end{aligned}
$$

Therefore, $\bar{\mathbf{B}}(\Omega) = E[U_i(\alpha)U_i'(\alpha)] + \sum_{YV} P(Y,V)(\pi(Y,V)^{-1} - 1)\mathrm{var}_{X|YV}[U_i(\alpha)U_i'(\alpha)]$, and the asymptotic variance of the IPW estimator under BSS is

$$\bar{\mathbf{A}}(\Psi)^{-1} + \bar{\mathbf{A}}(\Psi)^{-1}\bar{\mathbf{B}}^*(\Omega)\ \bar{\mathbf{A}}(\Psi)^{-1}, \tag{4.1}$$

where

$$\bar{\mathbf{B}}^*(\Omega) = \sum_{Y,V} P(Y,V)\Big[\frac{N_{YV}}{n_{YV}} - 1\Big] \cdot \mathrm{var}_{X|Y,V}[U_i(\alpha)],$$

which is equivalent to the asymptotic variance for IPW_2 derived in 3.3.4 under VPS.

4.2 IPW Analysis for Multi-Phase Sampling

Consider now a multi-phase sampling procedure for selection of the expensive covariate X. Suppose that a total of n unique individuals can be selected for measurement of X in any of T rounds of sampling, where we use $R_i^{(t)}$ to indicate selection of an individual at time t, and individu-

als are only selected once so $R_i = \sum_{t=1}^{T} R_i^{(t)} = 1$ for those individuals selected for measurement of X. Analysis based on IPW estimating equations is quite appealing in this setting since the probability that an individual is selected for measurement of X is simply

$$\begin{aligned}
&P(R_i = 1|y_i, v_i) \\
&\quad = P(R_i^{(1)} = 1 \cup R_i^{(2)} = 1 \cup \cdots \cup R_i^{(T)} = 1|y_i, v_i) \\
&\quad = P(R_i^{(1)} = 1|y_i, v_i) + P(R_i^{(2)} = 1|y_i, v_i, R_i^{(1)} = 0) \cdot P(R_i^{(1)} = 0|y_i, v_i) \\
&\qquad + \cdots + P(R_i^{(T)} = 1|y_i, v_i, R_i^{(T-1)} = 0) \cdot P(R_i^{(T-1)} = 0|y_i, v_i) \\
&\quad = \frac{n_{yv}^{(1)}}{N_{yv}} + \frac{n_{yv}^{(2)}}{N_{yv} - n_{yv}^{(1)}} \frac{N_{yv} - n_{yv}^{(1)}}{N_{yv}} + \cdots + \frac{n_{yv}^{(T-1)}}{N_{yv} - n_{yv}^{(1)} - \cdots - n_{yv}^{(T-1)}} \frac{N_{yv} - n_{yv}^{(1)} - \cdots - n_{yv}^{(T-1)}}{N_{yv}} \\
&\quad = N_{yv}^{-1}[N_{yv} + n_{yv}^{(1)} + \cdots + n_{yv}^{(T-1)}] \\
&\quad = \frac{n_{yv}}{N_{yv}},
\end{aligned}$$

where $n_{yv} = \sum_{t=1}^{T} n_{yv}^{(t)}$ is the total number of individuals selected from the stratum defined by (y, v) and $n_{yv}^{(t)}$ is the corresponding number selected from the stratum at the tth stage of sampling, $t = 1, 2, \ldots, T$. The IPW estimating equation will therefore be the same whether the phase-II data are selected simultaneously or sequentially (Whittemore and Halpern, 1997).

In the previous chapter, we discussed using external pilot data to estimate the design parameters so that phase-II sampling could be optimally efficient. Here we consider a multi-phase sampling procedure such that each phase of sampling uses the previously collected data as an internal pilot study for estimating the optimal sampling procedure for the next phase of sampling. This idea of using internal pilot studies to modify the study design without discarding data was advocated by Wittes and Brittain (1990). Lohr (1990) used a similar approach, called *triple sampling*, in estimation of an unknown multivariate mean. Pepe et al. (1994) acknowledged the possibility of such an adaptive optimal IPW designs and stated that this approach needs further exploration.

4.3 Adaptive Multi-Phase Sampling

In Section 3.4, we considered VPS with the budgetary constraint

$$P(R = 1; \delta) = \sum_{Y,V} \pi(Y, V; \delta) \cdot N_{YV}/N = P_R,$$

and assumed that N_{YV} (the size of the phase-I sample strata) were known at the design stage. We focussed on finding the optimal $\pi(y, v; \delta)$, which amounted to finding the best expected stratum

sampling weights. With BSS, we focus on choosing the actual sample sizes n_{yv} for each stratum defined by (y,v). Here the budgetary constraint is based on the actual (not expected) cost of the study through specification of n, the number of individuals that can be sampled at phase-II, and we partition our phase-II sampling into multiple phases.

In its simplest form, this adaptive multi-stage phase-II design will consist of two stages of selection: *phase IIa* in which individuals are selected without any knowledge of the true parameters, and *phase IIb* in which we exploit the phase-IIa data to obtain parameter estimates to guide the sampling of remaining phase-II individuals. Phase IIa will involve selecting $n^{(a)}$ individuals under proportional or balanced stratified sampling. Phase-IIb will involve selecting the remaining $n^{(b)} = n - n^{(a)}$ individuals optimally for IPW analysis (that is, optimally assuming the parameter estimates from phase IIa are the true parameters). For simplicity, we continue to discuss a two-stage phase-II design, but this procedure can easily be generalised to a sampling design with an arbitrary number of stages. The effect on efficiency of the choice of the number of stages for sampling and the relative sizes of each sampling stage will be explored.

Our sampling plan now involves deciding how to distribute the $n = n^{(a)} + n^{(b)}$ individuals amongst the phase-I strata under the budgetary constraint

$$n = \sum_{Y,V} [n_{YV}^{(a)} + n_{YV}^{(b)}],$$

and we wish to minimize the asymptotic variance of the estimator of interest as defined by the $[2,2]$ entry in (4.1). Therefore, we wish to find the $\{n_{YV}^{(b)}\}$ that minimise

$$\begin{aligned}\Lambda = &\left\{ \bar{\mathbf{A}}(\Psi)^{-1} + \bar{\mathbf{A}}(\Psi)^{-1} \{ \sum_{Y,V} P(Y,V) \big[\frac{N_{YV}}{n_{YV}^{(a)} + n_{YV}^{(b)}} - 1 \big] \cdot \mathrm{var}_{X|Y,V}[U_i(\alpha)] \} \bar{\mathbf{A}}(\Psi)^{-1} \right\}_{[2,2]} \\ &+ \lambda [\sum_{Y,V} (n_{YV}^{(a)} + n_{YV}^{(b)}) - n],\end{aligned}$$

so the optimal $n_{YV}^{(b)}$ is the solution to

$$\frac{\partial \Lambda}{\partial n_{YV}^{(b)}} = \left\{ \bar{\mathbf{A}}(\Psi)^{-1} \{ \frac{N_{YV}}{(n_{YV}^{(a)} + n_{YV}^{(b)})^2} P(Y,V) \mathrm{var}_{X|Y,V}[U_i(\alpha)] \} \bar{\mathbf{A}}(\Psi)^{-1} \right\}_{[2,2]} + \lambda = 0.$$

At the true Ω, given the phase-IIa sample sizes $\{n_{YV}^{(a)}\}$, the optimal stratum-specific phase-IIb sampling sizes for IPW estimation of α_x satisfy

$$n_{yv}^{(b)} = \lambda^* \big[N_{yv} P(y,v) \big]^{1/2} \left\{ \bar{\mathbf{A}}(\Psi)^{-1} \mathrm{var}_{X|y,v} \big[U_i(\alpha) \big] \bar{\mathbf{A}}(\Psi)^{-1} \right\}_{[2,2]}^{1/2} - n_{yv}^{(a)},$$

so the optimal design is

$$n_{yv}^{(b,\text{opt})} = \frac{n \cdot N_{yv}^{1/2} P(y,v)^{1/2} \cdot \left\{ \bar{\mathbf{A}}(\Psi)^{-1} \text{var}_{X|y,v}\left[U_i(\alpha)\right] \bar{\mathbf{A}}(\Psi)^{-1} \right\}_{[2,2]}^{1/2}}{\sum_{Y,V} N_{YV}^{1/2} P(Y,V)^{1/2} \cdot \left\{ \bar{\mathbf{A}}(\Psi)^{-1} \text{var}_{X|Y,V}\left[U_i(\alpha)\right] \bar{\mathbf{A}}(\Psi)^{-1} \right\}_{[2,2]}^{1/2}} - n_{yv}^{(a)} \tag{4.2}$$

under the constraints $0 \leq n_{yv}^{(b)} \leq N_{yv} - n_{yv}^{(a)}$. We can solve (4.2) along the boundary as in Reilly and Pepe (1995) (i.e. if $n_{yv}^{(b)} > N_{yv} - n_{yv}^{(a)}$, for some (y,v), then we set $n_{yv}^{(b)} = N_{yv} - n_{yv}^{(a)}$ and solve for the others using an updated version of (4.2)). Here we need not worry about the degenerate case $n_{yv}^{(b)} = 0$, but we must simply ensure that $n_{yv}^{(a)} + n_{yv}^{(b)} > 0$.

4.4 Empirical Properties of Adaptive Multi-phase Designs

We compare here the empirical efficiency of adaptive two-phase designs to that of the optimal design (which is based on the true, unknown parameters) for estimation of α_x in the each of the three main settings considered in Chapter 3: a binary expensive covariate, a continuous expensive covariate, and clustered binary response data.

4.4.1 Simulation Results for a Binary X

Here, we consider selection in phase-II in several steps: we first select n_a individuals using (i) proportional stratified sampling, or (ii) balanced sampling; and then select the remaining $n_b = n - n_a$ individuals according to the optimal IPW design based on the preliminary parameter estimates from the phase-IIa sample. Here, we consider basic stratified sampling, so that the number of individuals selected at phase II is not random. We considered a dataset of $N = 1000$ individuals at phase I, and phase-II sample sizes of $n = 250$ and 500; these values are chosen to correspond roughly to the total size of the psoriatic arthritis cohort of approximately 1000 individuals and settings with low and moderate budgets for biomarker studies. In both cases, we considered three possible choices for the proportion of the phase-II sample to select at phase IIa: 100% ($n_a = n$), 60%, and 20% (note that when $n_a = n$, $n_b = 0$ and this is not adaptive sampling but represents a scenario where the entire phase-II sample is chosen through proportional stratified sampling or balanced sampling). As in Pepe et al. (1994) and as with the external pilot data discussed in Chapter 3, we attempt to stabilise the estimation of the design parameters by adding one observation to each strata when analysing the phase-IIa data. Parameter values were chosen as in the previous chapter so that $\Psi_0 = (\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_v, \gamma_0)' = (-1.95, 1.00, 0.90, 1.05, -0.41, -0.04)'$.

Each design was employed for 1000 simulated datasets and the differences between these designs and the optimal designs are presented in Figures 4.1 and 4.2. Figure 4.1 displays the difference between the the optimal and the design-specific proportion of individuals selected for measurement of X from each stratum (i.e. $n_{YV}/N_{YV} - n_{YV}^{\text{opt}}/N_{YV}$) for $n = 250$; Figure 4.2 displays these differences in sampling fractions for $n = 500$. The optimal design here utilise the true parameters Ψ_0 that generate the data, while the adaptive designs attempt to approximate this design.

The empirical biases, empirical standard errors, and empirical coverage probabilities of the estimators resulting from the different designs are presented in Tables 4.1 and 4.2 for n=250 and 500, respectively.

The adaptive designs resulted in sampling fractions which were much closer to the optimal designs than either proportional or balanced sampling alone (see Figures 4.1 and 4.2). This near-optimality is reflected in Tables 4.1 and 4.2, where it can be seen that the adaptive designs increased the empirical efficiency of the resultant estimators of α_x. These adaptive designs also resulted in an increase in efficiency for estimation of α_0; the efficiency with which α_v was estimated changed little between designs.

The non-adaptive proportional sampling design was quite inefficient here. The balanced design offered improved efficiency over proportional stratified sampling. However, the adaptive designs all had much greater empirical efficiency than both non-adaptive designs. In all cases, the adaptive design which allocated 20% of the sample to phase IIa resulted in empirical efficiencies that were very similar to the true optimal design. These designs did not result in any bias and the empirical coverage probabilities were very close to the nominal 95% level even when the sample size was small. It is interesting to note that the adaptive design that allocated 20% of the sample to phase IIa lead to greater efficiency than the design which allocated 60%, but the sampling probabilities were less variable for the latter design. These results demonstrate empirically that there is an important trade-off between the efficiency that can be achieved through increasing the size of the phase-IIb sample and the variability of the design that results from the less precise phase-IIa estimates that result from smaller phase-IIa samples. Here, however, allocating 20% of the sample to phase IIa was sufficient to achieve levels of efficiency which were very close to optimal. A similar phenomenon was demonstrated in Chapter 3 where small pilot studies were sufficient to derive near-optimal designs for IPW analysis.

In all cases, the empirical coverage probabilities were close to the nominal 95% level, which indicates that the estimated asymptotic standard errors for the adaptive design closely tracked the empirical standard errors. The estimated asymptotic standard errors from the adaptive designs were very similar to those from the asymptotically optimal design. However, the empirical standard errors of the adaptive designs were actually sometimes smaller than the empirical standard errors from the optimal design (see Table 4.1). Therefore, the regular variance estimators worked well for these adaptive designs, which achieved efficiency that was very similar to the unknowable optimal design.

In the next sections we will consider the efficiency of these same two-stage phase-II designs in

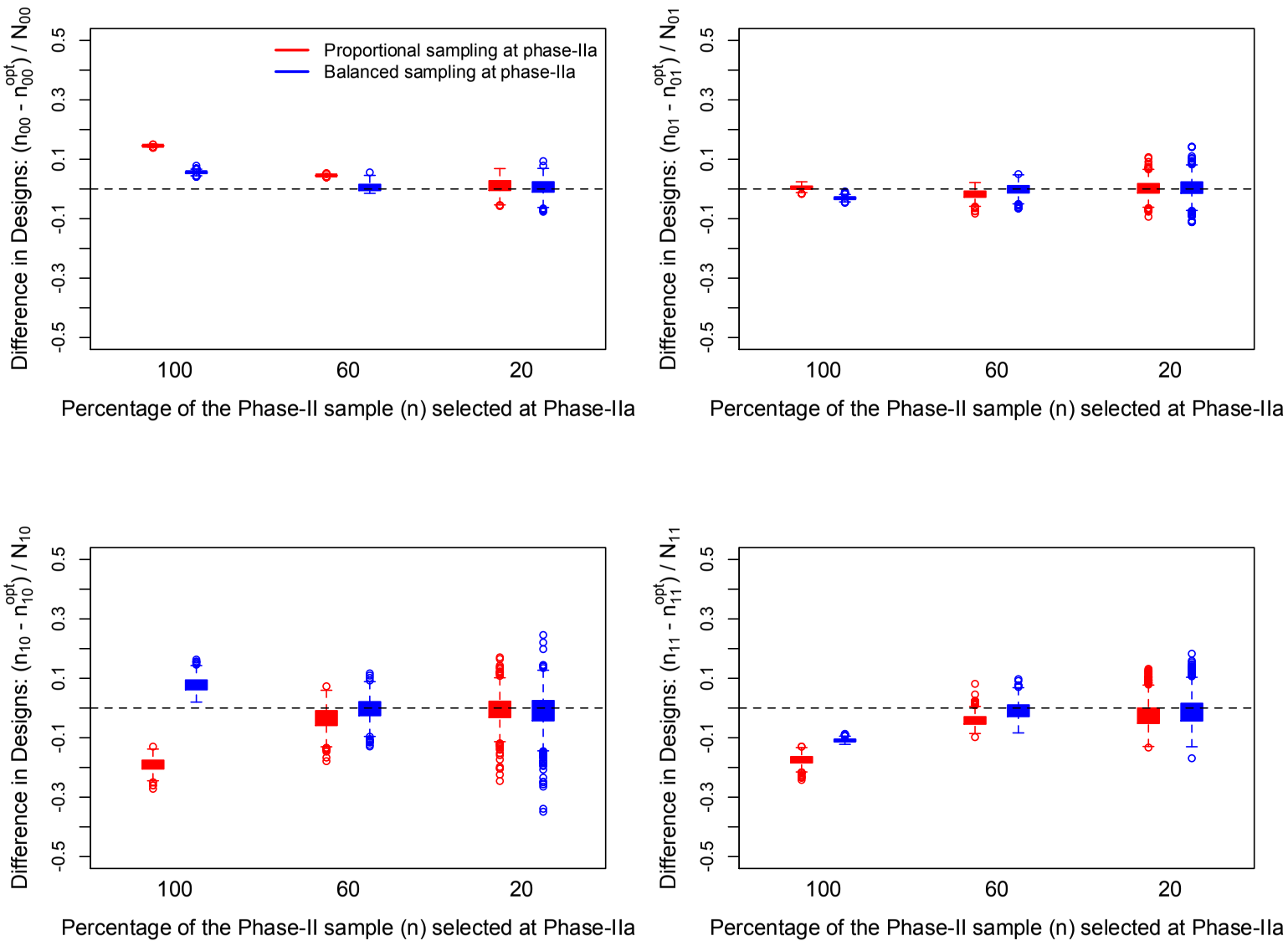

Figure 4.1: Boxplots displaying the differences in sampling fractions between the optimal and other designs for each simulated dataset with univariate responses and a binary expensive covariate when $N = 1000$, $n = 250$ and $(\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_v, \gamma_0)' = (-1.95, 1.00, 0.90, 1.05, -0.41, -0.04)'$.

