Semiparametric methods for multiphase response-selective samples

Gustavo Guimarães de Castro Amorim

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Statistics,
The University of Auckland, 2014.

Abstract

Missing data may arise due to happenstance, when some units fail to respond, or due to design, such as in multi-phase sampling schemes. The goal is to model Y in terms of a set of covariates $\boldsymbol{X}$. Efficient analysis for multiphase studies can be obtained by maximizing the resulting (full) likelihood. However, if the final (and fully observed) sample is outcome-dependent, the resulting likelihood involves the marginal distribution of an often high-dimensional $\boldsymbol{X}$. Modelling this marginal distribution may be hard or even unfeasible and methods that treat it non-parametrically are of interest.

Semiparametric methods in which only the conditional distribution of Y given $\boldsymbol{X}$ is treated parametrically have been widely discussed in the literature. Most methods, however, estimate the probability of providing full information through a saturated model, which may only be possible in specific scenarios. Moreover, fully efficient methods often do not take into account extra information that is not part of the model of interest. These data are discarded and approaches that make a better use of the whole information are needed.

Here we present a semiparametric method, denoted by CML+$\widetilde{\boldsymbol{S}}$, that copes with both situations. We first showed that it is consistent and asymptotically normal under mild conditions and later performed extensive simulated studies, for both discrete and continuous responses. Our simulations showed substantially gains in efficiency when

extra variables, not used for selecting the data, were taken into account. This was later extended to a wider class of designs, which encompasses the well-known case-control study and many others. The method was shown to be consistent and more efficient than the commonly used weighted approach in all cases analysed, but not as robust to model misspecifications.

The method is strongly connected to propensity scores and a discussion between their similarities and differences were also conducted. Both approaches were later combined, providing an alternative method for estimating treatment effects that could be applied in outcome-dependent problems.

Finally, we discussed its asymptotic efficiency by numerically deriving the semiparametric efficiency bound. The proposed estimator seemed to achieve, for the some specific scenarios, the semiparametric efficiency bound. For a discrete response the equality is mathematically guaranteed and the CML+$\widetilde{\boldsymbol{S}}$ method is thus semiparametric fully efficient.

Acknowledgements

I would like to thank my supervisors, Professors Chris Wild and Alastair Scott. As clichéd as it may sound, I cannot express my gratitude enough. I was only able to start, continue and now conclude my degree because of their support over all these years. Thank you for all the guidance and insights, for your patience and encouragement, for the endless proofreading and countless English lessons, but above all for the opportunity to work closely with some of the best researchers in this field.

I would also like to thank Professors Thomas Lumley and Alan Lee. Their help, support and suggestions greatly improved this work and completing it would not be possible without their help. Thanks also to Yannan for all her help and to the IT team for all the IT assistances they have provided.

My thanks to the Marsden Fund and to the University of Auckland Doctoral Scholarship for the financial support that enabled me to complete this research and report its findings in national and international conferences.

On a personal note, I would like to thank all my friends and colleagues: To all my Park Road friends (especially Sai), to all my friends in Brazil (especially Rodrigo), to Jon, Sam, Serg and to all my friends, lecturers and staff from the Department of Statistics. Thanks to Alex and Nancy, who have been more than just colleagues, but also good friends.

I also would like to thank my parents, my brother and Ju, for always being near despite being far. And finally my wife, who embraced this challenge and changed her life in order to help me to conclude my studies. Thank you for always being supportive, kind and helpful, for the warm hugs after tough days and cheerful laughter that helped me to overcome any difficulties on the way.

Lastly, thanks to God, for always including only the best people in my way.

Contents

List of Figures

List of Tables

List of Abbreviations

AIPW	Augmented Inverse-Probability Weighted
CML	Conditional Maximum Likelihood
Emp.SE	Empirical Standard Error
Est.SE	Estimated Standard Error
MAR	Missing at random
MCAR	Missing complete at random
MSE	Mean Squared Error
NMAR	Not missing at random

1

Introduction

1.1 Biased samples

Biased samples, the result of selection processes that give rise to sample distributions that differ systematically from the true target distribution, can be found in most data collection and, if not treated properly, can lead to inconsistent estimates.

Biased sampling may occur by design. Suppose that we are interested in predicting the probability that an individual will contract a rare disease. Partial information such as the disease status may be known for all individuals in the study (cohort), but more predictive information may still be required. For economic reasons, however, not all elements can be selected for full observations and so a common approach is to randomly choose a reasonable number of cases (subjects with the disease) and a similar number of controls (disease-free individuals). Notice that this leads to subjects with incomplete information and also to a ratio of cases/controls in the completely-observed sample which is very different from the true one. The sample is, therefore, biased and by not taking this into consideration one might get very inconsistent estimates.

Biased sampling may also occur by happenstance. A survey, for example, where men have higher probability of responding than women leads to a biased sample and if the outcome of interest is somehow linked to sex, we must be cautious when making inference to the entire population.

Note that if complete information had been obtained for both cases, inference could be made in standard ways. In real studies, however, it is common to have data with missing information. These missing data points may behave differently from the observed data in such way that ignoring it may lead to badly biased estimates, making missing data an important topic of discussion.

1.2 Missing data

Data sets are usually presented in matrix format where entries are values corresponding to the observed values of variables. Ideally, all entries of such a matrix would be filled. However, in most data collecting process it is unlikely to have full information observed for all individuals in the study, i.e., the matrix will most likely have empty cells. Since this missing information may bias future inferences, many methods have then been proposed. These methods rely on strong assumptions regarding how the data is missing, more specifically, regarding the missing data patterns and mechanisms.

1.2.1 Missing data patterns

Little and Rubin (2002) describe different missing data patterns and mechanisms that lead to datasets with incomplete information. The major missing patterns are:

- **General:** Here the observations are randomly missing, in a sense that the missingness is not following any kind of pattern.

- **Monotone:** Let S_1 be a set of individuals with fully observed variables. Consider the following process of collecting data. First, a subset of S_1, S_2, say, is taken and extra variables are observed for these sampled individuals only. Next, a subset of S_2, S_3, say, is now taken and extra information collected for these units. This procedure can have as many steps as necessary, but the point is that each subset has more information than the previous set. That is, the missingness is monotonically decreasing.

- **Univariate:** As the name suggests, only one variable has missing data, while all remaining variables are fully observed. If the number of partially observed variables is greater than 1 and they all are missing on the same set of subjects, we have a **multivariate missing pattern**.

1.2.2 Missing data mechanisms

The missing information can be related to the data and are categorized by Little and Rubin (2002) in three groups as follows:

- *Missing completely at random* (MCAR): Here the missingness does not depend on any value of the data. Each missing observation is just as likely to be missing as any other piece of data. Such missingness can be ignored when making inference to the population.

- *Missing at random* (MAR): Here the probability of missingness depends only on the observed components. Since this is less restrictive than a MCAR assumption, MAR is a common assumption in statistical analysis. It is commonly found in studies where the missingness is due to design, such as the expensive covariate problem discussed in section 1.1. Here, the expensive covariate is purposefully

observed for a sample of individuals selected from the total population or cohort. Selection into this sample is based on variables that have been fully observed for the entire cohort, so that information is MAR.

- *Missing not at random* (NMAR): Here the probability of missingness depends in part on the missing values of the data. It is quite common in real problems and results in biased inference. Non-response regarding family income, for instance, usually depends on the size of that income, resulting in a sample MNAR.

In real data problems we cannot really tell whether the missing data are MAR or MNAR, but we can in principle distinguish between MCAR and MAR. If the missing observations are MNAR, there is not much that can be done apart from forms of sensitivity analysis. For convenience, the missingness is usually assumed to be independent of the missing observations (MAR), enabling the modelling of missingness as an input to making inferences. This will be the case discussed throughout this thesis.

1.3 Basic methods

We now introduce a few methods still commonly used in real data problems to handle missing data. These methods are easy to implement and use standard complete-data statistical software to make inference to the entire population. In what follows we will focus on complete case analysis and multiple imputation.

1.3.1 Complete-case analysis

The simplest and most commonly used approach to missing is the so-called complete-data analysis. Here, the partially observed units are discarded so that only the completely observed subjects are used together with a standard method of analysis. This

is valid for the MCAR case, but otherwise may lead to inconsistent estimates and be seriously inefficient, particularly if the missingness depends on the response variable. Consider, for example, a survey where people with high income are less likely to report their income level. The MCAR assumption is clearly not valid and by considering only completely observed units, the resulting inference will be biased in favour of patterns amongst those with lower incomes.

Complete-case analysis should also be avoided if there is appreciable missingness and partial information is available from the partially observed subjects. Otherwise, information is being discarded leading to inefficient estimates. Valid applications of complete-case analysis tend, therefore, to be limited to a small class of problems.

1.3.2 Single and multiple imputation

A broader approach that makes better use of the data is imputation. Single imputation consists in estimating one value for each missing observation, leading to a complete dataset composed of the observed elements and the estimated ones. The parameter of interest can be easily estimated by conducting a standard analysis for complete data. Different methods can be used to estimate the missing values. Mean imputation, for example, replaces the missing values for a variable by the mean of the observed values. Regression imputation uses the observed values to fit a regression model and predict the missing values. Note that in both cases only a single value is estimated for each missing unit. The imputed values are then treated as if they were known values, ignoring a source of uncertainty and thus leading to an underestimation of the variance.

Multiple imputation (Rubin, 1987) on the other hand, accounts for imputation uncertainty. Instead of filling in each missing observation with only one value, it generates multiple values reflecting prediction uncertainties given an imputation model. Multi-

ple imputation analysis are easy to implement but it demands an intensive computing effort, since we have to impute many data sets, estimate $\boldsymbol{\beta}$, the parameter of interest, from each one and combine the results into one summary. The method work as follows. M values are generated (see Shieh (2003)) for each missing observation, resulting in M completed datasets and in M estimates of $\boldsymbol{\beta}$ and its variance. These M estimates are combined using "Rubin rule's" (see Little and Rubin (2002)) to produce a single combined estimate to be used for inference. Further applications can be found in Zhou et al. (2001), Shen (2007) and Stuart et al. (2009).

The success of multiple imputation, however, relies on the model assumed to impute the data and a wrong model can lead to biased estimates. Another criticism sometimes made is that it "creates data" instead of using only the observed data.

1.4 Multiphase sampling

Multiphase sampling is useful when collecting full information from all subjects in the study may be too expensive or even unfeasible. It can be used as a way to reduce costs while improving statistical efficiency. A 2-phase sampling scheme consists, for example, in collecting information on variables that are cheap or easy to measure for a large sample or even entire finite population, and additional variables for a subsample. The phase-1 data consist of all-subjects information on always-observed variables. The phase-2 sample consists of those subjects selected for further data collection.

The 2-phase sampling scheme was first introduced in the context of case-control studies by White (1982). Here the goal was to study an association between a binary response and a binary exposure variable, adjusted for discrete covariates. The author proposed to sample individuals from each stratum defined by the combination of response/exposure and measuring covariate information for the sampled subjects. Since

then, 2-phase sampling schemes have become widely used for cost savings, particularly when some covariates are expensive or hard to measure.

2-phase sampling schemes can also be used for efficiency gains. Zhao et al. (2012), for example, discuss analytical and numerical approaches to compare 1-phase and 2-phase designs with respect to their efficiency. The authors assume that the response, as well as a X-surrogate variable, were fully observed at phase-1 and that X, the expensive covariate, was only measured for a sample selected from the phase-1 individuals. If the phase-2 sample is obtained via a simple random sampling taken from the phase-1 population, Zhao et al. (2012) show that the 2-phase sampling scheme is, in general, more efficient than the 1-phase sampling design. It is especially more efficient when the relative cost, defined as

$$R = \frac{C_I}{C_C},$$

where C_C and C_I are the cost for observing the complete (response, X and X-surrogate) and incomplete datasets (individuals with only response and X-surrogate measured), respectively, is low and the X and its surrogate are strongly correlated. Moreover, assuming only discrete variables, they also show that a stratified 2-phase design (where the phase-2 sample is random selected from each stratum defined by the response and the X-surrogate variables) is more efficient than the "balancing design", where the probability of selecting individuals into phase-2 are proportional to the inverses of stratum sizes (Breslow and Cain, 1988), and that the "balancing design" is more efficient than the 2-phase design with a simple random sampling scheme. Gains in efficiency are greater when fine X-surrogates are observed for the phase-1 sample.

Non-response problems can also be viewed as 2-phase problem in cases where the response is assumed to be random with a probability that depends only on information available for the phase-1 population. Thus, whether the data is missing by design or

by happenstance, the same techniques can be applied to solve both problems.

Multiphase sampling is an obvious extension of the 2-phase sampling, where more detailed information is collected at each phase from a sample of subjects selected from the previous phase. Subsamples, as pointed out by Lawless et al. (1999) are often selected based on one of the following schemes:

1. **Basic stratified sampling (BSS)** : Here the study population is divided in K strata and a pre-specified number of subjects or a pre-specified fraction of subjects from each stratum is selected for the next phase.

2. **Variable probability sampling (VPS)** : Here units are inspected sequentially as they arise independently and are selected for the next phase with probability π. For a stratified VPS, each unit has its stratum identified and when the ith unit belongs to stratum S_j, it is selected for the next phase with probability π_j.

Note that, for BSS the sample size n_j selected from stratum S_j is considered fixed given the $N_j s$, the total number of individuals in stratum S_j, while for the VPS the $n_j s$ are random. Following the same argument given in the Appendix B of Scott and Wild (2001), in the stratified sampling case both methods lead to the same likelihood which is described in the next section.

1.4.1 Sampling scheme

In most of what follows, we are interested in explaining the response $\boldsymbol{Y}$ in terms of potential covariates $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, assuming a parametric model for the conditional distribution $f(\mathbf{y}|\mathbf{x};\boldsymbol{\beta})$, termed the *model of interest*. The vector parameter $\boldsymbol{\beta}$ will be termed the *parameter of interest*.

Consider the following 2-phase sampling scheme (see figure 1.1). Let $(\boldsymbol{Y}, \boldsymbol{X}_1)$ be

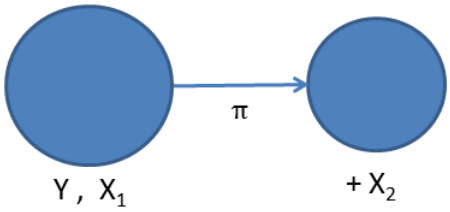


Figure 1.1: Sampling scheme for a 2-phase study, where the response $\boldsymbol{Y}$ and a covariate $\boldsymbol{X}_1$ are fully observed at phase-1, but another covariate $\boldsymbol{X}_2$ is only measured at phase-2.

fully observed for all N individuals, which either is, or is regarded as, a sample from an infinite population. This population or cohort is termed the phase-1 sample. A phase-2 sample is taken from the phase-1 sample with selection depending upon $\boldsymbol{Y}$ and possibly on $\boldsymbol{X}_1$ as well and $\boldsymbol{X}_2$ is observed. That is, $\boldsymbol{X}_2$ has only been partially observed and full information regarding $(\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2)$ is only available for those individuals selected into phase-2. Denote by R_i an indicator variable that is equal to 1 if the ith unit has been selected for full observation and 0 otherwise and let $\pi_i = \pi(\mathbf{x}_i, \mathbf{y}_i)$ be the probability of observing $R_i = 1$ and $(\boldsymbol{X}, \boldsymbol{Y})$, and $1 - \pi(\mathbf{x}_i, \mathbf{y}_i)$ the probability of observing $R_i = 0$ and $(\boldsymbol{X}_1, \boldsymbol{Y})$.

The complete or full likelihood L_F which encompasses both complete and incomplete data under the missing at random assumption is given by

$$L_F = \prod_i [\pi(\mathbf{x}_i, \mathbf{y}_i) f(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\beta}) g(\mathbf{x}_{2i}|\mathbf{x}_{1i})]^{R_i} [(1 - \pi(\mathbf{x}_i, \mathbf{y}_i)) f(\mathbf{y}_i|\mathbf{x}_{1i})]^{1-R_i}$$

where

$$\pi_i = \text{pr}(R_i = 1|\mathbf{x}_i, \mathbf{y}_i) = E(R_i|\mathbf{x}_i, \mathbf{y}_i), \quad f(\mathbf{y}_i|\mathbf{x}_{1i}) = \int_{\mathbf{x}_2} f(\mathbf{y}_i|\mathbf{x}_2, \mathbf{x}_{1i}; \boldsymbol{\beta}) g(\mathbf{x}_2|\mathbf{x}_{1i}) d\mathbf{x}_2,$$

and g is the conditional distribution of $\boldsymbol{X}_2$ given the observed $\boldsymbol{X}_1$ with cumulative

distribution function G. Notice that the likelihood depends on the selection probability $\pi(\mathbf{x}_i, \mathbf{y}_i)$, the probability that the ith individual is selected for full observation, which is assumed to be greater than zero ($\pi_i > 0$) for all units in the population. If the probability of selecting a subject depends on the outcome, the sampling scheme is called response-selective, response-biased or outcome-dependent, which will be the case throughout this thesis.

Since the selection probability does not involve $\boldsymbol{\beta}$ and we can write the full likelihood as

$$L_F(\boldsymbol{\beta}, g) \propto \prod_i [f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\beta}) g(\mathbf{x}_{2i}|\mathbf{x}_{1i})]^{R_i} \left[\int f(\mathbf{y}_i|\mathbf{x}_2, \mathbf{x}_{1i};\boldsymbol{\beta}) g(\mathbf{x}_2|\mathbf{x}_{1i}) d\mathbf{x}_2\right]^{1-R_i} \tag{1.1}$$

and its score function $\boldsymbol{S}_F(\boldsymbol{\beta}, g)$ with respect to $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\boldsymbol{S}_F(\boldsymbol{\beta}, g) &= \frac{\partial}{\partial \boldsymbol{\beta}} \log L_F(\boldsymbol{\beta}, g) \\ &= \sum_{i:R_i=1} \frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\beta}) + \sum_{i:R_i=0} \frac{\partial}{\partial \boldsymbol{\beta}} \log \int f(\mathbf{y}_i|\mathbf{x}_2, \mathbf{x}_{1i};\boldsymbol{\beta}) g(\mathbf{x}_2|\mathbf{x}_{1i}) d\mathbf{x}_2.\end{aligned}$$

By setting $\boldsymbol{S}_i = \partial \log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, we have that

$$\begin{aligned}\boldsymbol{S}_F(\boldsymbol{\beta}, g) &= \sum_{i:R_i=1} \boldsymbol{S}_i(\boldsymbol{\beta}) + \sum_{i:R_i=0} \frac{\int_{\mathbf{x}_2} \boldsymbol{S}_i(\boldsymbol{\beta}) f(\mathbf{y}_i|\mathbf{x}_2, \mathbf{x}_{1i};\boldsymbol{\beta}) g(\mathbf{x}_2|\mathbf{x}_{1i}) d\mathbf{x}_2}{\int_{\mathbf{x}_2} f(\mathbf{y}_i|\mathbf{x}_2, \mathbf{x}_{1i};\boldsymbol{\beta}) g(\mathbf{x}_2|\mathbf{x}_{1i}) d\mathbf{x}_2} \\ &= \sum_{i:R_i=1} \boldsymbol{S}_i(\boldsymbol{\beta}) + \sum_{i:R_i=0} E\left[\boldsymbol{S}_i(\boldsymbol{\beta})|\mathbf{y}, \mathbf{x}_1\right]. \end{aligned} \tag{1.2}$$

The first term of equation (1.2) is the score function of the completely observed data and the second one uses additional information obtained from the incomplete observations. If all models are correctly specified, methods that are based on equation (1.1) are more efficient than the methods discussed previously and will be presented next. Use of (1.2) requires either g known or some sort of consistent estimation of g.

1.5 Likelihood based methods

Working with the full likelihood (1.1) will generally result in more efficient estimates than obtained via previous methods. If all data were known, we could just maximize the likelihood or solve $\sum_i \boldsymbol{S}_i(\boldsymbol{\beta}) = 0$ which does not involve g so inference would be straightforward. For the case of missing data, the second term of equation (1.2) depends on the conditional distribution $g(\mathbf{x}_2|\mathbf{x}_1)$, which is usually unknown. One way to work is to treat the likelihood (1.1) non-parametrically with respect to $g(\mathbf{x}_2|\mathbf{x}_1)$.

In a seminal paper, Robins et al. (1994) defined *Augmented Inverse-Probability Weighted* (AIPW) estimators, a class of estimators that solve the weighted estimating equation

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i} \boldsymbol{S}_i(\boldsymbol{\beta}) + \sum_{i=1}^{N} \left(1 - \frac{R_i}{\pi_i}\right) A_i(\boldsymbol{\beta}) = 0, \tag{1.3}$$

where $\boldsymbol{S}_i(\boldsymbol{\beta})$ is defined as before and A_i is an arbitrary function of the phase-1 data. The most efficient choice for A_i is

$$A_i = \mathbb{E}\left(\boldsymbol{S}_i(\boldsymbol{\beta}) | \text{ observed data}\right) \tag{1.4}$$

which is equivalent to the second term of (1.2). It can be derived in some cases, but generally it is hard or even unfeasible to obtain. Methods that are efficient or nearly efficient and can be implemented are thus of interest. We discuss here pseudo-likelihood and the empirical likelihood methods.

1.5.1 Pseudo-likelihood methods

Also known as estimated likelihood, the pseudo-likelihood method consists of replacing G in likelihood (1.1) by an empirical version and solving the associated score

function for $\boldsymbol{\beta}$.