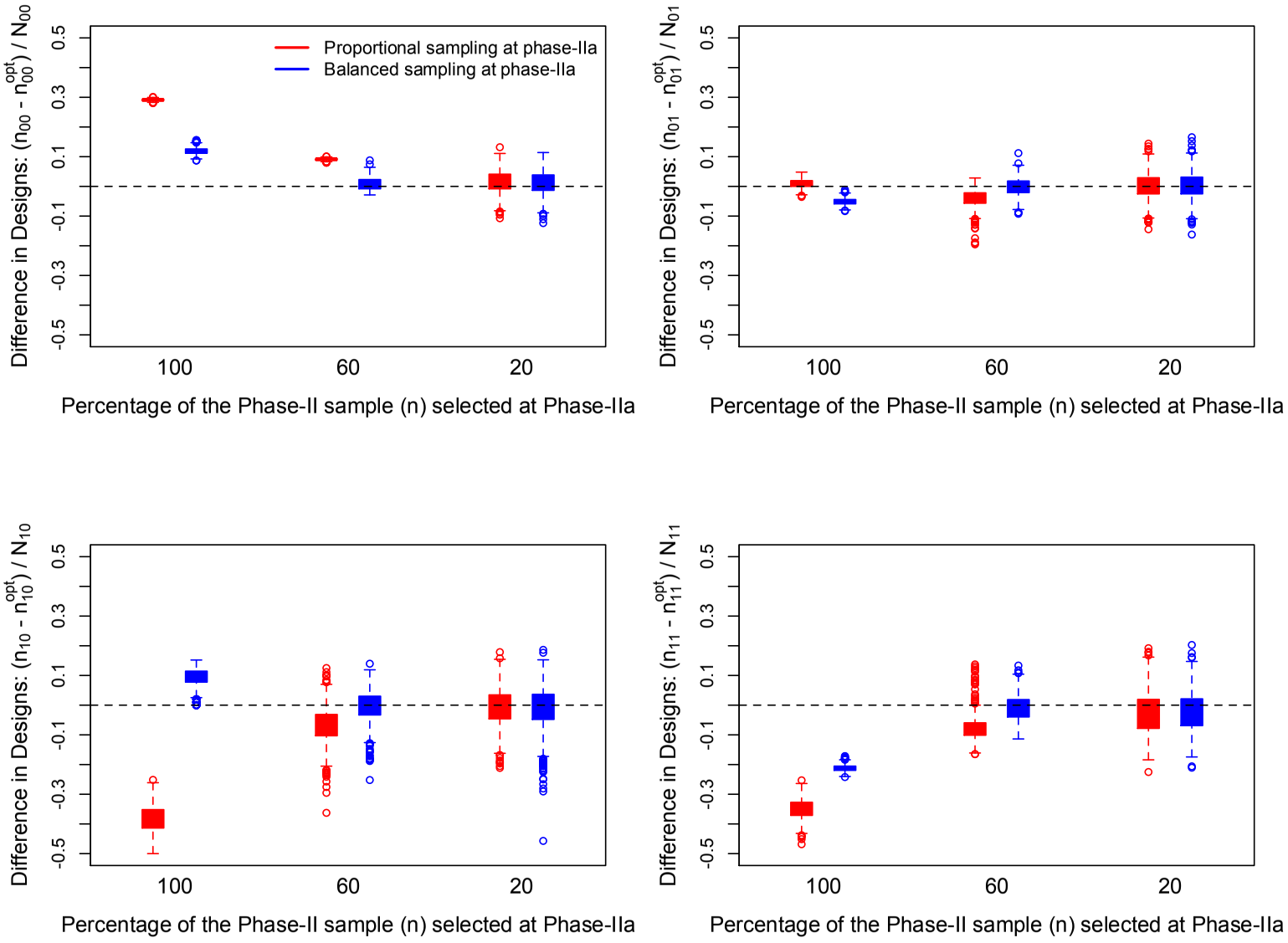

Figure 4.2: Boxplots displaying the differences in sampling fractions between the optimal and other designs for each simulated dataset with univariate responses and a binary expensive covariate when $N = 1000, n = 500$ and $(\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_v, \gamma_0)' = (-1.95, 1.00, 0.90, 1.05, -0.41, -0.04)'$.

the context of data involving a continuous expensive covariate and clustered responses before returning to this univariate setting to consider the effect of using more stages of sampling.

Table 4.1: Empirical biases, standard errors, and coverage probabilities for estimators based on 1000 simulated datasets with binary data when the overall phase-II sample size is $n = 250$ and phase-II sampling is carried out in two stages – phase-IIa based on proportional or balanced sampling, and phase IIb designed for optimally efficient estimation of α_x assuming that parameter estimates from phase-IIa data are correct. The efficiency of an optimal design based on true (unknown) parameters[†] is presented for comparison.

		Percentage of the Phase-II sample (n) selected at Phase-IIa								
Phase-IIa		100 %			60 %			20 %		
Design	Parameter	Bias	ESE	ECP	Bias	ESE	ECP	Bias	ESE	ECP
Optimal	α_0	-0.013	0.270	95.0	–	–	–	–	–	–
	α_x	0.012	0.305	94.8	–	–	–	–	–	–
	α_v	0.001	0.163	94.8	–	–	–	–	–	–
Proportional	α_0	-0.025	0.310	95.8	-0.017	0.272	95.1	-0.012	0.262	96.6
	α_x	0.017	0.337	96.3	0.014	0.297	95.4	0.005	0.294	96.2
	α_v	0.007	0.160	94.8	0.005	0.162	94.8	0.003	0.160	94.7
Balanced	α_0	-0.027	0.283	96.0	-0.009	0.267	95.8	-0.013	0.267	95.6
	α_x	0.022	0.314	95.6	0.004	0.300	95.5	0.008	0.300	95.6
	α_v	0.006	0.161	94.9	0.001	0.161	94.6	0.003	0.159	95.4

[†] $(\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_v, \gamma_0)' = (-1.95, 1.00, 0.90, 1.05, -0.41, -0.04)'$.

Table 4.2: Empirical biases, standard errors, and coverage probabilities for estimators based on 1000 simulated datasets with binary data when the overall phase-II sample size is $n = 500$ and phase-II sampling is carried out in two stages – phase-IIa based on proportional or balanced sampling, and phase IIb designed for optimally efficient estimation of α_x assuming that parameter estimates from phase-IIa data are correct. The efficiency of an optimal design based on true (unknown) parameters[†] is presented for comparison.

		Percentage of the Phase-II sample (n) selected at Phase-IIa								
Phase-IIa		100 %			60 %			20 %		
Design	Parameter	Bias	ESE	ECP	Bias	ESE	ECP	Bias	ESE	ECP
Optimal	α_0	-0.004	0.205	94.6	–	–	–	–	–	–
	α_x	0.004	0.210	95.3	–	–	–	–	–	–
	α_v	-0.002	0.147	95.2	–	–	–	–	–	–
Proportional	α_0	-0.013	0.231	95.0	-0.011	0.208	94.8	0.002	0.205	95.1
	α_x	0.010	0.241	95.4	0.011	0.220	94.8	-0.007	0.219	93.6
	α_v	0.002	0.145	95.3	0.001	0.147	94.9	-0.001	0.146	95.0
Balanced	α_0	-0.010	0.214	94.8	-0.003	0.199	95.7	0.001	0.200	95.6
	α_x	0.008	0.224	94.1	0.000	0.209	94.4	-0.005	0.214	94.6
	α_v	0.002	0.146	95.6	0.000	0.144	95.0	-0.002	0.143	95.3

[†] $(\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_v, \gamma_0)' = (-1.95, 1.00, 0.90, 1.05, -0.41, -0.04)'$.

4.4.2 Simulation Results for a Continuous X

Here we consider the case of a continuous expensive exposure variable where $X|V$ arises from a gamma distribution with shape β_0 and scale $\beta_1 + \beta_v V$, as in (3.20); the response Y and auxiliary covariate V were again generated according to the models (3.1) and (3.19). Because Y and V are still discrete, strata are defined as before at phase I. The parameter values used in generating these data were, as in Section 3.4.2, $\Psi_{c0}' = (\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_1, \beta_v, \gamma_0)' =$ (-2.18, 0.03, .84, 1.40, 10, 5, -.04)$'$, where values of β were chosen to reflect the distribution of actual MMP-3 values given ESR status seen in the PsA pilot data.

One thousand datasets were generated and the empirical biases and estimated asymptotic standard errors of the estimators resulting from the use of the different sampling designs were recorded. The differences between the sampling fractions for the adaptive and optimal designs are displayed in Figures 4.3 and 4.4 for selection of $n = 250$ and 500 out of $N = 1000$ individuals, respectively. The empirical biases, standard errors, and coverage probabilities of the estimators resulting from the different designs are reported in Tables 4.3 and 4.4, for n=250 and 500, respectively.

Again it can be seen that the adaptive designs produce sampling fractions that are much closer to the optimal designs than those from proportional or balanced sampling alone. These adaptive designs are also more efficient than the non-adaptive designs for estimation of α_x; in fact, both versions of the adaptive design were more efficient than either non-adaptive design in the estimation of both α_x and α_0 for both $n = 250$ and $n = 500$. The adaptive designs were also more efficient than proportional sampling for estimation of α_v. Here the adaptive design that allocates 20% of the sample to phase IIa was not always preferable to the adaptive design that allocates 60% of the sample to phase IIa in terms of empirical efficiency in estimation of α_x. There were also situations here where the empirical efficiency of the adaptive designs was actually greater than that from the asymptotically optimal designs for estimation of α_x. In all cases, the empirical coverage probabilities were near the nominal 95% level.

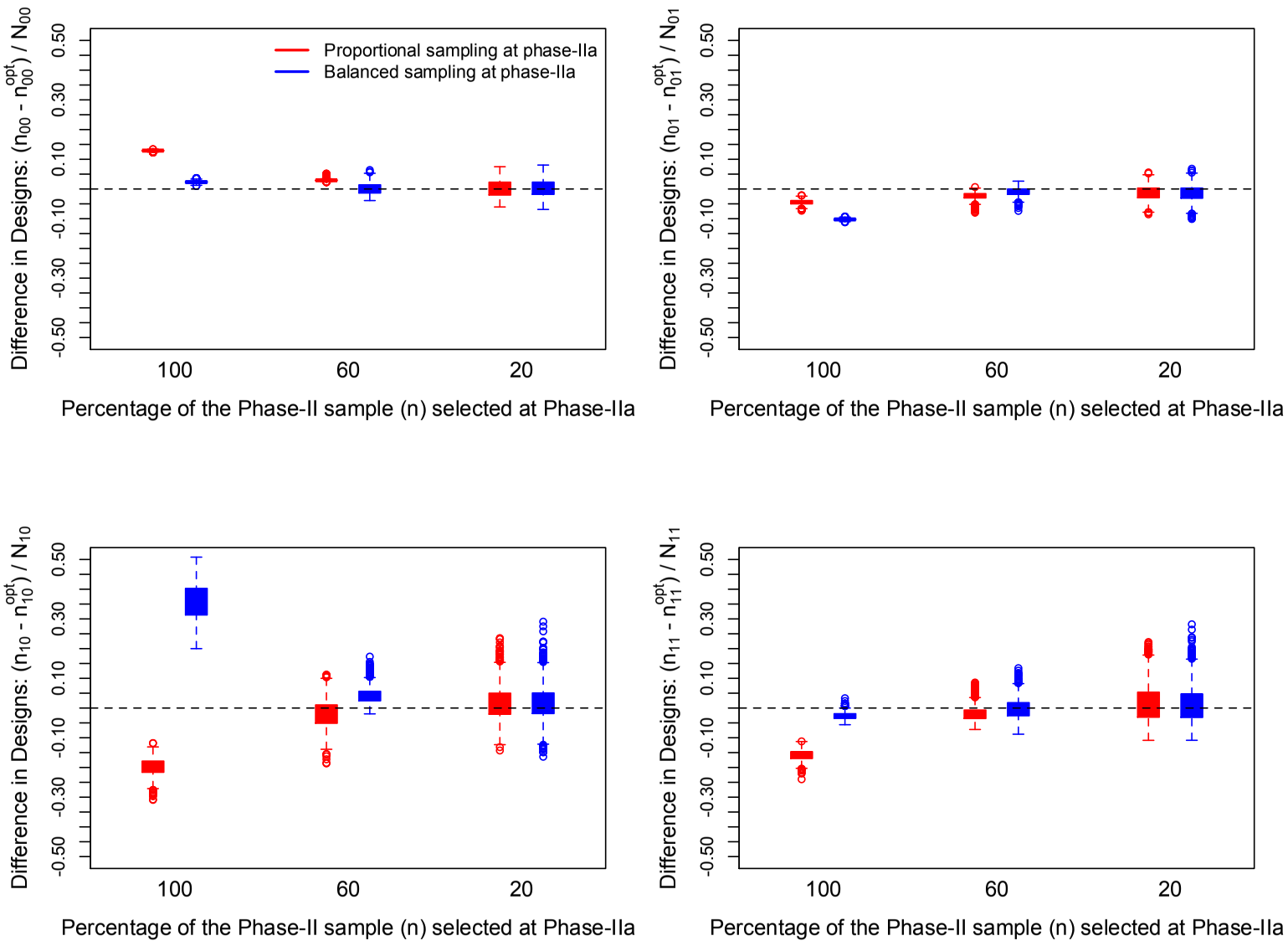

Figure 4.3: Boxplots displaying the differences in sampling fractions between the optimal and other designs for each simulated dataset with univariate responses and a continuous expensive covariate when $N = 1000$, $n = 250$ and $(\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_1, \beta_v, \gamma_0)' =$ (-2.18, 0.03, .84, 1.40, 10, 5, -.04)′

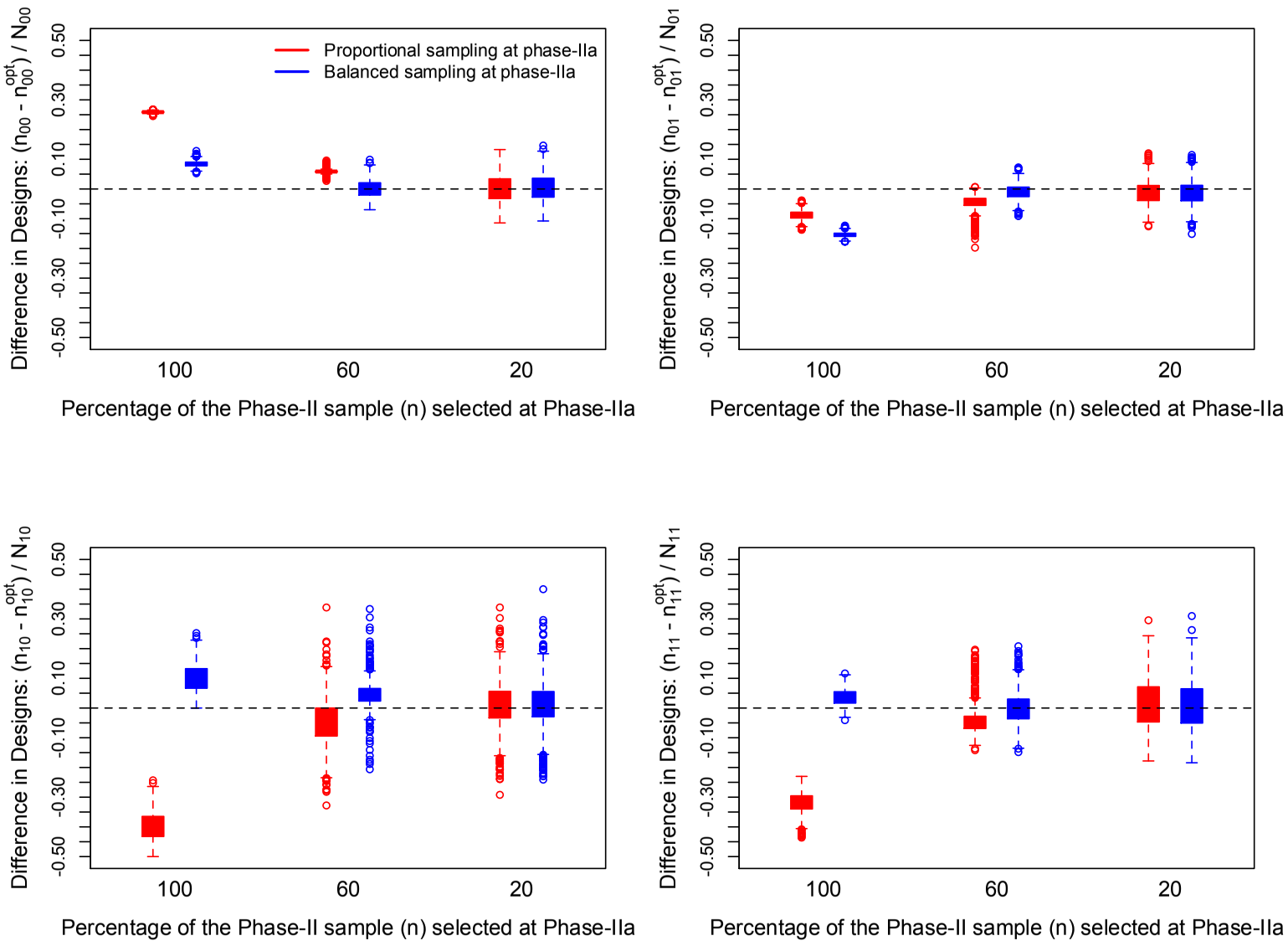

Figure 4.4: Boxplots displaying the differences in sampling fractions between the optimal and other designs for each simulated dataset with univariate responses and a continuous expensive covariate when $N = 1000, n = 500$ and $(\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_1, \beta_v, \gamma_0)' =$ (-2.18, 0.03, .84, 1.40, 10, 5, -.04)′

Table 4.3: Empirical biases, standard errors, and coverage probabilities for estimators based on 1000 simulated datasets with a continuous X when the overall phase-II sample size is $n = 250$ and phase-II sampling is carried out in two stages – phase-IIa based on proportional or balanced sampling, and phase IIb designed for optimally efficient estimation of α_x assuming that parameter estimates from phase-IIa data are correct. The efficiency of an optimal design based on true (unknown) parameters[†] is presented for comparison.

Phase-IIa		Percentage of the Phase-II sample (n) selected at Phase-IIa								
		100 %			60 %			20 %		
Design	Parameter	Bias	ESE	ECP	Bias	ESE	ECP	Bias	ESE	ECP
Optimal	α_0	-0.02151	0.19346	95.6	–	–	–	–	–	–
	α_x	0.00129	0.00923	95.4	–	–	–	–	–	–
	α_v	-0.00953	0.17793	95.7	–	–	–	–	–	–
Proportional	α_0	-0.01291	0.21185	95.4	-0.01872	0.19388	94.7	-0.02784	0.19080	95.1
	α_x	0.00025	0.00984	95.9	0.00106	0.00887	96.0	0.00174	0.00916	95.5
	α_v	0.00221	0.18004	96.2	-0.00844	0.17818	95.1	-0.01332	0.18252	95.2
Balanced	α_0	-0.02133	0.19751	95.4	-0.01912	0.19331	93.7	-0.02176	0.19211	96.1
	α_x	0.00123	0.00961	95.6	0.00110	0.00899	95.2	0.00141	0.00933	95.0
	α_v	-0.00970	0.17770	95.4	-0.00754	0.17710	95.2	-0.01162	0.17851	94.9

[†] $(\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_1, \beta_v, \gamma_0)' = (-2.18, 0.03, .84, 1.40, 10, 5, -.04)$

Table 4.4: Empirical biases, standard errors, and coverage probabilities for estimators based on 1000 simulated datasets with a continuous X when the overall phase-II sample size is $n = 500$ and phase-II sampling is carried out in two stages – phase-IIa based on proportional or balanced sampling, and phase IIb designed for optimally efficient estimation of α_x assuming that parameter estimates from phase-IIa data are correct. The efficiency of an optimal design based on true (unknown) parameters[†] is presented for comparison.

Phase-IIa		Percentage of the Phase-II sample (n) selected at Phase-IIa								
		100 %			60 %			20 %		
Design	Parameter	Bias	ESE	ECP	Bias	ESE	ECP	Bias	ESE	ECP
Optimal	α_0	-0.00228	0.15875	96.2	–	–	–	–	–	–
	α_x	0.00019	0.00619	94.9	–	–	–	–	–	–
	α_v	-0.00665	0.17359	94.1	–	–	–	–	–	–
Proportional	α_0	-0.00557	0.17080	96.1	-0.00405	0.16196	96.1	-0.00744	0.15805	96.0
	α_x	0.00023	0.00674	95.6	0.00029	0.00613	96.2	0.00057	0.00611	94.9
	α_v	-0.00405	0.17285	93.3	-0.00708	0.17009	94.3	-0.00933	0.17239	94.4
Balanced	α_0	-0.00639	0.16464	94.9	-0.00409	0.15709	95.9	-0.00611	0.15669	96.3
	α_x	0.00046	0.00648	95.9	0.00040	0.00603	96.0	0.00040	0.00606	96.5
	α_v	-0.00720	0.17046	94.5	-0.00869	0.17255	93.9	-0.00689	0.17012	94.7

[†] $(\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_1, \beta_v, \gamma_0)' = (-2.18, 0.03, .84, 1.40, 10, 5, -.04)$

4.4.3 Simulation Results for Clustered Responses

We now turn our attention to the setting of multiphase designs involving clustered responses. The more elaborate response model and the difficulty in specifying even tentative values for the association parameters motivated, in part, the exploration of the adaptive phase-II designs we consider in this chapter. Here we consider a balanced sampling design based on the eight strata defined by the eight possible values of (Y_1, Y_2, V), and a balanced sampling based on six strata defined by the six possible values of $(Y_1 + Y_2, V)$, as discussed in Chapter 3. Parameter values were chosen so that $E[Y_1] = 0.4, E[Y_2] = 0.4, \alpha_x = \log 1.25, \alpha_v = \log 1.5, E[X] = 0.6, \beta_v = \log 1.5$, and $\gamma_0 = 0$, as in Section 3.7. We considered both a mild and a strong within-cluster association defined by the conditional odds ratios ψ=1.25 and ψ=3.00, where ψ is the conditional odds ratio

$$\psi = \frac{P(Y_1 = 1, Y_2 = 1|X,V)/P(Y_1 = 0, Y_2 = 1|X,V)}{P(Y_1 = 1, Y_2 = 0|X,V)/P(Y_1 = 0, Y_2 = 0|X,V)}.$$

In each case, 1000 datasets were simulated and corresponding phase-II designs were selected. The distribution of the difference between the stratum-specific sampling fractions for each design and the optimal design can be found in Figures 4.5 and 4.6 for the case of the mild within-cluster association $\psi = 1.25$ and in Figures 4.7 and 4.8 for $\psi = 3.00$. In order to preserve space in these figures, the sampling fractions are only presented for the four strata with $V = 1$; the trends within the other strata are very similar. The empirical biases, empirical standard errors, and empirical coverage probabilities resulting from IPW analysis of the designs can be found in Tables 4.5 - 4.8.

The non-adaptive designs were generally quite dissimilar to the optimal designs. Out of these non-adaptive designs, the design balanced on 6 strata was generally the most similar to the optimal design. Both forms of adaptive sampling resulted in stratum-specific sampling fractions that were much closer to the optimal designs (Figures 4.5 - 4.8). The adaptive designs using $n_a = 0.20n$ were very similar to the optimal designs. The distribution of stratum-specific sampling probabilities from the balanced design based on 8 strata were generally further from the optimal designs than the proportional design. However, when only $n_a = 0.20n$ individuals were selected at phase-IIa, the three designs were quite similar.

None of the non-adaptive designs was consistently more efficient than the others for estimation of α_x; the balanced designs were often less efficient than the proportional stratified sampling design (see Tables 4.5 - 4.8). The adaptive designs generally resulted in much more efficient estimators of α_x than the non-adaptive designs. The adaptive designs using $n_a = 0.20n$ were generally very similar to the (unknown) optimal design in terms of efficiency regardless of the design used at phase-IIa. These designs were generally more efficient than the adaptive design using with $n_a = 0.60n$, which generally offered a substantial improvement over the non-adaptive design in terms of efficiency of estimating α_x. The adaptive designs also generally increased the efficiency with which the other parameters were estimated. In one setting, the non-adaptive design which

balanced on 6 strata achieved an empirical efficiency which was very similar to the optimal design 4.5; even in this setting, the adaptive approach still resulted in smaller empirical standard errors and again demonstrated a superefficiency compared to the asymptotically optimal design. In all cases the empirical coverage probabilities were near the nominal 95% level. These adaptive two-phase designs generally increased the efficiency of parameter estimates in the presence of an expensive covariate without artificially inflating the variance estimates. In general, the adaptive designs were most efficient when most individuals were selected at phase-IIb.

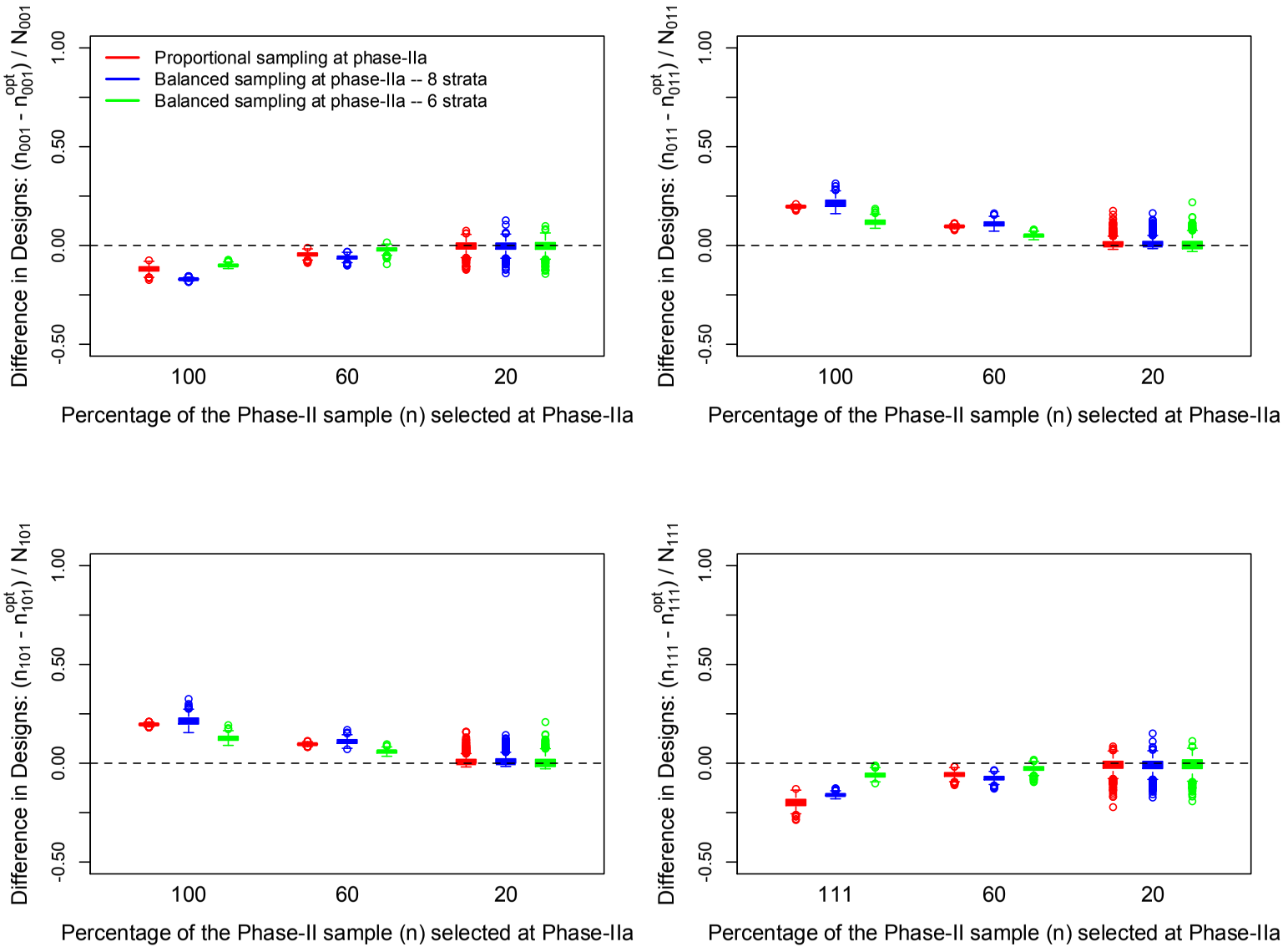

Figure 4.5: Boxplots displaying the differences in sampling fractions between the optimal and other designs in the four strata with $V = 1$ for each simulated dataset with bivariate responses and a binary expensive covariate when $N = 1000$, $n = 250$ and $E[Y_1] = 0.4; E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.5; \alpha_x = \log 1.25; \alpha_v = \log 1.5; \beta_v = \log 1.5; \psi = 1.25$

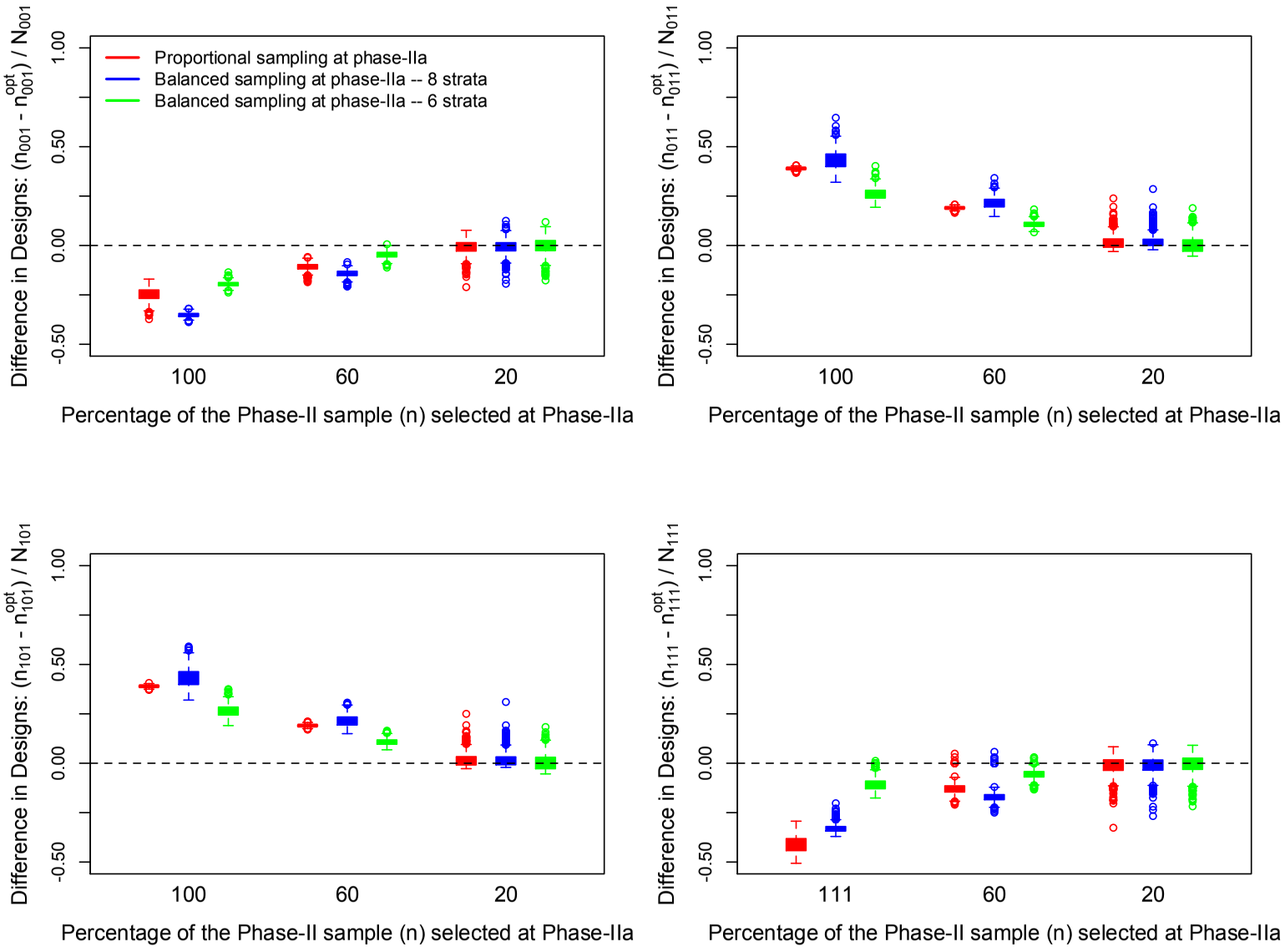

Figure 4.6: Boxplots displaying the differences in sampling fractions between the optimal and other designs in the four strata with $V = 1$ for each simulated dataset with bivariate responses and a binary expensive covariate when $N = 1000, n = 500$ and $E[Y_1] = 0.4; E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.5; \alpha_x = \log 1.25; \alpha_v = \log 1.5; \beta_v = \log 1.5; \psi = 1.25$