Weaver and Zhou (2005) developed an estimator for outcome-dependent sampling designs with a continuous response. They used a similar sampling scheme to the one presented earlier, but without a known $\boldsymbol{X}_1$ (so $\boldsymbol{X} = \boldsymbol{X}_2$). Thus, only the response variable $\boldsymbol{Y}$ was known for the phase-1 population. It was used to define strata from which samples are selected for full observation. In order to define these strata, the response variable was divided into K mutually exclusive intervals C_k, $k = 1, \ldots, K$. A biased sample will be generated and so a simple global empirical distribution cannot be used to estimate $G(\mathbf{x})$. But since

$$G(\mathbf{x}) = \text{pr}(\boldsymbol{X} \leq \mathbf{x}) = \sum_k \text{pr}(\mathbf{y} \in C_k)\, \text{pr}(\boldsymbol{X} \leq \mathbf{x} | \mathbf{y} \in C_k),$$

an estimative G^* for G can be given by

$$G^*(\mathbf{x}) = \sum_{k=1}^{K} \frac{N_k}{N} \hat{G}_k(\mathbf{x}),$$

where N_k/N is the proportion of elements in the kth interval and $\hat{G}_k(\mathbf{x})$ is the empirical cumulative distribution function of $\boldsymbol{X}$ given that it belongs to the kth interval. Lawless et al. (1999) considered a similar method, but for a slightly different likelihood. They assume that stratum membership information would be retained for the whole data, not for only the units in the sample, resulting in a slightly different likelihood but the same procedure is used to estimate G.

A more sophisticated method was developed by Chatterjee et al. (2003). The requirement of a positive selection probability for every unit in the sample is no longer necessary, but they assume that $\int \pi(\mathbf{y}, \mathbf{x}) dy > 0$ and $f(\mathbf{y}|\mathbf{x}) > 0$, almost surely, in a neighbourhood of the true parameter values. Their method is computationally simple

for both continuous and discrete outcomes when $\boldsymbol{X}_1$ is discrete. It consists in solving the pseudo-score function

$$\boldsymbol{S}_P(\boldsymbol{\beta};\pi) = \sum_{i:R_i=1} \boldsymbol{S}_i(\boldsymbol{\beta}) + \sum_{j:R_j=0} \sum_{i:R_i=1} \frac{\boldsymbol{S}_j(\boldsymbol{\beta}) h(\mathbf{y}_j, \mathbf{x}_{2i}, \mathbf{x}_{1j}; \boldsymbol{\beta}) I(\mathbf{x}_{1j} = \mathbf{x}_{1i})}{\sum_{l:R_l=1} h(\mathbf{y}_j, \mathbf{x}_{2l}, \mathbf{x}_{1j}; \boldsymbol{\beta}) I(\mathbf{x}_{1j} = \mathbf{x}_{1l})} = 0 \quad (1.5)$$

where

$$h(\mathbf{y}, \mathbf{x}; \boldsymbol{\beta}) = \frac{f(\mathbf{y}|\mathbf{x}; \boldsymbol{\beta})}{\pi(\mathbf{x})}$$

and $I(\boldsymbol{X}_{1j} = \boldsymbol{X}_{1l})$ is an indicator variable equal to 1 if $\boldsymbol{X}_{1j} = \boldsymbol{X}_{1l}$ and 0 otherwise. Note that the second term in equation (1.5) is, in a sense, a weighted version of the score function $\boldsymbol{S}$. Although this equation can be solved by a Newton-Raphson algorithm, the authors suggest a different algorithm that, under regularity conditions and starting with a known consistent estimate, converges to a solution of $S_P(\boldsymbol{\beta}; G_N, \widehat{\pi}) = 0$, where $\widehat{\pi}$ is an estimate of π. This method is also almost fully efficient in their simulations, as shown by Chatterjee et al. (2003) and Zhao et al. (2009), except for extreme parameter values.

Both methods are most useful when the phase-1 variables are discrete, since G can be easily estimated. For more general situations, McLeish and Struthers (2006) suggest using kernel density estimators and Monte Carlo methods to estimate (1.4), using $\boldsymbol{X}$ generated from a suitable family of distributions and averaging the results or evaluating the integral using the observed values of $\boldsymbol{X}$. Chatterjee and Chen (2007), following Chatterjee et al. (2003), suggested estimating $\boldsymbol{\beta}$ by solving an estimating equation given by

$$\boldsymbol{S}_P(\boldsymbol{\beta};\pi) = \sum_{i:R_i=1} \boldsymbol{S}_i(\boldsymbol{\beta}) + \sum_{j:R_j=0} \frac{E\left\{\boldsymbol{S}_j(\boldsymbol{\beta}) h(\mathbf{y}_j|\mathbf{x}_1, \mathbf{x}_2; \boldsymbol{\beta})|\mathbf{x}_{1j}, R=1\right\}}{E\left\{h(\mathbf{y}_j|\mathbf{x}_1, \mathbf{x}_2; \boldsymbol{\beta})|\mathbf{x}_{1j}, R=1\right\}} = 0 \quad (1.6)$$

using the kernel smoothing approach to estimate the conditional expectations. Their

approach is similar to the one developed by Carroll and Wand (1991), except that Chatterjee and Chen propose to partition $\boldsymbol{X}_1$ into a fixed number of strata and to apply the kernel smoothing approach within each stratum separately.

1.5.2 Semiparametric maximum likelihood methods

Semiparametric maximum likelihood methods are usually the most efficient ones, achieving full semiparametric efficiency in some important special cases. The idea behind this approach is to estimate G non-parametrically and solve the resulting score functions obtaining consistent estimates for $\boldsymbol{\beta}$, the regression coefficients.

Zhang and Rockette (2005) studied two different ways to maximize the complete likelihood function considering the covariate distribution unspecified. The first one, denoted by the global maximum likelihood estimates (MLE), is obtained by maximizing (1.1) over G and $\boldsymbol{\beta}$ simultaneously, which might be very difficult, since G can be distributed in rather arbitrary ways. A simpler approach is to consider a restricted MLE, where the covariate distribution is restricted to the set of probability measures concentrated on the observed values. The existence of both MLEs was shown under simple conditions, and some asymptotic results were established. The authors showed, for example, that both estimators are strongly consistent, asymptotically normal and semiparametric efficient (Zhang and Rockette, 2005, 2007), which makes the restricted MLE a good option in semiparametric maximum likelihood estimation.

As pointed out by Zhao et al. (2009), there are three choices for the support of $G(\mathbf{x}_{2i}|\mathbf{x}_{1i})$:

- the whole sample space $\mathcal{X}$ (resulting in the global MLE);
- the observed sample $\mathcal{X} = \{\mathbf{x}_i : R_i = 1\}$ (resulting in the restricted MLE);

- the observed sample given that $\boldsymbol{X}_{1i} = \mathbf{x}_1$, i.e., $\mathcal{X} = \{\mathbf{x}_i : R_i = 1 \text{ and } \boldsymbol{X}_{1i} = \mathbf{x}_1\}$ (resulting in the so-called doubly restricted MLE).

All three supports are very similar if $\boldsymbol{X}$ is discrete and contains only a few categories and so the first support can be used to maximize g. Otherwise, if $\boldsymbol{X}$ is continuous or discrete with too many levels, it is better to choose the second or third ones as the support of $G(\mathbf{x}_{2i}|\mathbf{x}_{1i})$. From Zhang and Rockette (2005), we have that all three supports are asymptotically equivalent if $\boldsymbol{X}_1$ is discrete. McLeish and Struthers (2006) use Monte Carlo simulations to estimate the integral (1.4), using the global MLE and the restricted MLE. The authors show that these two methods are closely related. Zhao et al. (2009) use the EM algorithm for computation and compared the performance of the restricted and doubly-restricted likelihood for a finite sample size in situations where certain variables are difficult or expensive to measure. The authors discuss three different situations: the "expensive covariate" setting, where partial information is measured only for a sample of the phase-1 data; the "expensive response" problem, where only partial covariates and an auxiliary variable related to $\boldsymbol{Y}$ are measured at the phase-1 and $\boldsymbol{X}$ is observed only for selected individuals; and finally a combination of both situations. The authors noticed that the results were very similar and so there is hardly any difference between using the restricted or the doubly-restricted likelihood.

Empirical likelihood

By restricting the support of G to the observed values, we are, in a sense, using an empirical estimate of it. This is closely related to the empirical likelihood method introduced in biased sampling problems by Qin (1993). It soon became an alternative method for solving response-selective and missing data problems. Qin and Lawless (1994), for example, studied estimating functions and empirical likelihood in problems

related to incomplete information in semiparametric models and Zhou et al. (2002, 2007) proposed a new estimator based on the empirical likelihood approach for response-selective problems.

The idea of the empirical likelihood is to maximize

$$l_f = \log L_F = \sum_{i:R_i=1} \log\left(f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\beta})\right) + \sum_{i:R_i=1} \log\left(g(\mathbf{x}_{2i}|\mathbf{x}_{1i})\right) + \sum_{i:R_i=0} \log\left(f(\mathbf{y}_i|\mathbf{x}_{1i})\right) \quad (1.7)$$

over g under the following constraints

$$g(\mathbf{x}_i|\mathbf{x}_{1i}) > 0, \quad \sum_i g(\mathbf{x}_{2i}|\mathbf{x}_{1i}) = 1 \quad \text{and} \quad f(\mathbf{y}_i|\mathbf{x}_{1i}) = \int f(\mathbf{y}_i|\mathbf{x}_{1i},\mathbf{x}_2;\boldsymbol{\beta}) g(\mathbf{x}_2|\mathbf{x}_{1i}) d\mathbf{x}_2. \quad (1.8)$$

Maximization of g should be carried over all distributions whose support contains the observed values but, following Owen (1988, 1990), only discrete distributions with jumps at each of the observed points need to be considered.

The regression coefficients are obtained by maximizing the resulting likelihood $L(\boldsymbol{\beta},\widehat{G})$ over $\boldsymbol{\beta}$, where $\widehat{G}$ is an estimate of G. It avoids the estimation problem of an infinite-dimensional parameter by transforming it into a problem of a finite number of parameters, which is, in the worst case, of the same size as the sample.