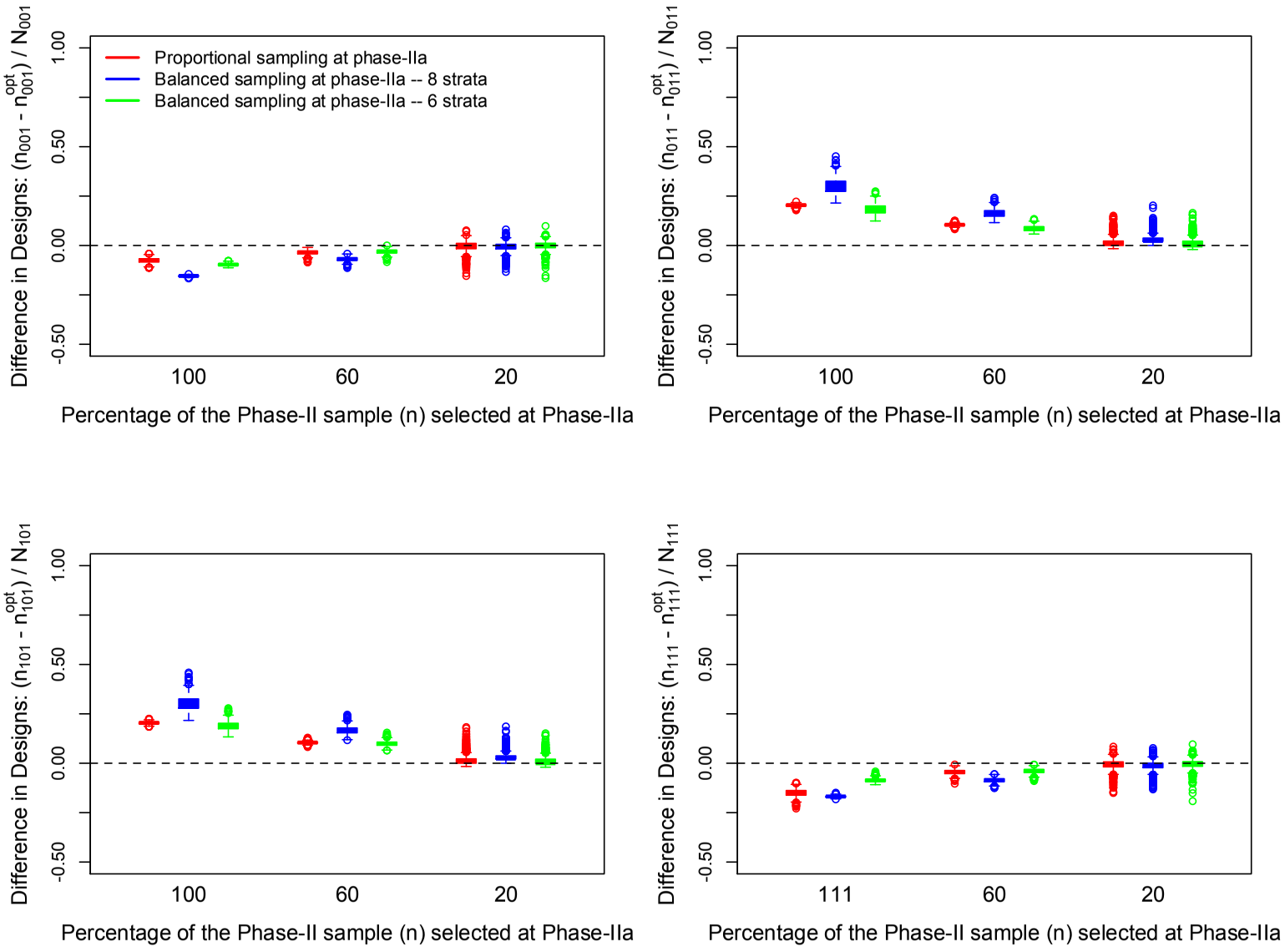

Figure 4.7: Boxplots displaying the differences in sampling fractions between the optimal and other designs in the four strata with $V = 1$ for each simulated dataset with bivariate responses and a binary expensive covariate when $N = 1000$, $n = 250$ and $E[Y_1] = 0.4; E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.5; \alpha_x = \log 1.25; \alpha_v = \log 1.5; \beta_v = \log 1.5; \psi = 3.00$

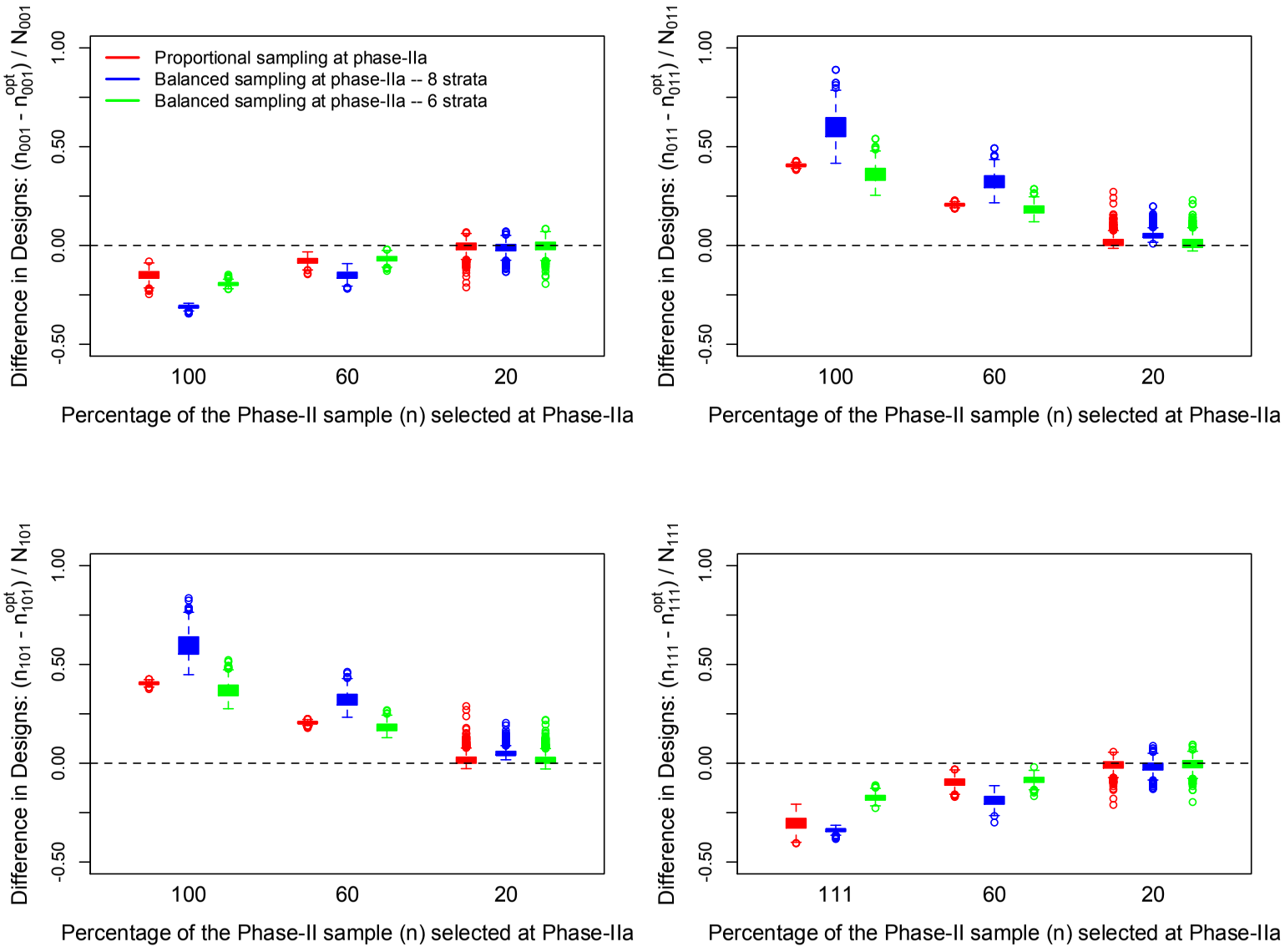

Figure 4.8: Boxplots displaying the differences in sampling fractions between the optimal and other designs in the four strata with $V = 1$ for each simulated dataset with bivariate responses and a binary expensive covariate when $N = 1000, n = 500$ and $E[Y_1] = 0.4; E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.5; \alpha_x = \log 1.25; \alpha_v = \log 1.5; \beta_v = \log 1.5; \psi = 3.00$

Table 4.5: Empirical biases, standard errors, and coverage probabilities for estimators based on 1000 simulated datasets with bivariate response when the overall phase-II sample size is $n = 250$, $\psi = 1.25$, and phase-II sampling is carried out in two stages – phase-IIa based on proportional or balanced sampling, and phase IIb designed for optimally efficient estimation of α_x assuming that parameter estimates from phase-IIa data are correct. The efficiency of an optimal design based on true (unknown) parameters[†] is presented for comparison.

		Percentage of the Phase-II sample (n) selected at Phase-IIa								
Phase-IIa		100 %			60 %			20 %		
Design	Parameter	Bias	ESE	ECP	Bias	ESE	ECP	Bias	ESE	ECP
Optimal	α_0	-0.006	0.120	94.8	–	–	–	–	–	–
	α_x	0.007	0.173	94.4	–	–	–	–	–	–
	α_v	0.001	0.101	94.6	–	–	–	–	–	–
	ψ	-0.005	0.164	94.2	–	–	–	–	–	–
Proportional	α_0	-0.002	0.126	95.9	0.001	0.118	95.5	-0.006	0.118	95.0
	α_x	0.001	0.190	95.2	-0.004	0.169	95.5	0.009	0.166	95.5
	α_v	0.001	0.100	94.9	0.001	0.099	94.9	0.001	0.099	95.8
	ψ	-0.006	0.165	94.8	-0.003	0.164	94.9	-0.004	0.164	95.3
Balanced 8	α_0	-0.004	0.128	96.3	0.004	0.120	94.6	-0.003	0.118	94.6
	α_x	0.009	0.201	96.0	-0.007	0.172	95.5	0.002	0.167	94.4
	α_v	-0.001	0.099	95.1	0.002	0.097	94.8	0.003	0.098	95.3
	ψ	-0.008	0.163	94.7	-0.003	0.164	95.2	-0.004	0.163	94.8
Balanced 6	α_0	0.000	0.114	96.3	0.005	0.115	95.9	0.002	0.113	96.5
	α_x	0.001	0.173	96.1	-0.009	0.164	95.9	-0.005	0.161	95.9
	α_v	0.000	0.100	95.0	0.001	0.098	94.5	0.002	0.099	94.2
	ψ	-0.004	0.165	95.1	-0.002	0.164	94.8	-0.002	0.164	95.1

[†] $E[Y_1] = 0.4; E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.5; \alpha_x = \log 1.25; \alpha_v = \log 1.5; \beta_v = \log 1.5; \psi = 1.25$

Table 4.6: Empirical biases, standard errors, and coverage probabilities for estimators based on 1000 simulated datasets with bivariate response when the overall phase-II sample size is $n = 500$, $\psi = 1.25$, and phase-II sampling is carried out in two stages – phase-IIa based on proportional or balanced sampling, and phase IIb designed for optimally efficient estimation of α_x assuming that parameter estimates from phase-IIa data are correct. The efficiency of an optimal design based on true (unknown) parameters[†] is presented for comparison.

		Percentage of the Phase-II sample (n) selected at Phase-IIa								
Phase-IIa		100 %			60 %			20 %		
Design	Parameter	Bias	ESE	ECP	Bias	ESE	ECP	Bias	ESE	ECP
Optimal	α_0	-0.006	0.097	94.9	–	–	–	–	–	–
	α_x	0.003	0.122	93.5	–	–	–	–	–	–
	α_v	0.003	0.096	95.9	–	–	–	–	–	–
	ψ	0.001	0.170	93.7	–	–	–	–	–	–
Proportional	α_0	-0.006	0.104	95.1	-0.004	0.097	94.7	-0.007	0.096	95.2
	α_x	0.003	0.137	94.7	0.000	0.124	94.6	0.005	0.120	94.8
	α_v	0.003	0.095	95.9	0.004	0.095	95.7	0.003	0.095	96.3
	ψ	-0.001	0.170	93.7	0.001	0.170	93.8	0.001	0.170	94.0
Balanced 8	α_0	-0.002	0.104	95.2	-0.003	0.098	95.9	-0.003	0.097	95.4
	α_x	-0.001	0.143	95.9	-0.001	0.127	93.9	-0.002	0.121	94.1
	α_v	0.003	0.096	95.6	0.004	0.096	95.9	0.004	0.095	96.0
	ψ	-0.001	0.170	93.4	0.001	0.169	93.6	0.001	0.170	93.7
Balanced 6	α_0	-0.005	0.101	94.2	-0.003	0.098	94.0	-0.005	0.098	94.9
	α_x	0.003	0.131	94.2	-0.002	0.121	95.2	0.002	0.120	94.8
	α_v	0.003	0.095	95.5	0.004	0.095	95.7	0.004	0.097	95.8
	ψ	0.000	0.170	93.8	0.001	0.169	94.0	0.001	0.169	93.8

[†] $E[Y_1] = 0.4; E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.5; \alpha_x = \log 1.25; \alpha_v = \log 1.5; \beta_v = \log 1.5; \psi = 1.25$

Table 4.7: Empirical biases, standard errors, and coverage probabilities for estimators based on 1000 simulated datasets with bivariate response when the overall phase-II sample size is $n = 250$, $\psi = 3.00$, and phase-II sampling is carried out in two stages – phase-IIa based on proportional or balanced sampling, and phase IIb designed for optimally efficient estimation of α_x assuming that parameter estimates from phase-IIa data are correct. The efficiency of an optimal design based on true (unknown) parameters[†] is presented for comparison.

		Percentage of the Phase-II sample (n) selected at Phase-IIa								
Phase-IIa		100 %			60 %			20 %		
Design	Parameter	Bias	ESE	ECP	Bias	ESE	ECP	Bias	ESE	ECP
Optimal	α_0	0.003	0.135	94.0	–	–	–	–	–	–
	α_x	0.000	0.190	94.3	–	–	–	–	–	–
	α_v	-0.006	0.111	94.0	–	–	–	–	–	–
	ψ	0.008	0.423	94.2	–	–	–	–	–	–
Proportional	α_0	-0.003	0.144	95.8	0.007	0.133	95.1	-0.002	0.132	95.6
	α_x	0.011	0.212	94.9	-0.004	0.194	95.1	0.011	0.186	95.2
	α_v	-0.006	0.110	93.7	-0.006	0.110	93.6	-0.007	0.111	94.2
	ψ	0.001	0.422	94.0	0.007	0.424	94.0	0.005	0.422	94.2
Balanced 8	α_0	0.007	0.156	94.9	0.002	0.136	94.5	0.002	0.132	94.8
	α_x	-0.002	0.244	94.3	0.005	0.204	95.7	0.004	0.186	95.7
	α_v	-0.007	0.111	93.9	-0.008	0.111	93.7	-0.007	0.112	93.6
	ψ	-0.003	0.425	94.3	0.004	0.425	94.2	0.007	0.421	94.4
Balanced 6	α_0	0.008	0.145	93.7	0.005	0.135	94.7	0.006	0.133	94.8
	α_x	-0.005	0.218	94.9	-0.002	0.194	94.9	-0.005	0.185	95.2
	α_v	-0.007	0.111	93.6	-0.006	0.111	93.0	-0.006	0.109	94.0
	ψ	0.003	0.422	94.4	0.006	0.420	94.1	0.009	0.424	94.1

[†] $E[Y_1] = 0.4; E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.5; \alpha_x = \log 1.25; \alpha_v = \log 1.5; \beta_v = \log 1.5; \psi = 3.00$

Table 4.8: Empirical biases, standard errors, and coverage probabilities for estimators based on 1000 simulated datasets with bivariate response when the overall phase-II sample size is $n = 500$, $\psi = 3.00$, and phase-II sampling is carried out in two stages – phase-IIa based on proportional or balanced sampling, and phase IIb designed for optimally efficient estimation of α_x assuming that parameter estimates from phase-IIa data are correct. The efficiency of an optimal design based on true (unknown) parameters[†] is presented for comparison.

		Percentage of the Phase-II sample (n) selected at Phase-IIa								
Phase-IIa		100 %			60 %			20 %		
Design	Parameter	Bias	ESE	ECP	Bias	ESE	ECP	Bias	ESE	ECP
Optimal	α_0	-0.003	0.103	95.9	–	–	–	–	–	–
	α_x	0.003	0.131	94.0	–	–	–	–	–	–
	α_v	-0.003	0.107	94.3	–	–	–	–	–	–
	ψ	0.003	0.410	94.5	–	–	–	–	–	–
Proportional	α_0	-0.004	0.110	96.1	-0.001	0.101	96.3	0.000	0.102	95.7
	α_x	0.003	0.149	96.3	-0.001	0.130	96.0	-0.003	0.130	95.0
	α_v	-0.001	0.107	94.0	-0.002	0.106	94.5	-0.001	0.107	94.4
	ψ	0.001	0.411	94.5	0.004	0.412	94.4	0.004	0.410	94.6
Balanced 8	α_0	0.001	0.114	96.1	-0.004	0.104	95.6	-0.002	0.103	96.1
	α_x	-0.002	0.173	94.3	0.005	0.141	94.6	0.002	0.131	94.9
	α_v	-0.002	0.107	94.9	-0.002	0.107	94.6	-0.002	0.106	94.7
	ψ	-0.002	0.411	94.7	0.001	0.412	94.4	0.003	0.410	94.4
Balanced 6	α_0	-0.004	0.109	95.1	-0.001	0.102	96.0	-0.003	0.102	96.1
	α_x	0.007	0.151	94.1	-0.001	0.134	94.7	0.002	0.129	95.7
	α_v	-0.003	0.106	94.6	-0.001	0.106	94.6	-0.003	0.107	94.5
	ψ	0.000	0.410	94.1	0.003	0.411	94.5	0.004	0.411	94.5

[†] $E[Y_1] = 0.4; E[Y_2] = 0.4; E[X] = 0.6; E[V] = 0.5; \alpha_x = \log 1.25; \alpha_v = \log 1.5; \beta_v = \log 1.5; \psi = 3.00$

4.5 Considerations of Studies Involving Greater Numbers of Stages

We return now to the problem involving a univariate response and binary covariates. We consider again the problem of selecting $n = 250$ individuals out of $N = 1000$ for measurement of the expensive covariate. Now, instead of examining only a two-stage adaptive selection procedure, we allow for adaptive sampling procedures that involve selection at 1, 2, 5, or 10 stages. The sampling procedure with only 1 stage of sampling is not adaptive. Here, the design using 2 stages differs slightly from those considered above as here n is divided evenly between the two stages of sampling ($n_a = n_b = n/2$); the design using 5 and 10 stages also select the same number of individuals at each stage ($0.20n$ and $0.10n$, respectively). After each stage of sampling, the completely observed individuals from all previous stages of selection are analysed in order to estimate the design parameters necessary for deriving the optimal design for the next stage of sampling. These designs were implemented for 1000 simulated datasets generated using the parameters discussed above. The difference between the sampling probabilities and the optimal sampling probabilities are displayed in Figure 4.9. The empirical biases and standard errors of the estimators, as well as the average estimated asymptotic standard errors and the empirical coverage probabilities resulting from the different designs are presented in Table 4.9.

There appears to be little to be gained here from greatly increasing the number of stages used in adaptive sampling. The designs using many stages of sampling did result in more efficient estimation than non-adaptive designs, but this efficiency was also gained by using relatively few stages of sampling. Again we see that the empirical efficiency of these adaptive designs was very similar to those for the optimal design; in fact, these adaptive designs were, by chance, sometimes more efficient in this finite sample situation than the asymptotically optimal design. There was also no evidence of bias here and the empirical coverage probabilities were compatible with the nominal level. Further work may be needed, however, to ensure that this use of internal pilot studies will not inflate the type-I error rate more generally (Wittes and Brittain, 1990).

4.6 Conclusion

This adaptive two-phase sampling design addresses the important problem of specifying parameter values at the design stage of studies. We have demonstrated situations in which the balanced design can be less efficient than simple random sampling or proportional stratified sampling, so it is not clear which of the non-adaptive designs would be preferable for any given study. However, the adaptive designs consistently improved the efficiency of the estimator of interest. These adaptive designs were beneficial regardless of which design was used at phase IIa; the efficiencies of estimators under adaptive sampling were generally very similar to those found under the optimal design. These adaptive procedures are particularly important for elaborate response models

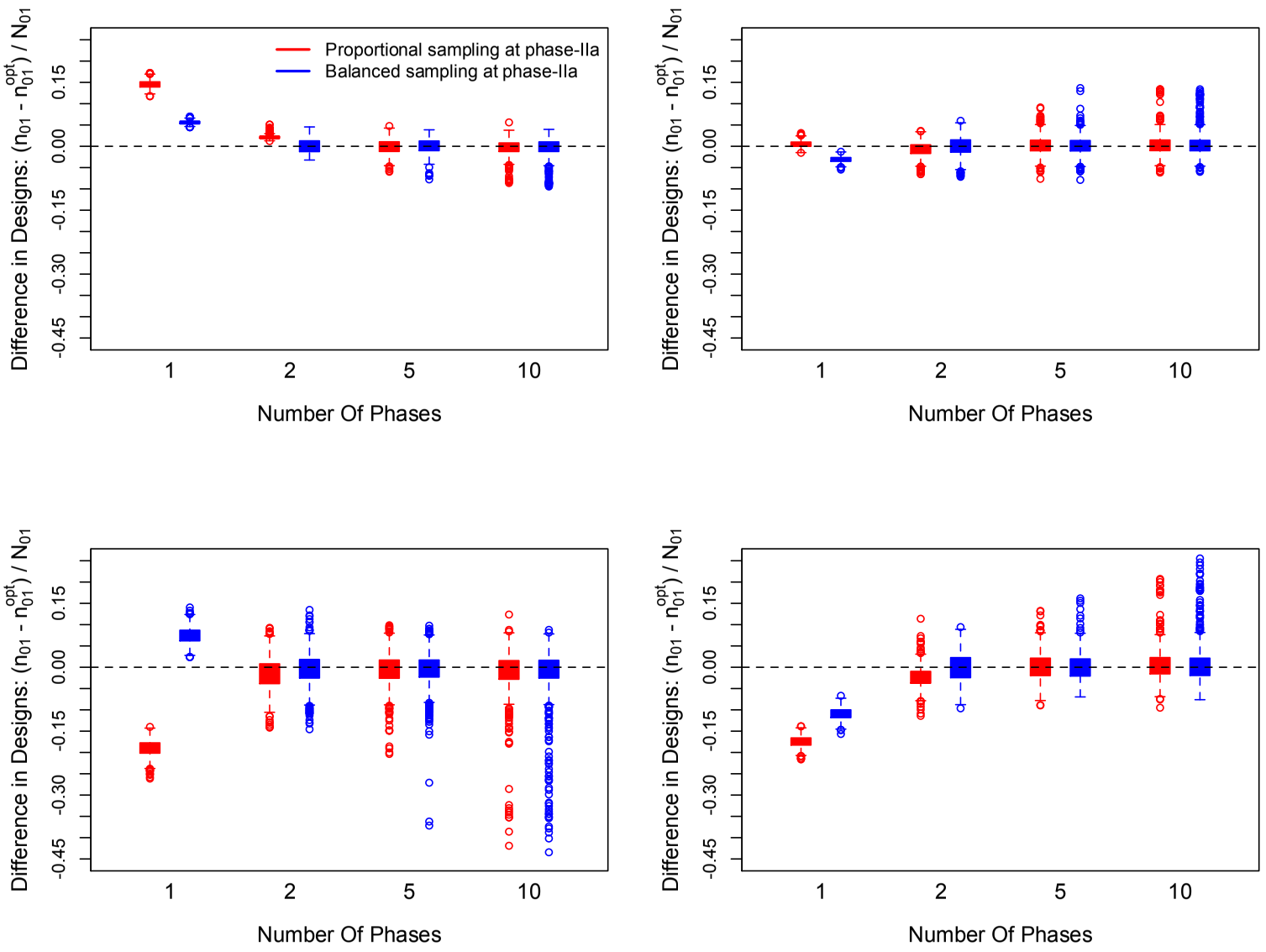

Figure 4.9: Boxplots displaying the differences in sampling fractions between the optimal and other designs and the optimal sampling fractions for each simulated dataset with univariate responses and a binary expensive covariate when $N = 1000$, $n = 250$ and $(\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_v, \gamma_0)' = (-1.95, 1.00, 0.90, 1.05, -0.41, -0.04)'$.

where many parameters must be specified in order to derive the optimal design; a simple example of this was the data involving clustered responses. Near-optimal designs were achieved without *a priori* knowledge of the parameters and without the necessity for costly external pilot studies. It was also demonstrated that adaptive designs using relatively small phase-IIa samples were generally quite efficient. There does appear to be a trade-off between the potential efficiency of a large phase-IIb sample and the decreased precision of estimates of the design parameter values when phase-IIa is small, however, the adaptive design that collected 20% of the individuals at phase IIa was generally quite effective at generating near-optimal sampling fractions.

We have focussed here on analyses using inverse probability weighted pseudo-likelihoods, however it was demonstrated in Chapter 3 that the optimal IPW design increased efficiency for a variety of methods of analysis. These adaptive designs have been shown to approximate the optimal IPW design, so the properties previously observed for the optimal design – improved efficiency for a variety of estimators, robustness to a misspecified covariate model – should also hold for this adaptive approach. It is of interest to explore the potential efficiency gains for more complex datasets such as those discussed in Chapter 5.

Table 4.9: Empirical biases, standard errors, and coverage probabilities for estimators based on 1000 simulated datasets when the overall phase-II sample size is $n = 250$ and phase-II sampling is carried out in multiple stages. The efficiency of an optimal design based on true (unknown) parameters[†] is presented for comparison.

		Number of Phases of Sampling											
		1			2			5			10		
Design		Bias	ESE	ECP	Bias	ESE	ECP	Bias	ESE	ECP	Bias	ESE	ECP
Optimal	α_0	-0.017	0.271	95.3	–	–	–	–	–	–	–	–	–
	α_x	0.024	0.298	95.3	–	–	–	–	–	–	–	–	–
	α_v	0.004	0.155	95.8	–	–	–	–	–	–	–	–	–
Proportional	α_0	-0.040	0.323	95.1	-0.007	0.265	94.6	0.001	0.272	94.6	-0.012	0.271	94.3
	α_x	0.045	0.357	95.8	0.010	0.298	95.1	-0.003	0.306	94.7	0.008	0.296	95.1
	α_v	0.006	0.151	96.5	0.001	0.151	96.1	0.005	0.154	95.4	0.013	0.156	95.2
Balanced	α_0	-0.027	0.283	95.3	-0.007	0.273	94.6	0.000	0.263	94.4	0.001	0.274	94.0
	α_x	0.038	0.316	95.4	0.011	0.310	93.3	0.002	0.295	94.9	-0.005	0.301	95.5
	α_v	0.004	0.153	96.0	0.003	0.154	95.3	0.002	0.154	96.0	0.005	0.159	94.6

[†] $(\alpha_0, \alpha_x, \alpha_v, \beta_0, \beta_v, \gamma_0)' = (-1.95, 1.00, 0.90, 1.05, -0.41, -0.04)'$.

Chapter 5

Future Work

5.1 Implications of Research Findings

Optimal designs for IPW analyses have been shown to result in increased efficiency compared to standard methods for two-phase sampling for a variety of settings and frameworks for analysis. Specifically, these designs resulted in efficiency gains over simple random sampling and balanced sampling designs under maximum likelihood, semi-parametric maximum likelihood, multiple imputation, inverse probability weighted pseudo-likelihoods, inverse probability weighted pseudo-likelihoods with estimated weights, and efficient augmented inverse probability weighted pseudo-likelihood methods of analysis. These efficiency gains were seen when all covariates were univariate and binary, when the expensive covariate was continuous, and when responses were clustered. Moreover, this optimal design resulted in efficient estimation of the parameter of interest for both variable probability and basic stratified sampling.

This optimal IPW design has also proven to be quite robust. Efficiency gains were seen for the estimator of interest when using misspecified optimal designs instead of balanced or simple random sampling. These efficiency gains were demonstrated when design parameters were misspecified and when there was misspecification of the nuisance covariate distribution required for derivation of the optimal design. This optimal design was also demonstrated to achieve high levels of efficiency when the design parameters were estimated using external pilot studies. The robustness of the optimal IPW design was also demonstrated when it was seen that design parameters estimated using moderately-sized external pilot studies were quite generally sufficient to improve the efficiency of the estimator of interest over both simple random and balanced sampling designs.

The two-stage adaptive Opt_{ipw} sampling scheme discussed in Chapter 4 was found to be quite effective for approximating the optimal IPW design in a variety of settings. Relatively small internal pilot studies were demonstrated to be sufficient for estimating the design parameters

necessary to derive the optimal IPW design. Furthermore, it was also demonstrated that balanced designs can be less efficient than simple random or proportional stratified sampling designs when responses are clustered.

When designing the second phase of a two-phase study, it is not clear which form of non-adaptive design would be most efficient. Simple random sampling, proportional stratified sampling, and balanced sampling designs were all demonstrated to be particularly inefficient in different settings. It has been shown, however, that the optimal design for IPW can be applied quite generally to improve the efficiency of the estimator of interest. These designs have also been seen to generally improve the efficiency of the estimator of the intercept term. If efficient estimation of other parameters is also of primary importance than the definition of optimality can be adjusted to incorporate this. By using a two- or multi-stage adaptive approach, the optimal IPW design can be approximated quite well without any *a priori* knowledge of parameter estimates and without the need to discard data selected in external pilot studies.

The decision of how to analyse the resultant incomplete data is not straight-forward. We have reviewed several methods of analysis designed for use with data under a MAR mechanism, and we compared their empirical efficiencies in a number of simulations. We further explored the limiting values of multiple imputation and IPW estimators implemented using misspecified models. When models were misspecified, neither of these approaches to analysis were universally preferable; in fact, neither of these methods of analysis were universally preferable to complete-case analysis. However, our results appear to support the assertions of Carpenter et al. (2006) that MI is more useful in the presence of incomplete responses, while IPW techniques are better suited to studies involving missing covariates. The doubly robust AIPW analysis may offer some benefit here as consistent estimators can still be achieved if either of the two specified models are correct. However, this method of analysis may perform poorly if both models are slightly misspecified (Kang and Schafer, 2007). Furthermore, we observed quite poor small sample properties for the AIPW estimator, although we also saw that this could be rectified through iterative estimation of the augmenting term, or by using the IPW_2 estimator to find the necessary preliminary estimate. A semiparametric efficient AIPW estimator is theoretically possible, but its derivation can computationally expensive and it is rarely, if ever, implemented in practice (Carpenter et al., 2006). Other potentially efficient methods of analysis for two-phase data include the pseudo-score estimator (Chatterjee et al., 2003) and conditional maximum likelihood (Lawless et al., 1999; Scott and Wild, 2011b). Although the results were not presented in this thesis, the optimal design for IPW analysis was also shown to increase the efficiency of the pseudo-score estimator of the parameter of interest.

Whatever the chosen method of analysis, the optimal IPW design could be implemented through the adaptive sampling approach. Calculation of these optimal stratum-specific sampling probabilities is relatively straight-forward, and the resultant design was generally quite efficient and robust. This design could be quite useful for improving precision of estimates without increasing study costs in a wide variety of settings, including those of the PsA and CLSA studies that served

as our motivation.

The contents of this thesis have so far resulted in the preparation of four manuscripts. McIsaac and Cook (2013a) focuses on missing data in the context of clinical trials; commonly used methods for analysing incomplete data are summarised and critically assessed based on both their asymptotic and empirical properties. McIsaac et al. (2013) examines causes and effects of missing data in dermatology trials and provides practical recommendations for planning and drawing conclusions from studies which could involve incomplete data. McIsaac and Cook (2013c) examines the potential efficiency gains of optimal two-phase sampling designs in the context of clustered and longitudinal data. Finally, McIsaac and Cook (2013b) critically examines the utility of implementing optimal two-phase designs when parameter values are not known *a priori*; in particular, the Opt_{ipw} design is shown to work well in practice. Immediate next steps involve the study of two-phase design in the context of longitudinal binary data with time-varying exposure variables, as described in the next section.

5.2 Ongoing and Future Research Plans

In this section we highlight several areas of interest for continuing research.