Wang and Zhou (2006) use the empirical likelihood method to derive fully efficient estimates for the following 2-phase outcome dependent problem. Let $\boldsymbol{Y} = \{1,\ldots,L\}$ be a categorical response, $\boldsymbol{X} = \{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_J\}$ the covariate matrix and $\boldsymbol{X}_{1d} = \{1,\ldots,K\}$ a categorical auxiliary variable for $\boldsymbol{X}_1$. All variables are known for every subject in the study, except for $\boldsymbol{X}_1$, which is observed only for a phase-2 sample selected from the phase 1 subjects. This sample is obtained by a simple random sample of size n_0 from all of phase-1 population plus an additional sample from each strata $\{\boldsymbol{Y} = l, \boldsymbol{X}_{1d} = k\}$

of size n_{lk}, for $l = 1, \ldots, L$ and $k = 1, \ldots, K$. The resulting likelihood is given by

$$L(\boldsymbol{\phi}, G(\mathbf{x}, \mathbf{x}_{1d})) = \left\{ \prod_{i \in V} f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\beta}) \right\} \left\{ \prod_{i \in V} g(\mathbf{x}_i | \mathbf{x}_{1di}) \right\} \left\{ \prod_{k=1}^{K} \prod_{l=1}^{L} {h_{lk}}^{-n_{lk}} \right\},$$

where V denotes the set of all individuals selected into phase 2, $n_{lk} = \sum_i I(\boldsymbol{Y}_i = l, \boldsymbol{X}_{1di} = k)$, for $i \in V$, and $h_{lk} \equiv \text{pr}(\mathbf{y} = l | \mathbf{x}_{1d} = k) = \int \text{pr}(\mathbf{y} = l | \mathbf{x}) dG(\mathbf{x} | \mathbf{x}_{1d} = k)$. In order to estimate $\boldsymbol{\phi} = (\boldsymbol{\beta}, h_{lk})'$, we first profile the above likelihood with respect to $p_{ik} = g(\mathbf{x}_i | \mathbf{x}_{1di} = k)$ over all distributions whose support contains the observed $\boldsymbol{X}$ values, with the constraints (1.8). That is, we maximize

$$\begin{aligned} H(\boldsymbol{\phi}, p_{ik}) = & \sum_{i \in V} \log f(\mathbf{y}_i | \mathbf{x}_i) + \sum_{k=1}^{K} \sum_{i \in V_k} \log p_{ik} - \sum_{k=1}^{K} \sum_{l=1}^{L} n_{lk} \log h_{lk} + \\ & + \sum_{k=1}^{K} t_k \left(1 - \sum_{i \in V_k} p_{ik} \right) + \sum_{k=1}^{K} \lambda_k \sum_{i \in V_k} p_{ik} \left(h_{ik} - f(\mathbf{y}_i | \mathbf{x}_i) \right), \end{aligned}$$

where t_k and λ_k are the two Lagrange multipliers from the two constraints (1.8). Using the fact that p_{ik} is a probability, we get

$$t_k = n_k \quad \text{and} \quad p_{ik} = \frac{1}{n_k} \frac{1}{1 + \lambda_k \left(f(\mathbf{y}_i | \mathbf{x}_i) - h_{lk} \right)},$$

with restriction

$$\frac{1}{n_k} \sum_{i \in V_k} \frac{f(\mathbf{y}_i | \mathbf{x}_i) - h_{1k}}{1 + \lambda_k \left(f(\mathbf{y}_i | \mathbf{x}_i) - h_{1k} \right)} = 0, \tag{1.9}$$

where $\lambda_k = n_{1k}/(n_k h_{1k}) - n_{2k}/(n_k h_{2k})$. Since λ_k is not centred around zero, the authors make the change of variable $\nu_k = \lambda_k - n_{1k}/(n_k h_{1k}) + n_{2k}/(n_k h_{2k})$ so that ν_k is centred around zero. Estimates can now be obtained solving the score equations $\partial \log L / \partial \boldsymbol{\beta} = 0$, $\partial \log L / \partial h = 0$ and $\partial \log L / \partial \boldsymbol{\nu} = 0$.

Wang et al. (2009) follows the same procedure, but for a slightly different likelihood. They consider that stratum membership is known for every subject in the study so that

the exponent of h_{lk} is $N_{lk} - n_{lk}$, where N_{lk} is the total number of subjects with $\boldsymbol{Y} = L$ and $\boldsymbol{X}_{1d} = k$.

Scott and Wild approach

A close related procedure was used by Scott and Wild (1997, 2001, 2006) to obtain fully efficient estimators (see Lee and Hirose (2010)) for response-selective problems for empty $\boldsymbol{X}_1$. The authors suggest maximizing the likelihood (1.1) over p_j, where $p_j = g(\mathbf{x}_j)$, under the constraint $\sum_j p_j = 1$, and then maximize the resulting function over $\boldsymbol{\beta}$. Their work focuses on the case where both response and covariates are discrete and derive fully efficient estimates. We will now explain this approach in some detail. The simplest version consists of N individuals divided into two strata according to their disease-status. Random samples of sizes n_1 and n_2 are selected from the diseased and disease-free groups, respectively. Let $\boldsymbol{Y} = 1, 2, \ldots, I$. The likelihood is given by

$$L = \prod_{i=1}^{I} \prod_{j=1}^{n_i} \mathrm{pr}(\mathbf{x}_{ij} | \mathbf{y} = i),$$

which can be written as

$$\begin{aligned} L &= \prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{\mathrm{pr}(\mathbf{y} = i | \mathbf{x}_{ij}; \boldsymbol{\beta}) g(\mathbf{x}_{ij})}{\mathrm{pr}(\mathbf{y} = i)} \\ &= \prod_{i=1}^{I} \left\{ \prod_{j=1}^{n_i} \mathrm{pr}(\mathbf{y} = i | \mathbf{x}_{ij}; \boldsymbol{\beta}) g(\mathbf{x}_{ij}) \right\} \left\{ \int \mathrm{pr}(\mathbf{y} = i | \mathbf{x}; \boldsymbol{\beta}) dG(\mathbf{x}) \right\}^{-n_i}. \end{aligned} \tag{1.10}$$

If stratum membership is known for each element and supposing that the total population is composed of N_1 cases and N_2 controls, the likelihood can be written as

$$\begin{aligned} L &= \prod_{i=1}^{I} \left\{ \prod_{j=1}^{n_i} \text{pr}(\mathbf{x}_{ij}|\mathbf{y}=i) \right\} \text{pr}(\mathbf{y}=i)^{N_i} \\ &= \prod_{i=1}^{I} \left\{ \prod_{j=1}^{n_i} \text{pr}(\mathbf{y}=i|\mathbf{x}_{ij};\boldsymbol{\beta}) g(\mathbf{x}_{ij}) \right\} \left\{ \int \text{pr}(\mathbf{y}=i|\mathbf{x};\boldsymbol{\beta}) dG(\mathbf{x}) \right\}^{N_i - n_i} \end{aligned}$$

Note that, in both cases the likelihood depends on $g(\mathbf{x})$, which is of no interest in its own right, as we have noted, and usually too complicated for modelling. The standard form of analysis results from Anderson (1972) and Prentice and Pike (1979). They have shown that, for the binary logistic regression model with an intercept (later extended to a broader class called "multiplicative intercept models" by Scott and Wild (1997)), maximum likelihood estimates for all regression coefficients except for the intercept can be obtained by ignoring the case-control scheme, i.e., the case-control problem can be treated as a prospective one. The intercept, as shown by Scott and Wild (1986), can subsequently be adjusted by using an offset $\boldsymbol{\alpha}$ given by

$$\boldsymbol{\alpha} = \boldsymbol{\beta} + k\mathbf{e},$$

where $\boldsymbol{\beta}$ is the vector of logistic regression parameters, $\mathbf{e} = (1, \mathbf{0})^T$ and k is the log-ratio of the selection probability of cases to those of controls.

Scott and Wild (1997) derived fully efficient estimates based on the restricted likelihood for simple and stratified case-control studies, later extended to include situations where additional information from subjects selected for the case-control study is available (Scott and Wild, 2001). Both papers present an algorithm to obtain estimates of $\boldsymbol{\beta}$, which is also discussed and improved in a later work (Scott and Wild, 2006). Here, in addition to estimating $\boldsymbol{\beta}$ for problems involving missing data and response-biased

sampling, they discuss computational issues regarding estimation of G and propose different parametrizations that take care of constraints and that also reduce computational costs. The authors, for example, show that by just profiling the likelihood (1.7) with respect to G, we may end up with a very large array, which is of the dimension of $\boldsymbol{X}$ or, in the worst case where $\boldsymbol{X}$ has no replicated data or is continuous, of the same size of the data. For a discrete response, however, they show that a great reduction in dimensionality can be obtained. If $\boldsymbol{X}$ is also discrete and K is the multiplicity of $\boldsymbol{Y}$, it can be shown that the problem is now reduced to $(K-1)$-dimensions, which is usually much smaller than the dimension of $\boldsymbol{X}$.

Equivalence between Empirical likelihood and Scott and Wild's approach

Scott and Wild's approach and the Empirical Likelihood method for dealing with G seem to be essentially the same except for the constraints used to maximize equation (1.7) with respect to $g(\mathbf{x}_{2i}|\mathbf{x}_{1i})$. The latter method uses an extra constraint regarding the expected value of $f(\mathbf{y}_i|\mathbf{x}_{1i})$ and so it is worth verifying if both methods are essentially the same. That is, we want to check if this extra constraint is already satisfied using Scott and Wild's approach.

Notice that the likelihood used by Wang and Zhou (2006) reduces to the one used by Scott and Wild (2006) when $n_0 = 0$ (no SRS sample) and no auxiliary variable. In such a case, the restriction (1.9) obtained by Wang and Zhou becomes

$$\frac{1}{n}\sum_{i\in V}\left(\frac{f(\mathbf{y}_1|\mathbf{x}_i)-h_1}{1+\lambda\{f(\mathbf{y}_1|\mathbf{x}_i)-h_1)}\right\}=0, \tag{1.11}$$

or, after some algebra,

$$\frac{1}{n}\sum_{i\in V}\left(\frac{f(\mathbf{y}_1|\mathbf{x}_i)-h_1}{\frac{n_1}{nh_1}f(\mathbf{y}_1|\mathbf{x}_i)+\frac{n_2}{nh_2}f(\mathbf{y}_2|\mathbf{x}_i)}\right)=0, \tag{1.12}$$

where $h_i = \text{pr}(\mathbf{y} = i)$, $i = 1, 2$, n is the number of individuals selected for phase-2 and $f(\mathbf{y}_i|\mathbf{x}_i) = \text{pr}(\mathbf{y} = 1|\mathbf{x}_i)$. We want to show that this equation is already satisfied by Scott and Wild's approach.

Scott and Wild maximize the likelihood (1.10) with respect to $p_j = g(\mathbf{x}_j)$ subject to the constraint $\sum_j p_j = 1$, obtaining the restriction

$$\sum_j \left(\frac{n}{\mu_i} f(\mathbf{y}_i|\mathbf{x}_j) \frac{1}{\sum_l \mu_i f(\mathbf{y}_i|\mathbf{x}_j)} \right) = n_i \quad \text{where} \quad \mu_i = \frac{n_i}{\sum_j p_j f(\mathbf{y}_i|\mathbf{x}_j)}. \tag{1.13}$$

So, for $i = 1$, we have that

$$\sum_j \left(n \frac{n_1}{h_1} f(\mathbf{y}_1|\mathbf{x}_j) \frac{1}{\sum_l \frac{n_l}{h_l} f(\mathbf{y}_l|\mathbf{x}_j)} \right) = n_1$$

which reduces to

$$\sum_j \left(\frac{f(\mathbf{y}_1|\mathbf{x}_i) - h_1}{\frac{n_1}{h_1} f(\mathbf{y}_1|\mathbf{x}_j) + \frac{n_2}{h_2} f(\mathbf{y}_2|\mathbf{x}_j)} \right) = 0.$$

This is equivalent to (1.12). Scott and Wild's method is therefore equivalent to the empirical likelihood method.

In general, if the outcomes of interest $\boldsymbol{Y}$ and $\boldsymbol{X}_1$ are both discrete, most methods described before can be fully or nearly fully efficient. If $\boldsymbol{X}_1$ is continuous, the usual approach is to discretize it into different categories and use the exact values only at the second phase of the study. This clearly leads to losses of efficiency and is slight worse than using the kernel approach discussed earlier after equation (1.6) (Chatterjee and Chen, 2007).

Next we consider another two other methods that have an intuitive appeal: the Horvitz-Thompson estimator and the calibration method. Both methods belong to a much broader class of estimators, defined by Robins et al. (1994) as the *Augmented*

Inverse-Probability Weighted (AIPW) estimators, previously discussed. Both methods are commonly used in survey sampling and will also be considered throughout this thesis.