5.2.1 Repeated Measure Analyses of Longitudinal Data

Consider the analysis of binary data arising from a longitudinal study. Suppose that measurements can be taken at K prespecified time points $t_1,\ldots,t_K$ so that the complete data for individual i consists of $Y_i = (Y_{i1},\ldots,Y_{iK})'$, $X_i = (X_{i1},\ldots,X_{iK})'$, and $V_i = (V_{i1},\ldots,V_{iK})'$. However the expensive covariate X may not be observed for individual i at all time points. We let R_{ij} be the indicator that X_{ij} is observed. Then $R_i = (R_{i1},\ldots,R_{iK})'$ is the vector of indicators of subsample selection for individual i. We will assume that all individuals have complete information at the first time point so that $R_{i1} = 1$, for $i = 1,\ldots N$. We will indicate the history of a vector $Z_i = (Z_{i1},\ldots,Z_{iK})'$ up to a certain time point k with an overbar and a subscript k, so that $\overline{Z}_{ik} = (z_{i1},\ldots,z_{i,k-1})'$. We denote the probability of being selected for measurement of the expensive covariate by $\pi_{ik} = P(R_{ik} = 1|y_i,x_i,v_i)$.

The analysis in this section is motivated by a modification of the CLSA design problems and could have applications in other studies, including those at the PsA clinic. Suppose that it is of interest to determine at each time point of an ongoing longitudinal study which individuals should be selected for in-depth study. The sampling decision can then be based only on the observed history. That is, the researcher can choose whether to select in an individual for measurement of the expensive covariate based on past responses, past realisations of the auxiliary covariate, and those past values of the expensive covariate which were observed. This differs from the problem

considered in Section 3.8 in that the responses are repeated measures and covariates here are time-dependent.

We can consider marginal analysis and use generalised estimating equations of the form

$$U_{\mathscr{V}}(\alpha,\eta)=\sum_{i=1}^{N}D_i'(\alpha)\mathscr{V}_i^{-1}(\eta)(Y_i-\mu_i(\alpha))=0,$$

for these longitudinal data, where $\mathscr{V}_i(\eta)=A_i^{1/2}R(\eta)A_i^{1/2}$, and η can be replaced by $\widehat{\eta}(\alpha)$, a $\sqrt{N}$-consistent estimator of η given α. In order to obtain consistent estimators of α in the presence of covariates which are MAR, we must introduce a weighting to the estimating function. Since we are discussing a prospective study where data are missing by design, the probability of selection can only depend on data which are available to the experimenter at the time of selection. That is, if we consider the values of covariate X observed prior to time k in $\overline{x}_{ik}^r=(x_{i1}^r,\ldots,x_{i,k-1}^r)'$ where $x_{ij}^r=(R_{ij},R_{ij}x_{ij})'$, then

$$P(R_{ik}=1|y_i,x_i,v_i,\overline{R}_{ik})=P(R_{ik}=1|y_{ik},v_{ik},\overline{y}_{ik},\overline{v}_{ik},\overline{x}_{ik}^r),\tag{5.1}$$

where, as in previous chapters, the two-phase design means that cheap covariates y_{ik} and v_{ik} are available at time k for use in deciding whether to sample x_{ik} (alternatively, one could consider the situation where only the covariates $\overline{y}_{ik},\overline{v}_{ik}$, and $\overline{x}_{ik}^r$ are available at time k).

Now, let $H_{ik}=(v_{ik},y_{ik},\overline{v}_{ik}',\overline{y}_{ik}',\overline{x}_{ik}^{r\prime})'$, so H_{ik} is the observed history for individual i up to time k. Given the designed nature of our missing data we will assume that

$$\lambda_{ik}\equiv\lambda_{ik}(\delta)=P(R_{ik}=1|y_i,v_i,x_i;\delta)=P(R_{ik}=1|H_{ik};\delta),\tag{5.2}$$

is a known function of the parameter δ, which has been selected by the experimenter. That is, the probability of observing individual i at time k is a known function (specified by the experimenter) which depends only on the data available for individual i at time k. Then the conditional probability of the selection process for individual i given the data (y_i,x_i,v_i) is

$$\pi_i\equiv\pi_i(\delta)=P(R_i=r_i|y_i,x_i,v_i)=\prod_{k=1}^{K}\lambda_{ik}^{r_{ik}}(1-\lambda_{ik})^{1-r_{ik}}.\tag{5.3}$$

We can then use the inverse probability weighted generalised estimating equation

$$U_1(\alpha,\widehat{\eta}(\alpha),\delta)=\sum_{i=1}^{N}D_i'(\alpha)\mathscr{V}_i^{-1}(\widehat{\eta}(\alpha))\Delta_i(\delta)\cdot(Y_i-\mu_i(\alpha))=0,\tag{5.4}$$

where, as suggested by Fitzmaurice et al. (1995) we use the cluster-level weights
$\Delta_i(\delta)=diag\{r_{i1}I(R_i=r_i)\ \pi_i^{-1},r_{i2}I(R_i=r_i)\ \pi_i^{-1},\cdots,r_{iK}I(R_i=r_i)\ \pi_i^{-1}\}$.

For any given vector of observations r_i, estimating equation (5.4) is unbiased and results in a $\sqrt{N}$-consistent estimator for α since

$$E\left[U_1(\alpha,\widehat{\eta}(\alpha),\delta)\right]=\sum_{i=1}^{N}E_{Y,X,V}\left[D_i'\mathscr{V}_i^{-1}E_{R|Y,X,V}\left\{\Delta_i\right\}\cdot(Y_i-\mu_i)\right]=0$$

when $E_{R|Y,X,V}\left\{\Delta_i\right\}$ does not depend on Y_i, and the model for μ_i is properly specified.

Under the regulatory conditions discussed in Robins et al. (1995) and assuming that η_0 is the true value for η, Taylor's theorem gives
$0=U_1(\widetilde{\alpha},\eta_0,\delta)=U_1(\alpha,\eta_0,\delta)+\frac{\partial}{\partial\alpha'}U_1(\alpha,\eta_0,\delta)(\widetilde{\alpha}-\alpha)+o_p(N^{-\frac{1}{2}})$, so
$(\widetilde{\alpha}-\alpha)=I^{-1}(\alpha)U_1(\alpha,\eta_0,\delta)+o_p(N^{-\frac{1}{2}})$, where $I(\alpha)=-\frac{\partial}{\partial\alpha'}U_1(\alpha,\eta_0,\delta)$. Thus,
$\sqrt{N}(\widetilde{\alpha}-\alpha)=\left[\frac{1}{N}I(\alpha)\right]^{-1}\left[\frac{1}{\sqrt{N}}U_1(\alpha,\eta_0,\delta)\right]+o_p(1)$. As $N\to\infty$,
$\frac{1}{N}I(\alpha)=-\frac{1}{N}\frac{\partial}{\partial\alpha'}\sum_{i=1}^{N}U_{1i}(\alpha,\eta_0,\delta)=-\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial}{\partial\alpha'}U_{1i}(\alpha,\eta_0,\delta)\right]\xrightarrow{p}-E\left[\frac{\partial}{\partial\alpha'}U_{1i}(\alpha,\eta_0,\delta)\right]$, and
$\frac{1}{\sqrt{N}}U_1(\alpha,\eta_0,\delta)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}U_{1i}(\alpha,\eta_0,\delta)\xrightarrow{D}MVN\left(\mathbf{0},\mathscr{I}_\alpha\right)$, where
$\mathscr{I}_\alpha=E\left[U_{1i}(\alpha,\eta_0,\delta)U_{1i}'(\alpha,\eta_0,\delta)\right]$. So, $\sqrt{N}(\widetilde{\alpha}-\alpha)\xrightarrow{D}MVN\left(0,\Gamma_\alpha^{-1}\mathscr{I}_\alpha[\Gamma_\alpha^{-1}]'\right)$ where $\Gamma_\alpha=E\left[\frac{\partial}{\partial\alpha'}U_{1i}(\alpha,\eta_0,\delta)\right]$. So, the estimator $\widetilde{\alpha}$ of α is consistent and $\sqrt{N}(\widetilde{\alpha}-\alpha)$ has asymptotic variance

$$\Gamma_\alpha^{-1}\mathscr{I}_\alpha[\Gamma_\alpha^{-1}]' \tag{5.5}$$

when η_0 is known.

If primary interest lies in modelling the association amongst the responses as well as the marginal means, then we may wish to proceed with a GEE2 type procedure where it is not assumed that the parameters α and η are orthogonal. As in Liang et al. (1992), consider $W_i=(Y_{i1}Y_{i2},Y_{i1}Y_{i3},\ldots,Y_{i,K-1}Y_{iK})'$ and $\omega_i=E[W_i|X_i,V_i]$. Then IPWGEE2 amounts to utilising the estimating equation for $\Omega=(\alpha',\eta')'$:

$$U_2(\alpha,\eta,\delta)=\sum_{i=1}^{N}U_{2i}(\alpha,\eta,\delta)=\sum_{i=1}^{N}\frac{\partial(\mu_i',\omega_i')}{\partial\Omega}\cdot cov^{-1}\begin{pmatrix}Y_i\\W_i\end{pmatrix}\cdot\Delta_i^*\cdot\begin{pmatrix}Y_i-\mu_i\\W_i-\omega_i\end{pmatrix},$$

where $\Delta_i^*\equiv\Delta_i^*(\delta)$ is a matrix of weights such that $E_{R|Y,X,V}\left\{\Delta_i^*\right\}$ does not depend on Y_i. For example, with subject-specific (cluster-level) weights we could have

$$\Delta_i^*(\delta)=diag\{r_{i1},\ldots,r_{iK},r_{i1}r_{i2},\ldots,r_{i,K-1}r_{iK}\}I(R_i=r_i)\pi_i^{-1}(\delta), \tag{5.6}$$

where $\pi_i(\delta)$ is defined as in (5.3). By similar arguments to those used above, it can be shown that with properly specified models for marginal means (μ) and marginal pairwise associations (ω), $\sqrt{N}\left(\widehat{\Omega}-\Omega\right)=\sqrt{N}\begin{pmatrix}\widetilde{\alpha}-\alpha\\\widehat{\eta}-\eta\end{pmatrix}\xrightarrow{D}N\left(0,\Gamma_\Omega^{-1}\mathscr{I}_\Omega[\Gamma_\Omega^{-1}]'\right)$, where
$\mathscr{I}_\Omega=E\left[U_{2i}(\alpha,\eta,\delta)U_{2i}'(\alpha,\eta,\delta)\right]$ and $\Gamma_\Omega=E\left[\frac{\partial}{\partial\Omega'}U_{2i}(\alpha,\eta,\delta)\right]$.

In this setting, we can explore optimal designs which result in an efficient estimator of α. One interesting feature of utilising these longitudinal data to explore marginal features is that the parameter estimates can be updated at each time point. Thus instead of requiring a pilot study to find estimates to use in developing our design, we could begin with a simple design and revise to more optimal sampling designs as our knowledge of the parameters increases. This is a natural extension of the adaptive samples considered in Chapter 4.

5.2.2 Current Status Failure Time Data

It is also interesting to explore the selection of individuals for biomarker testing based on current status data of progression. As shown in Figure 5.1, the response is an indicator of damage occurring before clinic entry. The auxiliary variable will be the time from disease diagnosis to clinic entry. These inexpensive variables will be incorporated in a selection model whose parameters will be optimally tuned to obtain the most informative sample possible for estimating the parameter of interest. This extension will be interesting as it seems reasonable to assume that the selection procedure will depend not only on the response, but also on the size of the censoring interval. For example, with current status data, where there is only one clinic visit (the enrolment visit), it seems reasonable to assume that failures occurring within large censoring intervals will not be as informative as failures that occur in short intervals since in the latter case we would have more precise information about when the event actually occurred. An estimate of the distribution of time from disease diagnosis to first damaged joint in PsA patients based on the current status information at clinic entry is shown in Figure 5.2.

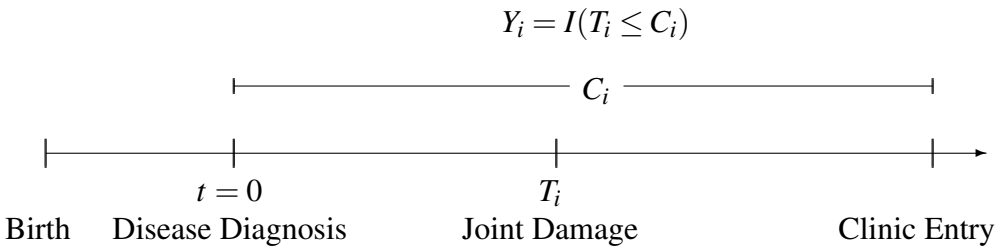

Figure 5.1: Timeline diagram for current status data on disease progression

Suppose that the responses Y_i and censoring times C_i are known for all individuals $i = 1, \ldots, N$ in the first phase of a two-phase study involving current status data. In the second phase, the covariates X_i are to be ascertained for a subset of the individuals. We once again will utilise the notation that R_i is an indicator of X_i being observed, and we will attempt to find the optimal parameters δ indexing $P(R_i = 1|y_i, c_i; \delta)$ which maximise the precision with which the parameter of interest, α_x, can be estimated.

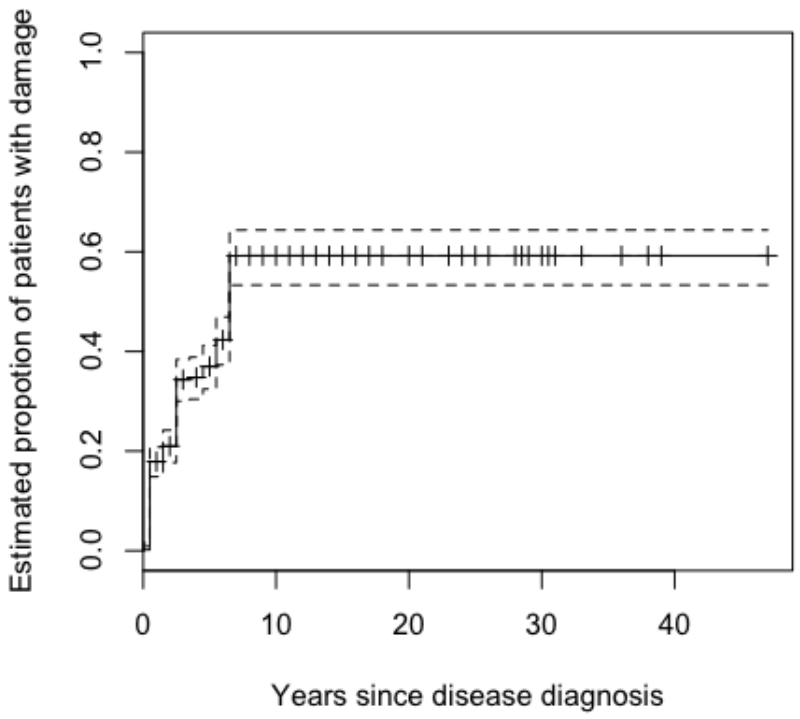

Figure 5.2: Kaplan-Meier estimate of the distribution of the time from disease diagnosis to first damaged joint for clinic patients with stored sera based on current status data at clinic entry.

We can consider efficient designs under two main methods of analyses that are analogous to the methods considered for cross-sectional data: a fully-specified likelihood approach, and an inverse-weighting approach similar to that proposed by Li and Nan (2011). We could consider the optimal selection strategies in both these circumstances and explore how these optimal designs work in practice; here optimal designs could be sensitive to misspecification of the nuisance baseline hazard function.

LIKELIHOOD WITH COMPLETE COVARIATES

Under the proportional hazards assumption $h(t|x) = h_0(t)e^{\alpha_x x}$ with complete covariates, the likelihood for current status data is

$$\begin{aligned} L(\alpha_x, H_0; x, c) &= \prod_{i=1}^{N} L_i(\alpha_x, H_0; x_i, c_i) \\ &= \prod_{i=1}^{N} [\exp\{-e^{\alpha_x x_i} H_0(c_i)\}]^{1-y_i} [1 - \exp\{-e^{\alpha_x x_i} H_0(c_i)\}]^{y_i} \end{aligned} \tag{5.7}$$

(Sun, 2006). Assuming a Weibull model, analysis of the the binary response $Y_i = I(T_i < c_i)$ can be carried out using standard statistical software for binomial data under a complementary log-log link.