1.5.3 Weighted method

The weighted method, also known as the Horvitz-Thompson method (Horvitz and Thompson, 1952) or the inverse probability weighting IPW, is a common choice in practical use because it is easy to implement and robust under model misspecification, in the sense that it estimates the same quantities that we would be estimating if applying the same models to data from the full cohort.

This method consists in weighting each unit by the inverse of its probability of being selected for full observation. The weights can be seen as the number of times that each sampled unit should be replicated to represent the entire population and are usually known by design. For a 2-phase study the pseudo-loglikelihood function can be written as

$$l_w(\boldsymbol{\beta}) = \sum_{i:R_i=1} \frac{1}{\pi_i} \log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\beta})$$

The weighted approach is also commonly used in multi-phase studies since it can be easily generalized. For example, consider a study with $J+1$ phases. The probability of a unit being selected for full observation is $\pi_i = \pi_{1i} \times \pi_{2i} \times \cdots \times \pi_{Ji}$, $i = 1,\ldots,n$, where π_{ji} is the probability of the ith subject currently in phase j being selected for phase $j+1$, $j = 1,\ldots,J$. For a 2-phase study, $J = 1$ and $\pi_i = \pi_{1i}$.

To make inferences about $\boldsymbol{\beta}$, the parameter of interest, we have to solve the score function

$$\boldsymbol{S}_w(\boldsymbol{\beta}) = \sum_{i:R_i=1} \frac{1}{\pi_i} \frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\beta}) = \sum_i \frac{R_i}{\pi_i} \boldsymbol{S}_i(\boldsymbol{\beta}).$$

Note that this method uses only the completely observed data ignoring the incomplete ones and so it can give quite inefficient estimates in some cases.

1.6 Calibration

Calibration is a common technique in survey sampling when the survey is affected by non-response and estimation of population totals is required (Estevao and Sarndal, 2006). Calibration has gained popularity in real applications because it relies on natural constraints and provides estimates that are easy to interpret and approximately unbiased. It makes use of auxiliary information to adjust the sampling weights, providing more precise estimates of $\boldsymbol{\beta}$, the parameter of interest.

Calibration is more easily understood in terms of the estimation of population totals. Here we are interested in estimating the population total of a variable y, which was only observed for individuals in a selected sample. Let, as before, R_i be an indicator variable equals to 1 if the ith was selected and 0 otherwise.

The Horvitz Thompson estimator, here denoted by T_H, is

$$\hat{T}_H = \sum_{i:R_i=1} \frac{1}{\pi_i} y_i.$$

It gives unbiased but often inefficient estimates of the population total $T = \sum_i y_i$. The calibration method uses an auxiliary variable X which is related to Y and whose population total is known to modify the weights so that the weighted sample-total of X-values gives the known population total exactly. That is, the *calibration weights* h_i must satisfy the *calibration constraints*

$$\sum_{i:R_i=1} w_i \mathbf{x}_{1i} = \sum_i \mathbf{x}_{1i}, \tag{1.14}$$

where $w_i = h_i/\pi_i$. The calibration weights make the total estimated and known population totals for X agree. As discussed by Lumley (2010), h_i will give more weight to cases where the weighted sample total is too small and downweight large values of X when the weighted sample total is too large. The calibration estimator $\hat{T}_C$ is then given by

$$\hat{T}_C = \sum_{i:R_i=1} w_i y_i \tag{1.15}$$

and is related to the Horvitz-Thompson estimator as

$$\hat{T}_C = \hat{T}_H + \sum_{i:R_i=1} (w_i - d_i) y_i.$$

Here d_i is the design weight $1/\pi_i$ and calculation of the w_is has yet to be discussed. Since the weighted estimator is unbiased, the resulting bias for the calibrated method is given by

$$E(\hat{T}_c) - Y = E\left(\sum_{i:R_i=1} (w_i - d_i) y_i \right).$$

Thus, provided that there are small deviations between the calibrated weights and the design weights, the resulting estimates are nearly unbiased. Therefore, the goal is to obtain $\{w_i\}$ to minimize some distance function, $Q(w_i, d_i)$, such that (Sarndal, 2007), for every $w_i > 0$, $i = 1, \ldots, N$,

- $Q(w_i, d_i) > 0$;
- $Q(d_i, d_i) = 0$;
- $Q(w_i, d_i)$ is strictly convex;
- $q(w_i, d_i) = \partial Q(w_i, d_i)/\partial w_i$, where q is continuous.

Different distance functions lead to different sets of calibration weights. The function $Q(w_i, d_i) = \sum_i (w_i - d_i)^2/2 d_i v_i$, for instance, where v_i is a positive scale factor, leads

to so-called regression estimator method. The calibration constraints are satisfied for any choice of positive scale factors v_i, but different v_is lead to different estimators. For example, if x_i is a positive scalar and if $v_i = 1/x_i$, we get the ratio estimator (Sarndal, 2007).

Some estimators may lead to negative weights, which can be avoided by using raking or post-stratification. If variables with known population totals were not used to stratify the study population, post-stratification adjusts the design weights d by correcting group sizes as they would be in stratified sampling. If there is more than one variable, however, we may not be able to perform post-stratification since a cross-classification of the variables would be required, generating many groups and increasing the chance of having no subject belonging from a specific group. We are then unable to adjust the weights. Raking solves this problem by post-stratifying on each set of variables at a time and repeating this process until the weights stop changing. It allows multiple grouping variables to be used without constructing a complete cross-classification (Lumley, 2010).

Each method is associated with a specific distance function and specific calibrated weights. As point out by Lumley (2010), these methods, as well as linear regression, have been used before the theory was formulated, but it is interesting to note that calibration encompasses not only these methods but many more. Further discussion regarding different distance functions and their relations can be found in Deville and Sarndal (1992).

1.6.1 Calibration for a 2-phase study

Calibration can also be applied to 2-phase problems where partial information is available for a finite population and full observations are provided for only a sample of

units (Deville and Sarndal, 1992; Breslow et al., 2009).

As noted by Lumley et al. (2011), the regression estimator

$$\hat{T}_{yreg} = \sum_i \frac{R_i}{\pi_i} y_i + \left(1 - \frac{R_i}{\pi_i}\right) \mathbf{x}_{1i}\hat{\boldsymbol{\beta}},$$

results in the augmented inverse-probability weighted (AIPW) estimator proposed by Robins et al. (1994) (see equation (1.3)) if y_i is replaced by $S_i(\boldsymbol{\beta})$. The optimal choice for the auxiliary variable which minimizes the variance of the parameter estimates is $E\{\boldsymbol{S}(\boldsymbol{\beta})|\mathbf{y},\mathbf{x}_1\}$, as given by equation (1.4), where $\boldsymbol{Y}$ and $\boldsymbol{X}_1$ are variables known for every subject in the study. However, since these optimal auxiliary variables depend on the marginal distribution of $\boldsymbol{X}$, they are not generally available and must be estimated. Breslow et al. (2009) describe a plug-in method for approximating the conditional expectation credited to Kulich and Lin (2004), which works as follows:

1. The first step consists in developing predictive models for each missing variable given variables known for all, using a linear or logistic regression model fitted by IPW to the phase two data.

2. Using the models from step 1, estimated values for the missing variables are imputed from the phase-1 data, resulting in a "complete" dataset formed by the estimated and the true values.

3. The model $f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\beta})$ is fitted to the "complete" dataset.

4. Use the influence functions from the model fitted in step three as auxiliary variables in calibration. These are the estimates of the optimal auxiliary variables discussed before and are used to construct the calibration equations. Adjusted weights $w_i s$ are function Q obtained by minimizing a given distance

5. Estimate $\boldsymbol{\beta}$ from the phase two data using the adjusted weights w_i obtained in step four.

Note that only step three requires iteration. Unlike previous methods, no iteration is required to estimate $\boldsymbol{\beta}$, since the parameters are estimated after the adjustment of the weights has been done.

1.7 Outline

Most of the work proposed in the literature for missing data is concerned with discrete response or discrete phase-1 variables. In this thesis we propose a new method that is more flexible in the sense that it does not require a discrete distribution for the fully observed variables. Unlike most methods that discard extra cheap and easy to observe variables that were not used for selecting the final sample nor fitting the model of interest, our proposed method makes use of all available information in a fairly simple way. It may also lead to large gains in efficiency. Zhao et al. (2012), for example, show that cheap surrogates can substantially improve efficiency of a stratified 2-phase design and conclude that "developing computation software for the stratified 2-phase sampling design is valuable". Our work attempts to overcome this issue.

The flexibility of the proposed method allows it to be applied to non-response problems, under the MAR assumption, and more generally to situations where the phase-2 sample was not obtained from a stratified sampling scheme. That is, the proposed method allows us to fit a saturated model for the selection probability, making use of the entire information available in the study.

This thesis is set out as follows. In chapter 2 we discuss a semiparametric estimator due to Scott and Wild (2011). Their work was restricted to the discrete response

problem and the authors did not discuss asymptotic properties of the proposed estimator. In this chapter we extend Scott and Wild (2011) approach, deriving unbiased estimating equations for both discrete and continuous responses and we conclude this chapter by deriving asymptotic results of the proposed estimator.

In chapter 3 we discuss the proposed method for 2 and 3-phase problems, assuming that the response is binary. We perform a more detailed analysis than previously done by Scott and Wild (2011), using simulated data to compare its performance with that of several commonly-used estimators on a variety of scenarios that had not been previously considered. We also discuss its performance under the assumption of correctly specified and misspecified models. Finally, the proposed method is used, for the first time, to analyse a real dataset from the Women Health Initiative study (Rossouw et al., 2002).

In chapter 4 we assume that the response $\boldsymbol{Y}$ is continuous, following a generalized normal, skew-normal or a T-distribution, and study its efficiency through simulations. Its robustness is also analysed through simulations, but a more theoretical approach is also considered. If all models are correctly specified, likelihood based methods can be fully efficient but slight model misspecifications lead to increasing bias. Thus, following Lumley (2013) approach, we compare the proposed method against the best AIPW (Robins et al., 1994) estimator using nearly-correct model, so that we could obtain a threshold where the more robust AIPW estimator becomes more efficient than the proposed one.

Chapter 5 discusses the similarities and differences between the proposed method and the propensity score approach. We start by reviewing how propensity scores have been applied in the literature and conclude the chapter by combining both approaches in order to increase robustness.

Chapter 6 generalizes previous chapters by considering a broader sampling scheme.

Here we allow the response $\boldsymbol{Y}$ to be partially observed, but a correlated variable $\boldsymbol{V}$ to be fully observed for all individuals. Unlike in Neuhaus et al. (2006) and Jiang et al. (2006) where both $\boldsymbol{Y}$ and $\boldsymbol{V}$ were considered discrete, we allow each or both variables to be continuous. We develop estimating equations for the more general sampling scheme and conclude this chapter by analyzing another real dataset.

In chapter 7 we study the efficiency of the proposed method for both discrete and continuous responses. We start by showing its equivalence to the Scott and Wild (1997) method for binary responses, concluding that the proposed method is thus fully efficient in this case. We also derive the efficiency bounds for any semiparametric estimator and use it to study the asymptotic efficiency of the proposed method in different scenarios.

In chapter 8 we give a general conclusion of this dissertation and discuss some potential topics to extended the propose methods in future research.

2

Conditional Maximum Likelihood

In this chapter we study another technique to deal with missing response-selective data, the so-called conditional maximum likelihood method (CML), which is often a simple and efficient way to deal with missing information. Some advantages are its easy implementation, especially for the important binary case, and the fact that it can be fully efficient if some assumptions are satisfied.

We start by presenting the CML method and discussing the extensions made by Scott and Wild (2011). The proposed method is shown to be very flexible, making use of information that is usually discarded by more common approaches, and thus producing better estimates, as will be shown through simulations in chapters 3 and 4.

Later we extend Scott and Wild (2011) approach by first deriving unbiased estimating equations for not only the discrete response case, as done by the authors, but also for the continuous case, and by deriving asymptotic properties for the proposed estimator.

2.1 Introduction

Suppose that we observe a response Y_i, $i = 1, 2, \ldots, N$, generated from the joint distribution $f(\mathbf{x}, y) = f(y|\mathbf{x}; \boldsymbol{\beta})g(\mathbf{x})$, where $\boldsymbol{X}$ is a set of covariates and N is the population size. The disease-status Y ($Y = 1$ denoting a case and $Y = 2$, a control) and exposure history are observed for all N individuals (know as phase-1 data). These individuals are further classified into different strata S_j, $j = 1, \ldots, J,$, based on the disease/exposure combination and samples (known as phase-2 sample) of size n_{ij} are taken from each stratum and the remaining information is measured. Note that only the phase-2 sample was fully observed. Our goal here is to predict the response Y in terms of the covariates $\boldsymbol{X}$.

To this end, Breslow and Cain (1988) suggest estimating the regression coefficients $\boldsymbol{\beta}$ using an adaptation and extension of the conditional likelihood method of Manski and McFadden (1981). The authors work with a pseudo-likelihood function based on the conditional probability of being a case (i=1) or control (i=0) given that a member from stratum j, with regression variables $\boldsymbol{X}$, was selected for the phase-2 of the study. Using Bayes theorem, this conditional probability can be rewritten as

$$\frac{\frac{n_{ij}}{n} \operatorname{pr}(\mathbf{x}|S = j, Y = i, \text{ sampled at phase-2})}{\sum_l \frac{n_{lj}}{n} \operatorname{pr}(\mathbf{x}|S = j, Y = l, \text{ sampled at phase-2})}. \tag{2.1}$$

Wild (1991) shows that the resulting likelihood from (2.1) is actually a conditional likelihood not for the standard case-control sampling but for a slightly different conditional sampling scheme. Instead of taking samples of fixed size n_i from each group, as in Breslow and Cain (1988), we first choose $Y = i$ with probability n_i/n, where $n = \sum_i n_i$, and then sample $\boldsymbol{X}$ from $\operatorname{pr}(\mathbf{x}|\, Y = i)$, resulting in samples of random rather than fixed size. The likelihood utilized by Breslow and Cain (1988) can be obtained by noticing

that they considered the finite population also to be a case-control sample rather than a prospective one.

Note that the probability (2.1) considers only the completely observed data and that the selection probability was estimated by the ratio of elements in stratum (i,j) among all individuals. By Bayes theorem, a more general equation can then be written as

$$f_c = \text{pr}(\, Y_i = 1 | R_i = 1, \mathbf{x}_i; \boldsymbol{\beta}) = \frac{\pi(y_i, \mathbf{x}_i) f(y_i | \mathbf{x}_i; \boldsymbol{\beta})}{\sum_l \pi(y_l, \mathbf{x}_i) f(y_l | \mathbf{x}_i; \boldsymbol{\beta})}, \quad l = 0, 1 \tag{2.2}$$

where

$$\pi(y_i, \mathbf{x}_i) = \text{pr}(R_i = 1 | y_i, \mathbf{x}_i)$$

and R_i is an indicator variable that equal to 1 if the ith unit has been selected into the next phase of the study or 0 otherwise. Here, and for everything that follows, we assume that each unit is independently selected for inclusion into the next phase of the study (i.e., R_i for $i = 1, \ldots, N$ are independent) and that each individual has a positive probability of being selected for further observation (i.e., $\pi(y, \mathbf{x}) > 0$).

Notice that in the important case of a logistic regression model, $\text{logit}\{\text{pr}(Y = y_i | \mathbf{x}_i; \boldsymbol{\beta})\} = \mathbf{x}\boldsymbol{\beta}$, the conditional probability still follows a logistic distribution, but with an offset added. That is, for a binary response,

$$\text{pr}(\, Y_i = 1 | R_i = 1, \mathbf{x}_i; \boldsymbol{\beta}) = \frac{e^{o_i + \mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{o_i + \mathbf{x}_i^T \boldsymbol{\beta}}}, \tag{2.3}$$

where o_i is an offset given by

$$o_i = \log \left(\frac{\pi(\, Y_i = 1, \mathbf{x}_{1i})}{\pi(\, Y_i = 0, \mathbf{x}_{1i})} \right) \tag{2.4}$$

where $\boldsymbol{X}_1$ is a covariate fully observed at phase-1. The remaining covariate $\boldsymbol{X}_2$ is

observed only at phase-2. It was assumed that $\pi(y, \mathbf{x}) = \pi(y, \mathbf{x}_1)$, i.e., the selection probability depends only on variables fully observed at phase-1. The selection probability π in many cases is controlled by the researcher and so inferences can be made by using standard logistic regression software that allows the inclusion of offsets. We can estimate the parameters of interest by solving

$$\boldsymbol{S}_0(\boldsymbol{\beta}, \pi) = \sum_i S_{0i} = \frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \pi) = 0, \tag{2.5}$$

for $\boldsymbol{\beta}$, where $\boldsymbol{S}_0$ is the score function and $\ell = \log \prod_i f$, the log-likelihood.

Hsieh et al. (1985) showed that, with fixed π, the estimates are consistent and asymptotically normal with variance $ACov$ given by

$$ACov(\boldsymbol{\beta}) = \boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{C}}_{00} \boldsymbol{\mathcal{I}}_{00}^{-1} \tag{2.6}$$

where $\boldsymbol{\mathcal{I}}_{00ij} = E\{-\partial \boldsymbol{S}_{0i} / \partial \boldsymbol{\beta}_j^T\}$, the element (i, j) of the information matrix $\boldsymbol{\mathcal{I}}_{00}$, and $\boldsymbol{\mathcal{C}}_{00ij} = Cov\{\boldsymbol{S}_{0i}, \boldsymbol{S}_{0j}\}$.

2.2 Selection probabilities unknown

Suppose now that we are interested in estimating the selection probability π. Robins et al. (1994) showed that gains in efficiency can be obtained by estimating selection probabilities π even if they are known by design. This result is not as paradoxical as it seems; by estimating the πs we are actually using extra information from the incomplete data that are not used when the true fixed-by-design probabilities are used. As noticed by Lumley et al. (2011), we introduce some error while estimating these probabilities, but the gain in precision is large enough to overcome this error so that this approach will be at least as efficient as using the true selection probabilities. This

result is particularly interesting for the weighted and conditional likelihood methods since both methods estimate $\boldsymbol{\beta}$ by solving the score equations of the form

$$\boldsymbol{S}_0(\boldsymbol{\beta},\pi) = \sum_i R_i W_i(\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\beta},\pi) = \sum_i S_{0i}, \tag{2.7}$$

where

$$W_i(\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\beta},\pi) = \begin{cases} \frac{1}{\pi_i}\frac{\partial}{\partial\boldsymbol{\beta}}\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\beta}), & \text{for the weighted method} \\ \frac{\partial}{\partial\boldsymbol{\beta}}\log f(\mathbf{y}_i|R_i=1,\mathbf{x}_i;\boldsymbol{\beta},\pi), & \text{for the conditional likelihood method} \end{cases}$$

Both approaches depend on the selection probability π and their efficiencies can be improved by estimating the selection probabilities.

2.2.1 Modelling the selection probabilities

To facilitate discussion of estimating selection probabilities, we assume a parametric model $\pi_i(\boldsymbol{\alpha}) = \pi(\mathbf{x}_{1i},\mathbf{y}_i;\boldsymbol{\alpha})$, where $\boldsymbol{X}_1$ and $\boldsymbol{Y}$ are observed for all phase-1 population so that $\boldsymbol{\alpha}$ can be estimated from the completely-observed data. The remaining variable $\boldsymbol{X}_2$ is only observed at phase-2.

We now need to estimate both the parameter of interest $\boldsymbol{\beta}$ and also $\boldsymbol{\alpha}$, the parameter of the selection model π. We do this by combining the estimating equation (2.7) with another estimating equation for $\boldsymbol{\alpha}$. This allows us to incorporate the uncertainty in the estimation of $\boldsymbol{\alpha}$ when estimating the variance of $\hat{\boldsymbol{\beta}}$. We further assume that units are selected for full observation independently. Since a unit is selected for full observation according to a Bernoulli distribution with parameter $\pi_i(\boldsymbol{\alpha})$, its score function $\boldsymbol{S}_1$ can

be written as

$$\begin{aligned} \boldsymbol{S}_1(\boldsymbol{\alpha}) &= \frac{\partial}{\partial \boldsymbol{\alpha}} \sum_i [R_i \log \pi_i + (1-R_i)\log(1-\pi_i)] \\ &= \sum_i \left[\frac{(R_i - \pi_i)}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \boldsymbol{\alpha}} \right] \\ &= \sum_i S_{1i}, \end{aligned} \tag{2.8}$$

and we estimate $\boldsymbol{\alpha}$ by solving $\boldsymbol{S}_1(\boldsymbol{\alpha}) = 0$. The resulting combined estimating equations are then given by

$$\boldsymbol{S}(\boldsymbol{\phi}) = \boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{pmatrix} \boldsymbol{S}_0(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ \boldsymbol{S}_1(\boldsymbol{\alpha}) \end{pmatrix} \tag{2.9}$$

and estimates $\hat{\boldsymbol{\phi}}^T = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\alpha}}^T)$ are obtained by setting $\boldsymbol{S}(\hat{\boldsymbol{\phi}}) = 0$. The resulting estimators are consistent and asymptotically normal under mild regularity conditions (Scott and Wild, 2011). The asymptotic covariance $ACov(\boldsymbol{\phi})$ matrix is given by

$$ACov(\boldsymbol{\phi}) = \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\mathcal{C}} \left(\boldsymbol{\mathcal{I}}^T \right)^{-1}.$$

Here,

$$\boldsymbol{\mathcal{I}} = E\left\{ \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{\phi}} \right\} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{00} & \boldsymbol{\mathcal{I}}_{01} \\ 0 & \boldsymbol{\mathcal{I}}_{11} \end{pmatrix},$$

and

$$\boldsymbol{\mathcal{C}} = Cov\{\boldsymbol{S}\} = \begin{pmatrix} \boldsymbol{\mathcal{C}}_{00} & \boldsymbol{\mathcal{C}}_{01} \\ \boldsymbol{\mathcal{C}}_{10} & \boldsymbol{\mathcal{C}}_{11} \end{pmatrix}.$$

The asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by

$$ACov(\hat{\boldsymbol{\beta}}) = \boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{C}}_{00} \boldsymbol{\mathcal{I}}_{00}^{-1} - \boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{C}}_{01} \boldsymbol{\mathcal{I}}_{11}^{-1} \boldsymbol{\mathcal{C}}_{01}^T \boldsymbol{\mathcal{I}}_{00}^{-1}. \tag{2.10}$$

Expression (2.10) can be simplified using the following result due to Scott and Wild (2011)

Result 2.1 (Scott and Wild (2011))**.** *Let* $\boldsymbol{S}_0$ *and* $\boldsymbol{S}_1$ *be defined as in equations* (2.7) *and* (2.8)*, respectively. Then,* $Cov(\boldsymbol{S}_0, \boldsymbol{S}_1) = \boldsymbol{\mathcal{I}}_{01}$*, where* $\boldsymbol{\mathcal{I}}_{01} = -\partial \boldsymbol{S}_0 / \partial \boldsymbol{\alpha}_1$*.*

Applying result 2.1, (2.10) becomes

$$ACov(\hat{\boldsymbol{\beta}}) = \boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{C}}_{00} \boldsymbol{\mathcal{I}}_{00}^{-1} - \boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{I}}_{01} \boldsymbol{\mathcal{I}}_{11}^{-1} \boldsymbol{\mathcal{I}}_{01}^{T} \boldsymbol{\mathcal{I}}_{00}^{-1}. \tag{2.11}$$

The first term of the equation (2.11) is obtained when the known π_is are used (see equation (2.6)) and so the second one represents the effect of estimating the selection probability. The last term is non-negative definite which means that the asymptotic variance will be, at worst, equal to the one obtained when the known selection probabilities are used. In other words, we are better off estimating the π_is instead of using the true values, even if they are known by design.

Insights about the role of including additional variables in the selection model follow from writing equation (2.11) as

$$ACov(\hat{\boldsymbol{\beta}}) = \boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{C}}_R \boldsymbol{\mathcal{I}}_{00}^{-1}, \quad \text{where} \quad \boldsymbol{\mathcal{C}}_R = \boldsymbol{\mathcal{C}}_{00} - \boldsymbol{\mathcal{I}}_{01} \boldsymbol{\mathcal{I}}_{11}^{-1} \boldsymbol{\mathcal{I}}_{01}^{T}. \tag{2.12}$$

As noticed by Scott and Wild (2011), $\boldsymbol{\mathcal{C}}_R$ is the covariance matrix of the residual vector when $\boldsymbol{S}_0(\boldsymbol{\phi})$ is regressed on $\boldsymbol{S}_1(\boldsymbol{\phi})$, i.e., $\boldsymbol{\mathcal{C}}_R = \inf_{\boldsymbol{B}} Cov\{\boldsymbol{S}_0 - \boldsymbol{B}\boldsymbol{S}_1\}$. The reduction depends not on the effect of estimating π, but on the predictive relationship between the score function from the selection model and $\boldsymbol{S}_0$. So, if $\boldsymbol{X}$ can take only a finite number of values, the optimum solution would be to include a saturated model of all variables into the selection model and this can be done if sample sizes are large enough so that over-parametrization is not an issue.

These results can be extended to a three-phase sampling problem. Here, partial information on the response variable as well as some covariates, $\boldsymbol{X}_1$, say, are known from all members, termed the phase-1 sample, constructed or treated as a sample of an infinite population. A sample is then taken from the phase-1 data and additional information, $\boldsymbol{X}_2$, say, is measured. A phase-3 subsample is finally taken from the phase-2 sample and the remaining information is collected. Note that there are now two selection probabilities to be estimated (phase-1 $\stackrel{R_1=1}{\longrightarrow}$ phase-2 $\stackrel{R_2=1}{\longrightarrow}$ phase-3) and only individuals with $R_1R_2=1$ will be fully observed. As before, we fit a parametric model for $\text{pr}(R_1=1|\mathbf{y},\mathbf{x}_1;\boldsymbol{\alpha}_1)$, which we will write as $\pi_1(\boldsymbol{\alpha}_1)=\pi_1(\mathbf{y},\mathbf{x}_1;\boldsymbol{\alpha}_1)$ and for $\text{pr}(R_2=1|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,R_1=1;\boldsymbol{\alpha}_2)$, which we will write as $\pi_2(\boldsymbol{\alpha}_2)=\pi_2(\mathbf{y},\mathbf{x}_1,\mathbf{x}_2;\boldsymbol{\alpha}_2)$. Fitting these two binary regression models results in score functions

$$\boldsymbol{S}_1(\boldsymbol{\alpha}_1)=\sum_i\left(\frac{R_{1i}-\pi_{1i}}{\pi_{1i}(1-\pi_{1i})}\frac{\partial\pi_{1i}}{\partial\boldsymbol{\alpha}_1}\right)\quad\text{and}\quad\boldsymbol{S}_2(\boldsymbol{\alpha}_2)=\sum_i\left(\frac{R_{2i}-\pi_{2i}}{\pi_{2i}(1-\pi_{2i})}\frac{\partial\pi_{2i}}{\partial\boldsymbol{\alpha}_2}\right).$$

The parameter $\boldsymbol{\phi}=(\boldsymbol{\beta}^T,\boldsymbol{\alpha}_1^T,\boldsymbol{\alpha}_2^T)^T$ can then be estimated by solving

$$\boldsymbol{S}(\boldsymbol{\phi})=S(\boldsymbol{\beta},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2)=\begin{pmatrix}\boldsymbol{S}_0(\boldsymbol{\beta},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2)\\ \boldsymbol{S}_1(\boldsymbol{\alpha}_1)\\ \boldsymbol{S}_2(\boldsymbol{\alpha}_2)\end{pmatrix}=\sum_i\boldsymbol{S}_i=0. \tag{2.13}$$

With the same argument as in result 2.1, we can also show that $\text{Cov}(\boldsymbol{S}_0,\boldsymbol{S}_2)=\boldsymbol{\mathcal{I}}_{02}$ and

$$\boldsymbol{\mathcal{I}}=\begin{pmatrix}\boldsymbol{\mathcal{I}}_{00}&\boldsymbol{\mathcal{I}}_{01}&\boldsymbol{\mathcal{I}}_{02}\\ 0&\boldsymbol{\mathcal{I}}_{11}&0\\ 0&0&\boldsymbol{\mathcal{I}}_{22}\end{pmatrix}.$$

The asymptotic covariance matrix can now be written as

$$ACov(\hat{\boldsymbol{\beta}}) = \boldsymbol{\mathcal{I}}_{00}^{-1}\boldsymbol{\mathcal{C}}_{00}\boldsymbol{\mathcal{I}}_{00}^{-1} - \boldsymbol{\mathcal{I}}_{00}^{-1}\boldsymbol{\mathcal{I}}_{01}\boldsymbol{\mathcal{I}}_{11}^{-1}\boldsymbol{\mathcal{I}}_{01}^{T}\boldsymbol{\mathcal{I}}_{00}^{-1} - \boldsymbol{\mathcal{I}}_{00}^{-1}\boldsymbol{\mathcal{I}}_{02}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{02}^{T}\boldsymbol{\mathcal{I}}_{00}^{-1}.$$

As before, the first term is obtained when the true values of π_is are used and the last two are due to estimating both selection probabilities. Again, both terms are non-negative and so it is always better to estimate the π_is rather than using the true values.

2.3 Additional information

In previous sections we estimated the parameter of interest $\boldsymbol{\beta}$ by setting $\boldsymbol{S}(\boldsymbol{\phi}) = 0$. Notice, however, that we have not used the entire information provided by the data. Since both the weighted and the conditional maximum likelihood methods depend on $\boldsymbol{\alpha}$ through π, the quantity

$$\widetilde{\boldsymbol{S}}_1 = \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \tag{2.14}$$

for the CML, where $fc = f(y_i|R_i = 1, \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\alpha})$, and

$$\widetilde{\boldsymbol{S}}_1 = \frac{\partial}{\partial \boldsymbol{\alpha}} \frac{\log f}{\pi} \tag{2.15}$$

for the weighted method, where $f = f(y_i|\mathbf{x}_i; \boldsymbol{\beta})$, can be used to improve our estimates. An important point that rises is how this extra information should be added. Scott and Wild (2011) use a linear combination, resulting in the estimating equation

$$\boldsymbol{S}_\lambda = \boldsymbol{S} + \lambda\widetilde{\boldsymbol{S}}, \quad \text{where} \quad \widetilde{\boldsymbol{S}} = \begin{pmatrix} \mathbf{0} \\ \widetilde{\boldsymbol{S}}_1 \end{pmatrix}, \tag{2.16}$$

for a 2-phase sampling scheme, where λ is assumed to be a scalar and $\boldsymbol{S}$ is given by (2.9). By checking for asymptotic efficiency, Scott and Wild (2011) showed that the optimum value of λ is -1. Their result, however, is only valid when the response is discrete. Here we provide a more general result, which is valid whether the outcome is discrete or continuous. We also provide asymptotic results that were not considered in Scott and Wild (2011). We will consider the two approaches, CML and the weighted method, separately.

2.3.1 CML method

Our goal here is to find λ that minimizes the asymptotic variance, allowing the response $\boldsymbol{Y}$ to be discrete or continuous. Semiparametric efficiency will then be discussed later in chapter 7.

By using the estimating equations $\boldsymbol{S}_\lambda = \boldsymbol{S} + \lambda\widetilde{\boldsymbol{S}}$, the information and covariance matrix are given by

$$\boldsymbol{\mathcal{I}}_\lambda = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{00} & \boldsymbol{\mathcal{I}}_{01} \\ \boldsymbol{\mathcal{I}}_{10} + \lambda\widetilde{\boldsymbol{\mathcal{I}}}_{10} & \boldsymbol{\mathcal{I}}_{11} + \lambda\widetilde{\boldsymbol{\mathcal{I}}}_{11} \end{pmatrix}, \tag{2.17}$$

where $\widetilde{\boldsymbol{\mathcal{I}}}_{10} = \partial\widetilde{\boldsymbol{S}}_1/\partial\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\mathcal{I}}}_{11} = \partial\widetilde{\boldsymbol{S}}_1/\partial\boldsymbol{\alpha}$, and

$$\boldsymbol{\mathcal{C}}_\lambda = \begin{pmatrix} Cov(\boldsymbol{S}_0, \boldsymbol{S}_0) & Cov(\boldsymbol{S}_1 + \lambda\widetilde{\boldsymbol{S}}_1, \boldsymbol{S}_0) \\ Cov(\boldsymbol{S}_1 + \lambda\widetilde{\boldsymbol{S}}_1, \boldsymbol{S}_0) & Cov(\boldsymbol{S}_1 + \lambda\widetilde{\boldsymbol{S}}_1, \boldsymbol{S}_1 + \lambda\widetilde{\boldsymbol{S}}_1) \end{pmatrix} \tag{2.18}$$

respectively. The asymptotic covariance matrix $ACov_\lambda$ is now given by $\boldsymbol{\mathcal{I}}_\lambda^{-1}\boldsymbol{\mathcal{C}}_\lambda\left(\boldsymbol{\mathcal{I}}_\lambda^T\right)^{-1}$ and the optimum value of λ is given by the following proposition.