LIKELIHOOD APPROACH WITH INCOMPLETE COVARIATES

When we do not have complete covariate information, Wen and Lin (2010) suggest the use of the likelihood

$$L_F(\alpha,\beta) = \prod_{i=1}^{N} \int_{x_i \in \mathscr{X}_i} L_i(\alpha; x_i, c_i) g(x_i;\beta)\, dx_i, \tag{5.8}$$

where $g(\cdot)$ is the density function associated with X, $\mathscr{X}_i$ is the set of possible values of x_i, and maximisation can be implemented through the use of an EM algorithm. Then the ML estimator of θ will satisfy

$U_F(\widehat{\theta}) = \partial \ell_F(\widehat{\theta})/\partial\theta \equiv \partial \ell_F(\theta)/\partial\theta \Big|_{\theta=\widehat{\theta}} = 0$, where

$$\begin{aligned} U_F(\theta) &= \sum_{i=1}^{N} U_{Fi}(\theta) \\ &= \sum_{i=1}^{N} \begin{pmatrix} R_i\, \partial \log L_i(\alpha; x_i, c_i)/\partial\alpha + (1-R_i)\, \partial \log\left(\int_x L_i(\alpha; x, c_i) g(x_i;\beta) dx\right)/\partial\alpha \\ R_i\, \partial \log g(x_i;\beta)/\partial\beta + (1-R_i)\, \partial \log\left[\int_x L_i(\alpha; x, c_i) g(x_i;\beta) dx\right]/\partial\beta \end{pmatrix} \end{aligned}$$

is the derivative of the logarithm of (5.8) with respect to θ.

Therefore, $\sqrt{N}(\widehat{\theta} - \theta) \xrightarrow{D} \text{MVN}\left(0, \mathscr{I}_\theta^{-1}\right)$, where

$$\mathscr{I}_\theta = E\left[U_{Fi}(\theta) U_{Fi}'(\theta)\right]. \tag{5.9}$$

The expectation in (5.9) will be taken over the R_i, so the asymptotic variance of the resultant estimator will be a function of δ.

INVERSE WEIGHTING APPROACH

With a Weibull hazard, the inverse weighting approach involves maximising the weighted complete-case log-likelihood

$$\ell_W(\alpha,\delta) = \sum_{i:R_i=1} \log L_i(\alpha; x_i, c_i) \times P(R_i = 1 | y_i, c_i; \delta)^{-1}. \tag{5.10}$$

If we let $U_{Wi}(\alpha,\delta)$ be the score function associated with (5.10), then the derivation of the asymptotic variance of the estimator resulting from the inverse weighting approach can be carried out

as before, and $\sqrt{N}(\widehat{\alpha}-\alpha) \xrightarrow{D} \text{MVN}\left(0, \Gamma_\alpha^{-1}\mathscr{I}_\alpha[\Gamma_\alpha^{-1}]'\right)$, where now

$$\Gamma_\alpha = E\big[\frac{\partial}{\partial \alpha'}U_{Wi}(\alpha,\delta)\big] \text{ and } \mathscr{I}_\alpha = E\big[U_{Wi}(\alpha,\delta)U'_{Wi}(\alpha,\delta)\big].$$

Once again, the resulting asymptotic variance will depend on δ, which is associated with the sampling design.

We can consider different classes of selection models incorporating continuous and discretised censoring time values and explore the relative efficiency of optimal designs. We can further explore the robustness of these designs to misspecification of design models and parameters. It will also be interesting to explore the relative efficiency of such designs when analysis is carried out using a semi-parametric approach which does not assume a Weibull baseline hazard function.

5.2.3 Dealing with Data Missing by Design and Chance

Another area of research interest deals with data which are partially incomplete by happenstance and partially incomplete by design in a two-phase study. If data are complete at the first phase of sampling, then having two types of missingness changes the true probabilities of being observed so that, using the notation of this thesis,

$$\pi_i = P(R_i = 1|Y_i,V_i) = P(R_i^{(d)} = 1|Y_i,V_i)\cdot P(R_i^{(h)} = 1|Y_i,V_i,R_i^{(d)} = 1),$$

where $R_i^{(d)}$ is the indicator that an individual is selected by the design for measurement at the second phase of sampling, and $R_i^{(h)}$ is the indicator that the observation is not missing by happenstance. In terms of analysis, this missingness by happenstance at phase-II will change little as long as the missing data mechanism is MAR. If analysis is to be carried out using likelihood methods (e.g. ML, SPML, or MI), then the model for missingness need not be specified; only the MAR assumption is required. If analysis is carried out using weighted estimating equations, then the true missingness model need not be known as more efficient estimators can be found by using estimated selection probabilities, which can easily be found for discrete phase-I data as $\sum_i I(R_i,Y_i = y,V_i = v)/\sum_i I(Y_i = y,V_i = v)$. Here, consistency is only guaranteed if both missing data mechanisms are MAR and if $P(R_i^{(d)} = 1|Y_i,V_i)\cdot P(R_i^{(h)} = 1|Y_i,V_i,R_i^{(d)} = 1)$ is bounded away from zero. Optimal designs are somewhat complicated as the optimal selection probabilities should be inflated by a factor of $P(R_i^{(h)} = 1|Y_i,V_i,R_i^{(d)} = 1)^{-1}$ since we are optimising for π_i and not $P(R_i^{(d)} = 1|Y_i,V_i)$. There is little that can be done at the design stage to estimate the probability of phase-II missingness, however an internal pilot study could now serve the dual purpose of estimating the design parameters and estimating $P(R_i^{(h)} = 1|Y_i,V_i,R_i^{(d)} = 1)$. It would be interesting to examine the efficiency of such a design.

If phase-I data are incomplete, then the situation is potentially more complicated. As before, knowledge of the true parameters allows for derivation of the optimal complete-observation probabilities π_i, only now

$$\pi_i = P(R_i = 1|Y_i, V_i) = P(R_i^{(h)} = 1|Y_i, V_i) \cdot P(R_i^{(d)} = 1|Y_i, V_i, R_i^{(h)} = 1),$$

since individuals can only be completely observed if their phase-I values are observed. In this setting, a small validation sample could be selected for observation of the missing phase-I data, which would allow for verification of the MAR assumption and allow for estimation of $P(R_i^{(h)} = 1|Y_i, V_i)$, which when combined with the optimal π_i would allow for estimation of the optimal phase-II selection probabilities among those available for study $P(R_i^{(d)} = 1|Y_i, V_i, R_i^{(h)} = 1) = \pi_i \cdot P(R_i^{(h)} = 1|Y_i, V_i)^{-1}$. This validation sample could be combined with the idea of the adaptive designs, so that the internal pilot study consists of actively seeking out individuals with missing phase-I data for complete observation of Y, X, V; these data would provide enough information to approximate an optimal sampling design for the remainder of the phase-II sample. There is likely to be a cost difference between complete observation of an individual who was originally missing and phase-II observation of a previously-observed individual. There will be an interesting decisions to be made about how many incomplete phase-I individuals to pursue. It may again be of interest to inflate these selection probabilities by an estimate of the stratum-specific probability of non-response.

These data fall under the framework of the three-stage case-control studies (Scott and Wild, 2011b) and multi-phase designs (Lee et al., 2010). Efficient estimators based on conditional maximum likelihood have been discussed by Scott and Wild (2011b) in this setting. However, further efficiency gains would be possible here by implementing optimal or near-optimal selection procedures. Easily-calculated optimal designs for IPW analyses may again offer a robust and efficient method for selecting individuals at each of the sampling phases regardless of the chosen method of analysis. Adaptive designs such as those discussed in Chapter 4 will be useful for approximating these designs without requiring *a priori* specification of design parameters.

Bibliography

P. S. Albert and D. Follmann. Shared-parameter models. In G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, editors, *Longitudinal Data Analysis*, pages 433–452. CRC Press, Boca Raton, FL, 2009.

M. Alirezai, S. A. George, I. Coutts, D. I. Roseeuw, J. P. Hachem, N. Kerrouche, F. Sidou, P. Soto, et al. Daily treatment with adapalene gel 0. 1% maintains initial improvement of acne vulgaris previously treated with oral lymecycline. *European Journal of Dermatology*, 17(1): 45–51, 2007.

P. D. Allison. *Missing Data*. Number no. 136 in Quantitative Applications in the Social Sciences. SAGE Publications, 2001. ISBN 9780761916727.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

J. Barnard and D. B. Rubin. Miscellanea. small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955, 1999.

C. Beunckens, G. Molenberghs, and M. G. Kenward. Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials*, 2(5):379–386, 2005.

N. E. Breslow and K. C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20, 1988.

N. E. Breslow and N. Chatterjee. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Applied Statistics*, 48(4):457–468, 1999.

N. E. Breslow and R. Holubkov. Weighted likelihood, psuedo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine*, 16(1): 103–116, 1997.

J. R. Carpenter, M. G. Kenward, and S. Vansteelandt. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society A*, 169(3):571–584, 2006.

V. Chandran, R. J. Cook, J. Edwin, H. Shen, F. J. Pellett, S. Shanmugarajah, C. F. Rosen, and D. D. Gladman. Soluble biomarkers differentiate patients with psoriatic arthritis from those with psoriasis without arthritis. *Rheumatology*, 49(7):1399–1405, 2010a.

V. Chandran, D. C. Tolusso, R. J. Cook, and D. D. Gladman. Risk factors for axial inflammatory arthritis in patients with psoriatic arthritis. *The Journal of Rheumatology*, 37(4):809–815, 2010b.

N. Chatterjee, Y. Chen, and N. E. Breslow. A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168, Mar 2003.

B. Chen and R. J. Cook. Strategies for bias reduction in estimation of marginal means with data missing at random. In P. Pardalos and T. F. Coleman amd P. Xanthopoulos, editors, *Optimization and Data Analysis on Biomedical Informatics*. American Mathematics Society, 2012.

B. Chen, G. Y. Yi, and R. J. Cook. Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Associaton*, 105:336–353, 2010.

D. Clayton, D. Spielgelhalter, G. Dunn, and A. Pickles. Analysis of longitudinal binary data from multi-phase sampling. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 60(1):71–87, 1998.

Canadian Longitudinal Study On Aging (CLSA). Etude longitudinale canadienne sur le vieillissement (elcv). `http://www.clsa-elcv.ca/en/welcome/index.html`, 2009.

R. J. Cook, L. Zeng, and G. Y. Yi. Marginal analysis of incomplete longitudinal binary data: A cautionary note on locf imputation. *Biometrics*, 60:820–828, September 2004.

D. R. Cox. The analysis of multivariate binary data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 21(2):113–120, 1972.

M. Crowder. On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika*, 82:407–410, 1995.

S. Demissie, M. P. LaValley, N. J. Horton, R. J. Glynn, and L. A. Cupples. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Statistics In Medicine*, 22(4):545–557, 2003.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38, 1977.

L. Dubertret, W. Sterry, J. D. Bos, S. Chimenti, S. Shumack, C. G. Larsen, N. H. Shear, and K. A. Papp. Clinical experience acquired with the efalizumab (raptiva®)(clear) trial in patients with moderate-to-severe plaque psoriasis: Results from a phase iii international randomized, placebo-controlled trial. *British Journal Of Dermatology*, 155(1):170–181, 2006.

AF Emery and A. V. Nenarokomov. Optimal experiment design. *Measurement Science and Technology*, 9(6):864, 1999.

G. M. Fitzmaurice, G. Molenberghs, and S. R. Lipsitz. Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(4):691–704, 1995.

D. D. Gladman and V. Chandran. Observation cohort studies: Lessons learnt from the university of toronto psoriatic arthritis program. *Rheumatology*, 50:25–31, 2011.

R. J. Glynn, N. M. Laird, and D. B. Rubin. Multiple imputation in mixture models for nonignorable nonresponse with followups. *Journal of the American Statistical Association*, pages 984–993, 1993.

K. B. Gordon, R. G. Langley, C. Leonardi, D. Toth, M. A. Menter, S. Kang, M. Hefferman, B. Miller, R. Hamlin, L. Lim, J. Zhong, R. Hoffman, and M. M. Okun. Clinical response to adalimumab treatment in patients with moderate to severe psoriasis: Double-blind, randomized controlled trial and open-label extension study. *Journal of the American Academy of Dermatology*, 55:598–606, 2006a.

K. B. Gordon, R. G. Langley, C. Leonardi, D. Toth, M. A. Menter, S. Kang, M. Heffernan, B. Miller, R. Hamlin, L. Lim, et al. Clinical response to adalimumab treatment in patients with moderate to severe psoriasis: Double-blind, randomized controlled trial and open-label extension study. *Journal of the American Academy of Dermatology*, 55(4):598–606, 2006b.

K. B. Gordon, R. G. Langley, A. B. Gottlieb, K. A. Papp, G. G. Krueger, B. E. Strober, D. A. Williams, Y. Gu, and J. M. Valdes. A phase iii, randomized, controlled trial of the fully human il-12/23 mab briakinumab in moderate-to-severe psoriasis. *Journal of Investigative Dermatology*, 2011.

A. B. Gottlieb, R. Evans, S. Li, L. T. Dooley, C. A. Guzzo, D. Baker, M. Bala, C. W. Marano, and A. Menter. Infliximab induction therapy for patients with severe plaque-type psoriasis: A randomized, double-blind, placebo-controlled trial. *Journal of the American Academy of Dermatology*, 51(4):534–542, 2004.

C. E. M. Griffiths, B. E. Strober, P. van de Kerkhof, V. Ho, R. Fidelus-Gort, N. Yeilding, C. Guzzo, Y. Xia, B. Zhou, S. Li, et al. Comparison of ustekinumab and etanercept for moderate-to-severe psoriasis. *New England Journal of Medicine*, 362(2):118–128, 2010.

G. Grimmett and D. Stirzaker. *Probability and Random Processes, 3rd Edition*. Oxford University Press Inc, New York, 2007.

P. J. Heagerty. Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, 58(2):342–351, 2002.

P. J. Heagerty and S. L. Zeger. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, 15(1):1–26, 2000.

T. Herzog and D. B. Rubin. Using multiple imputations to handle nonresponse in sample surveys. In W. G. Madow, I. Olkin, and D. B. Rubin, editors, *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliography*. Academic Press, New York, 1983.

J. W. Hogan, J. Roy, and C. Korkontzelou. Handling drop-out in longitudinal studies. *Statistics In Medicine*, 23(9):1455–1497, 2004.

N. J. Horton and N. M. Laird. Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics*, 57:34–42, 2001.

L. Joseph, P. Bélisle, H. Tamim, and J. S. Sampalis. Selection bias found in interpreting analyses with missing data for the prehospital index for trauma. *Journal Of Clinical Epidemiology*, 57 (2):147–153, 2004.

J. D. Y. Kang and J. L. Schafer. Demystifying double robustness. *Statistical Science*, 22(4): 523–539, 2007.

M. J. Knol, K. J. M. Janssen, A. R. T. Donders, A. C. G. Egberts, E. R. Heerdink, D. E. Grobbee, K. G. M. Moons, and M. I. Geerlings. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: An empirical example. *Journal Of Clinical Epidemiology*, 63(7):728–736, 2010.

G. G. Krueger, K. A. Papp, D. B. Stough, and et al. A randomized, double-blind, placebo-controlled phase iii study evaluating efficacy and tolerability of 2 courses of alefacept in patients with chronic plaque psoriasis. *Journal of the American Academy of Dermatology*, 47: 821–833, 2002.

M. Kulich and D. Y. Lin. Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*, 99(467):832–844, 2004.

R. G. B. Langley, G. G. Krueger, and C. E. M. Griffiths. Psoriasis: Epidemiology, clinical features, and quality of life. *Annals Of The Rheumatic Diseases*, 64(suppl 2):ii18–ii23, 2005.

A. Laupacis, D. L. Sackett, and R. S. Roberts. An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine*, 318:1728–1733, 1998.

J. F. Lawless. Likelihood and pseudo likelihood estimation based on response-biased observations. In I. S. Basawa, V. P. Godambe, and R. L. Taylor, editors, *Studies in Item Analysis and Prediction*, pages 43–55. Institute of Mathematical Statistics, 1997.

J. F. Lawless and J. D. Kalbfleisch. Discussions. *International Statistical Review*, 79(2):225–228, 2011.

J. F. Lawless, J. D. Kalbfleisch, and C. J. Wild. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 61(2):413–438, 1999.

M. Lebwohl, E. Christophers, R. Langley, J. P. Ortonne, J. Roberts, and C. E. M. Griffiths. An international, randomized, double-blind, placebo-controlled phase 3 trial of intramuscular alefacept in patients with chronic plaque psoriasis. *Archives Of Dermatology*, 139(6):719, 2003.

A. J. Lee, A. J. Scott, and C. J. Wild. Efficient estimation in multi-phase case-control studies. *Biometrika*, 97(2):361–374, 2010.

C. Leonardi, R. G. Langley, K. Papp, S. K. Tyring, N. Wasel, R. Vender, K. Unnebrink, S. R. Gupta, W. C. Valdecantos, and J. Bagel. Adalimumab for treatment of moderate to severe chronic plaque psoriasis of the hands and feet: Efficacy and safety results from reach, a randomized, placebo-controlled, double-blind trial. *Archives of Dermatology*, 147(4):429, 2011.

C. L. Leonardi, J. L. Powers, R. T. Matheson, B. S. Goffe, R. Zitnik, A. Wang, and A. B. Gottlieb. Etanercept as monotherapy in patients with psoriasis. *New England Journal of Medicine*, 349 (21):2014–2022, 2003.

C. L. Leonardi, K. A. Papp, K. B. Gordon, A. Menter, S. R. Feldman, I. Caro, P. A. Walicke, P. G. Compton, A. B. Gottlieb, et al. Extended efalizumab therapy improves chronic plaque psoriasis: Results from a randomized phase iii trial. *Journal of the American Academy of Dermatology*, 52(3 Pt 1):425, 2005.

C. L. Leonardi, A. B. Kimball, K. A. Papp, N. Yeilding, C. Guzzo, Y. Wang, S. Li, L. T. Dooley, K. B. Gordon, et al. Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with psoriasis: 76-week results from a randomised, double-blind, placebo-controlled trial (phoenix 1). *Lancet*, 371(9625):1665, 2008.

Z. Li and B. Nan. Relative risk regression for current status data in case-cohort studies. *Canadian Journal of Statistics*, 39(4):557–577, 2011.

K. Liang, S. L. Zeger, and B. Qaqish. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1):3–40, 1992.

K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika Trust*, 73(1):13–22, April 1986.

S. R. Lipsitz, N. M. Laird, and D. P. Harrington. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika*, 78(1):153–160, March 1991.

S. R. Lipsitz, M. Parzen, and L. P. Zhao. A degrees-of-freedom approximation in multiple imputation. *Journal of Statistical Computation and Simulation*, 72(4):309–318, 2002.

R. J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Associaton*, 88:125–134, 1993.

R. J. A. Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Associaton*, 90:1112–1121, 1995.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data Second Edition*. John Wiley & Sons, New York, 2002.

S. L. Lohr. *Sampling: Design And Analysis*. Thomson, 2009.

S. L. Lohr. Accurate multivariate estimation using triple sampling. *The Annals of Statistics*, pages 1615–1633, 1990.

T. Lumley, P. A. Shaw, and J. Y. Dai. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79(2):200–220, 2011.

D. E. Matthews and V. T. Farewell. *Using and Understanding Medical Statistics, 3rd Revised Edition*. Karger, Basel, Switzerland, 1996.

B. Mayer, R. Muche, and K. Hohl. Software for the handling and imputation of missing data – an overview. *Journal of Clinical Trials*, 2012.

P. McCullagh and J. A. Nelder. *Generalised Linear Models*. Chapman & Hall, New York, 1989.

M. A. McIsaac and R. J. Cook. Statistical models and methods for incomplete data in randomized clinical trials. In K. van Montfort, J. Oud, and W. Ghidey, editors, *Developments in Statistical Evaluation of Clinical Trials*. Springer, 2013a.

M. A. McIsaac and R. J. Cook. Biomarkers for disease progression in rheumatology: A review and empirical study of two-phase designs. *Canadian Journal of Statistics*, 2013b. Submitted.

M. A. McIsaac and R. J. Cook. Two-phase designs leading to incomplete exposure information in clustered or longitudinal data. *Springer Lecture Notes*, ISS2012 Symposium proceedings volume, 2013c. In press.

M. A. McIsaac, R. J. Cook, and M. Poulin-Costello. Incomplete data in randomized dermatology trials: Consequences and statistical methodology. Reviewed and resubmitted to Dermatology, 2013.

A. Menter, K. Gordon, W. Carey, T. Hamilton, S. Glazer, I. Caro, N. Li, W. Gulliver, et al. Efficacy and safety observed during 24 weeks of efalizumab therapy in patients with moderate to severe plaque psoriasis. *Archives Of Dermatology*, 141(1):31, 2005.

A. Menter, S. R. Feldman, G. D. Weinstein, K. Papp, R. Evans, C. Guzzo, S. Li, L. T. Dooley, C. Arnold, A. B. Gottlieb, et al. A randomized comparison of continuous vs. intermittent infliximab maintenance regimens over 1 year in the treatment of moderate-to-severe plaque psoriasis. *Journal Of The American Academy Of Dermatology*, 56(1):31–e1, 2007.

A. Menter, S. K. Tyring, K. Gordon, A. B. Kimball, C. L. Leonardi, R. G. Langley, B. E. Strober, M. Kaul, Y. Gu, M. Okun, et al. Adalimumab therapy for moderate to severe psoriasis: A randomized, controlled phase iii trial. *Journal of the American Academy of Dermatology*, 58 (1):106, 2008.

G. Molenberghs and M. Kenward. *Missing Data in Clinical Studies*. John Wiley & Sons Ltd, West Sussex, England, UK, 2007.

G. Molenberghs, H. Thijs, I. Jansen, C. Beunckens, M. G. Kenward, C. Mallinckrodt, and R. J. Carroll. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5(3):445–464, 2004.

Panel on Handling Missing Data in Clinical Trials National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press, 2010. ISBN 9780309158145.

J. Neyman. Contribution to the theory of sampling from human populations. *Journal of the American Statistical Association*, 33:101–116, 1938.

K. A. Papp, S. Tyring, M. Lahfa, J. Prinz, CEM Griffiths, A. M. Nakanishi, R. Zitnik, and P. C. M. Van De Kerkhof. A global phase iii randomized controlled trial of etanercept in psoriasis: Safety, efficacy, and effect of dose reduction. *British journal of dermatology*, 152 (6):1304–1312, 2005.

K. A. Papp, P. Fonjallaz, F. Casset-Semanaz, J. G. Krueger, and K. M. Wittkowski. Analytical approaches to reporting long-term clinical trial data. *Current Medical Research and Opinion®*, 24(7):2001–2008, 2008a.

K. A. Papp, P. Fonjallaz, F. Casset-Semanaz, J. G. Krueger, and K. M. Wittkowski. Analytical approaches to reporting long-term clinical trial data. *Current Medical Research and Opinion®*, 24(7):2001–2008, 2008b.

K. A. Papp, R. G. Langley, M. Lebwohl, G. G. Krueger, P. Szapary, N. Yeilding, C. Guzzo, M. C. Hsu, Y. Wang, S. Li, et al. Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with psoriasis: 52-week results from a randomised, double-blind, placebo-controlled trial (phoenix 2). *The Lancet*, 371(9625):1675–1684, 2008c.

M. S. Pepe. Inference using surrogate outcome data and a validation sample. *Biometrika*, 79(2): 355–365, June 1992.

M. S. Pepe, M. Reilly, and T. R. Fleming. Auxiliary outcome data and the mean-score method. *Journal of Statical Planning and Inference*, 42:137–160, 1994.

A. Pickles, G. Dunn, and J. L. Vazquez-Barquero. Screening for stratification in two-phase ('two-stage') epidemiological surveys. *Statistical Methods in Medical Research*, 4:73–89, 1995.

A. Prakash, R. C. Risser, and C. H. Mallinckrodt. The impact of analytic method on interpretation of outcomes in longitudinal clinical trials. *International Journal Of Clinical Practice*, 62(8): 1147–1158, 2008.

P. Rahman, D. D. Gladman, R. J. Cook, Y. Zhou, G. Young, and D. Salonen. Radiological assessment in psoriatic arthritis. *Rheumatology*, 37(7):760–765, 1998.

P. Rahman, T. Snelgrove, L. Peddle, F. Siannis, V. Farewell, C. Schentag, and D. Gladman. A variant of the il4 i50v single-nucleotide polymorphism is associated with erosive joint disease in psoriatic arthritis. *Arthritis & Rheumatism*, 58(7):2207–2208, 2008.

P. S. Raina, C. Wolfson, S. A. Kirkland, and L. E. Griffith et al. The canadian longitudinal study on aging (clsa). *Canadian Journal on Aging*, 28(3):221–229, 2009.

K. Reich, F. O. Nestle, K. Papp, J. P. Ortonne, R. Evans, C. Guzzo, L. T. Dooley, and C. E. M. for the EXPRESS Study Investigators Griffiths. Infliximab induction and maintenance therapy for moderate-to-severe psoriasis: A phase iii, multicentre, double-blind trial. *Lancet*, 366:1367–1374, 2005a.

K. Reich, F. O. Nestle, K. Papp, J. P. Ortonne, R. Evans, C. Guzzo, S. Li, L. T. Dooley, and C. E. M. Griffiths. Infliximab induction and maintenance therapy for moderate-to-severe psoriasis: A phase iii, multicentre, double-blind trial. *The Lancet*, 366(9494):1367–1374, 2005b.

K. Reich, R. G. Langley, K. A. Papp, J. P. Ortonne, K. Unnebrink, M. Kaul, and J. M. Valdes. A 52-week trial comparing briakinumab with methotrexate in patients with psoriasis. *New England Journal of Medicine*, 365(17):1586–1596, 2011.

M. Reilly. Optimal sampling strategies for two phase studies. *American Journal of Epidemiology*, 143:92–100, 1996.

M. Reilly and M. S. Pepe. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314, June 1995.

M. Reilly and M. S. Pepe. The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medicine*, 16:5–19, 1997.

S. I. Resnick. *A Probability Path*. Birkhauser, New York, 1999.

J. M. Robins and Y. Ritov. Toward a curse of dimensionality approximate (coda) asymptotic theory for semiparameric models. *Statistics in Medicine*, 16:285–319, 1997.

J. M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87:113–124, 2000.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, September 1994.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, March 1995.

J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidermiology*, 11(5):550–560, 2000.

K. J. Rothman and S. Greenland, editors. *Modern Epidemiology, Second Edition*. Lippincott Williams & Wilkins, Philadelphia, 1998.

A. Rotnitzky and D. Wypij. A note on the bias of estimators with missing data. *Biometrics*, 50 (4):1163–1170, December 1994.

S. Roy, N. Chen, and M. Cifaldi. Pms70 a comparison of non-responder imputation and last-observation-carried-forward analysis methods in rheumatoid arthritis clinical trials. *Value in Health*, 14(3):A136–A136, 2011.

D. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.

D. B. Rubin and N. Schenker. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81: 366–374, 1986.

D. L. Sackett and M. Gent. Controversy in counting and attributing events in clinical trials. *New England Journal of Medicine*, 301(26):1410–1412, 1979.

J. H. Saurat, G. Stingl, L. Dubertret, K. Papp, R. G. Langley, J. P. Ortonne, K. Unnebrink, M. Kaul, A. Camez, and for the CHAMPION Study Investigators. Efficacy and safety results from the randomized controlled comparative study of adalimumab vs. methotrexate vs. placebo in patients with psoriasis (champion). *British Journal of Dermatology*, 158:558–566, 2007.

J. H. Saurat, G. Stingl, L. Dubertret, K. Papp, R. G. Langley, J. P. Ortonne, K. Unnebrink, M. Kaul, and A. Camez. Efficacy and safety results from the randomized controlled comparative study of adalimumab vs. methotrexate vs. placebo in patients with psoriasis (champion). *British Journal of Dermatology*, 158(3):558–566, 2008.

J. L. Schafer. Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8:3–15, 1999.

D. Schaubel, J. Hanley, J. P. Collet, J. F. Boivin, C. Sharpe, H. I. Morrison, and Y. Mao. Two-stage sampling for etiologic studies sample size and power. *American Journal Of Epidemiology*, 146(5):450–458, 1997.

N. Schenker and A. H. Welsh. Asymptotic results for multiple imputation. *The Annals of Statistics*, 16(4):1550–1566, 1988.

K. F. Schulz, D. G. Altman, D. Moher, et al. Consort 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMC medicine*, 8(1):18, 2010.

A. J. Scott and C. J. Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):57–71, 1997.

A. J. Scott and C. J. Wild. Discussions. *International Statistical Review*, 79(2):228–230, 2011a.

A. J. Scott and C. J. Wild. Fitting regression models with response-biased samples. *Canadian Journal of Statistics*, 39(3):519–536, 2011b.

S. R. Seaman, I. R. White, A. J. Copas, and L. Li. Combining multiple imputation and inverse-probability weighting. *Biometrics*, 68:129–137, 1999.

D. A. Sprott. *Statistical Inference in Science*. Springer, New York, 2000.

D. A. Sprott and V. T. Farewell. Randomization in experimental science. *Statistical Papers*, 34: 89–94, 1993.

M. J. Stampfer, W. C. Willett, G. A. Colditz, B. Rosner, F. E. Speizer, and C. H. Hennekens. A prospective study of postmenopausal estrogen therapy and coronary heart disease. *The New England Journal of Medicine 313(17)*, 313(17):1044–1049, 1985.

J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ: British Medical Journal*, 338, 2009.

B. E. Strober, J. J. Crowley, P. S. Yamauchi, M. Olds, and D. A. Williams. Efficacy and safety results from a phase iii, randomized controlled trial comparing the safety and efficacy of briakinumab with etanercept and placebo in patients with moderate to severe chronic plaque psoriasis. *British Journal of Dermatology*, 165(3):661–668, 2011.

J. Sun. *The Statistical Analysis of Interval-censored Failure Time Data*. Springer, Columbia, 2006.

B. C. Sutradhar and K. Das. On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika*, 86:459–465, 1999.

R. Sutradhar and R. J. Cook. A bivariate mover – stayer model for interval-censored recurrent event data: Application to joint damage in rheumatology. *Communications in Statistics – Theory and Methods*, 38(18):3389–3405, 2009.

D. Tolusso and R. J. Cook. Robust estimation of state occupancy probabilities for interval-censored multistate data: An application involving spondylitis in psoriatic arthritis. *Communications in Statistics – Theory and Methods*, 38(18):3307–3325, 2009.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Science + Business Media, New York, 2006.

A. A. Tsiatis and M. Davidian. Discussions. *International Statistical Review*, 79(2):221–223, 2011.

S. Tyring, K. B. Gordon, Y. Poulin, R. G. Langley, A. B. Gottlieb, M. Dunn, and A. Jahreis. Long-term safety and efficacy of 50 mg of etanercept twice weekly in patients with psoriasis. *Archives Of Dermatology*, 143(6):719, 2007.

University Health Network Centre For Prognosis Studies In The Rheumatic Diseases. Psoriatic arthritis clinic newsletter. `http://www.uhnresearch.ca/studies/cpsrd/psa/psa_nl10.html`, May 2007.

P. C. M. Van de Kerkhof, S. Segaert, M. Lahfa, T. A. Luger, Z. Karolyi, A. Kaszuba, G. Leigheb, F. M. Camacho, D. Forsea, C. Zang, et al. Once weekly administration of etanercept 50 mg is efficacious and well tolerated in patients with moderate-to-severe plaque psoriasis: A randomized controlled trial with open-label extension. *British Journal of Dermatology*, 159 (5):1177–1185, 2008.

D. van der Heijde, J. Sharp, S. Wassenberg, and D. D. Gladman. Psoriatic arthritis imaging: A review of scoring methods. *Annals Of The Rheumatic Diseases*, 64(suppl 2):ii61–ii64, 2005.

W. M. van der Wal and R. B. Geskus. Ipw: An r package for inverse probability weighting. *Journal of Statistical Software*, 43(i13), 2011.

E. Walter and L. Pronzato. Qualitative and quantitative experiment design for phenomenological models – a survey. *Automatica*, 26(2):195–213, 1990.

N. Wang and J. M. Robins. Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85:935–948, 1998.

C. C. Wen and C. T. Lin. Analysis of current status data with missing covariates. *Biometrics*, Nov 2010. no. doi: 10. 1111/j. 1541-0420. 2010. 01505. x.

H. A. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25, 1982a.

J. E. White. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128, 1982b.

A. S. Whittemore and J. Halpern. Multi-stage sampling in genetic epidemiology. *Statistics in Medicine*, 16:153–167, 1997.

G. K. Wilcock, S. Lilienfeld, and E. Gaens. Efficacy and safety of galantamine in patients with mild to moderate alzheimer's disease: Multicentre randomised controlled trial. *Bmj*, 321 (7274):1445, 2000.

C. J. Wild. Fitting prospective regression models to case-control data. *Biometrika*, 78:705–717, December 1991.

J. Wittes and E. Brittain. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9(1-2):65–72, 1990.

F. Xie and M. C. Paik. Multiple imputation methods for the missing covariates in generalized estimating equation. *Biometrics*, 53:1538–1546, 1997.

G. Y. Yi and R. J. Cook. Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association*, 97:1071–1080, December 2002.

M. Yu and B. Nan. A revisit of semiparametric regression models with missing data. *Statistica Sinica*, 16(4):1193, 2006.

Z. Zhang and H. E. Rockette. On maximum likelihood estimation in parametric regression with missing covariates. *Journal of Statistical Planning and Inference*, 134:206 – 223, 2005.

L. P. Zhao and R. L. Prentice. Correlated binary regression using a quadratic exponential model. *Biometrika*, 77(3):642–648, 1990.

Y. Zhao, J. F. Lawless, and D. L. McLeish. Likelihood methods for regression models with expensive variables missing by design. *Biometrical Journal*, 51(1):123–136, 2009.