Proposition 2.1. *The asymptotic covariance matrix $ACov_\lambda$ is minimized for $\lambda = -1$.*

Proof. For proof the proposition, we need two results.

Result 2.2. *Let* $f_c = f(\boldsymbol{y}|\boldsymbol{x}, R=1; \boldsymbol{\beta}, \boldsymbol{\alpha})$ *and* $\boldsymbol{S}_0$ *and* $\widetilde{\boldsymbol{S}}_1$ *be given as in* (2.5) *and* (2.14)*, respectively. Then,* $Cov\left(\boldsymbol{S}_0, \widetilde{\boldsymbol{S}}_1\right) = \boldsymbol{\mathcal{I}}_{01}$.

Proof. For the proof, first note that

$$\frac{\partial^2 \log f_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}} = \frac{\partial}{\partial \boldsymbol{\alpha}} \left(\frac{1}{f_c} \frac{\partial f_c}{\partial \boldsymbol{\beta}} \right) = \frac{-1}{f_c^2} \frac{\partial f_c}{\partial \boldsymbol{\alpha}} \frac{\partial f_c}{\partial \boldsymbol{\beta}} + \frac{1}{f_c} \frac{\partial^2 f_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}}. \tag{2.19}$$

Now,

$$\begin{aligned}
Cov\left(\boldsymbol{S}_0, \widetilde{\boldsymbol{S}}_1\right) &= E\left(\boldsymbol{S}_0 \widetilde{\boldsymbol{S}}_1 \middle| \mathbf{x}\right) \\
&= \int \frac{\partial \log f_c}{\partial \boldsymbol{\beta}} \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \operatorname{pr}(R=1, \mathbf{y}|\mathbf{x};)\, d\mathbf{y} \\
&= \int \pi \frac{\partial \log f_c}{\partial \boldsymbol{\beta}} \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} f(\mathbf{y}|\mathbf{x}; \boldsymbol{\beta})\, d\mathbf{y} \\
&= \int \pi \left(\frac{-\partial^2 \log f_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}} + \frac{1}{f_c} \frac{\partial^2 f_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}} \right) f(\mathbf{y}|\mathbf{x}; \boldsymbol{\beta})\, d\mathbf{y} \quad \text{(from equation (2.19))} \\
&= -\int \pi \frac{\partial^2 \log f_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}} f(\mathbf{y}|\mathbf{x}; \boldsymbol{\beta})\, d\mathbf{y} + \int \frac{\pi}{f_c} \frac{\partial^2 f_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}} f(\mathbf{y}|\mathbf{x}; \boldsymbol{\beta})\, d\mathbf{y}.
\end{aligned}$$

For the first term, we have that

$$\int \pi \frac{\partial^2 \log f_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}} f(\mathbf{y}|\mathbf{x}; \boldsymbol{\beta})\, d\mathbf{y} = E\left(R \frac{\partial^2 \log f_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}} \middle| \mathbf{x} \right) = -\boldsymbol{\mathcal{I}}_{01},$$

and so we just have to show that the second is zero. This follows because

$$\int \frac{\pi}{f_c} \frac{\partial^2 f_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}} f(\mathbf{y}|\mathbf{x}; \boldsymbol{\beta})\, d\mathbf{y} = \int \frac{\partial^2 f_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}} \frac{\operatorname{pr}(R=1|\mathbf{x})}{f(\mathbf{y}|R=1, \mathbf{x})} d\mathbf{y} = \pi(\mathbf{x}; \boldsymbol{\alpha}) \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}} \int f_c d\mathbf{y} = 0.$$

□

Result 2.3. $Cov\left(\boldsymbol{S}_1, \widetilde{\boldsymbol{S}}_1\right) = \widetilde{\boldsymbol{\mathcal{I}}}_{11}$.

Proof. Since

$$\mathbb{E}\left(\widetilde{\boldsymbol{S}}_1 \middle|\ \mathbf{x}\right) = \int \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \pi f\left(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}\right) d\mathbf{y} = 0,$$

we have that

$$\frac{\partial}{\partial \boldsymbol{\alpha}} \mathbb{E}\left(\widetilde{\boldsymbol{S}}_1 \middle|\ \mathbf{x}\right) = \int \frac{\partial^2 \log f_c}{\partial \boldsymbol{\alpha}^2} \pi f\left(y|\ \mathbf{x}\right) d\mathbf{y} + \int \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \frac{\partial \pi}{\partial \boldsymbol{\alpha}} f\left(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}\right) d\mathbf{y} = 0.$$

The first term of the right-hand side above can be written as

$$\int \frac{\partial^2 \log f_c}{\partial \boldsymbol{\alpha}^2} \pi f\left(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}\right) d\mathbf{y} = \mathbb{E}\left(R \frac{\partial^2 \log f_c}{\partial \boldsymbol{\alpha}^2}\right) = \mathbb{E}\left(\frac{\partial \widetilde{S}_1}{\partial \boldsymbol{\alpha}} \middle|\ \mathbf{x}\right) = -\widetilde{\boldsymbol{\mathcal{I}}}_{11},$$

and the second one, as

$$\begin{aligned}
\int \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \frac{\partial \pi}{\partial \boldsymbol{\alpha}} f\left(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}\right) d\mathbf{y} &= \int \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \frac{1}{\pi} \frac{\partial \pi}{\partial \boldsymbol{\alpha}} \pi f\left(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}\right) d\mathbf{y} \\
&= \mathbb{E}\left(R \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \frac{1}{\pi} \frac{\partial \pi}{\partial \boldsymbol{\alpha}} \middle|\ \mathbf{x}\right) \\
&= \mathbb{E}\left(R \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \left(\frac{R}{\pi} - \frac{1-R}{1-\pi}\right) \frac{\partial \pi}{\partial \boldsymbol{\alpha}} \middle|\ \mathbf{x}\right) \\
&= \mathbb{E}\left(\boldsymbol{S}_1 \widetilde{\boldsymbol{S}}_1 \middle|\ \mathbf{x}\right)
\end{aligned}$$

Hence,

$$Cov\left(\boldsymbol{S}_1, \widetilde{\boldsymbol{S}}_1\right) = \widetilde{\boldsymbol{\mathcal{I}}}_{11}.$$

□

From results 2.2 and 2.3, we can rewrite the covariance matrix (2.18) as

$$\boldsymbol{\mathcal{C}}_\lambda = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{00} & \boldsymbol{\mathcal{I}}_{01} + \lambda \boldsymbol{\mathcal{I}}_{01} \\ \boldsymbol{\mathcal{I}}_{01}^T + \lambda \boldsymbol{\mathcal{I}}_{01}^T & \boldsymbol{\mathcal{I}}_{11} + \lambda \left(2\widetilde{\boldsymbol{\mathcal{I}}}_{11} + \lambda \widetilde{\boldsymbol{\mathcal{I}}}_{11}\right) \end{pmatrix}.$$

The inverse of $\boldsymbol{\mathcal{I}}_\lambda$ (see equation (2.17)) is

$$\boldsymbol{\mathcal{I}}_\lambda^{-1} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{00}^* & \boldsymbol{\mathcal{I}}_{01}^* \\ \boldsymbol{\mathcal{I}}_{10}^* & \boldsymbol{\mathcal{I}}_{11}^* \end{pmatrix}$$

where

$$\boldsymbol{\mathcal{I}}_{00}^* = \boldsymbol{\mathcal{I}}_{00}^{-1} + \lambda \boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{I}}_{01} \boldsymbol{\mathcal{A}}^{-1} \widetilde{\boldsymbol{\mathcal{I}}}_{10} \boldsymbol{\mathcal{I}}_{00}^{-1},$$

$$\boldsymbol{\mathcal{I}}_{01}^* = -\boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{I}}_{01} \boldsymbol{\mathcal{A}}^{-1},$$

$$\boldsymbol{\mathcal{I}}_{10}^* = -\lambda \boldsymbol{\mathcal{A}}^{-1} \widetilde{\boldsymbol{\mathcal{I}}}_{10} \boldsymbol{\mathcal{I}}_{00}^{-1},$$

$$\boldsymbol{\mathcal{I}}_{11}^* = \boldsymbol{\mathcal{A}}^{-1}$$

and

$$\boldsymbol{\mathcal{A}} = -\lambda \widetilde{\boldsymbol{\mathcal{I}}}_{10} \boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{I}}_{01} + \boldsymbol{\mathcal{I}}_{11} + \lambda \widetilde{\boldsymbol{\mathcal{I}}}_{11}.$$

The upper left-hand block of $\boldsymbol{\mathcal{I}}_\lambda^{-1} \boldsymbol{\mathcal{C}}_\lambda \left(\boldsymbol{\mathcal{I}}_\lambda^T \right)^{-1}$ is then equal to

$$\begin{aligned} &\boldsymbol{\mathcal{I}}_{00}^* \boldsymbol{\mathcal{I}}_{00} \left(\boldsymbol{\mathcal{I}}_{00}^*\right)^T + (1+\lambda) \boldsymbol{\mathcal{I}}_{00}^* \boldsymbol{\mathcal{I}}_{01} \left(\boldsymbol{\mathcal{I}}_{01}^*\right)^T + (1+\lambda) \boldsymbol{\mathcal{I}}_{01}^* \boldsymbol{\mathcal{I}}_{10} \left(\boldsymbol{\mathcal{I}}_{00}^*\right)^T \\ &+ \boldsymbol{\mathcal{I}}_{01}^* \boldsymbol{\mathcal{I}}_{11} \left(\boldsymbol{\mathcal{I}}_{01}^*\right)^T + \lambda \boldsymbol{\mathcal{I}}_{01}^* (2\widetilde{\boldsymbol{\mathcal{I}}}_{11} + \lambda \widetilde{\boldsymbol{\mathcal{I}}}_{11}) \left(\boldsymbol{\mathcal{I}}_{01}^*\right)^T \end{aligned}$$

which is minimized with respect to λ. From this, we get $\lambda = -1$. □

The problem corresponds to maximizing the pseudo-loglikelihood

$$\ell(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_i R_i \log(f_c) - \sum_i \Big(R_i \log(\pi_i) + (1 - R_i) \log(1 - \pi_i) \Big). \tag{2.20}$$

We propose to estimate the parameters of interest by solving the estimating equations associated to the pseudo-likelihood (2.20) so that we use the extra information regarding $\widetilde{\boldsymbol{S}}$ (see equation (2.14)) in the most efficient way. We denote this new estimator by

CML+$\widetilde{\boldsymbol{S}}$.

2.3.2 Weighted method

For the weighted method, the score function $\boldsymbol{S}_0$ with respect to $\boldsymbol{\beta}$ is also a function of $\boldsymbol{\alpha}$. We could, in principle, use the extra information in the same way as used for the CML method. First, we obtain

$$\begin{aligned}\widetilde{\boldsymbol{S}}_1 &= \frac{\partial}{\partial \boldsymbol{\alpha}} \sum_i \left(\frac{R_i}{\pi_i} \log f(\mathbf{y}_i|\mathbf{x}_{1i};\boldsymbol{\beta}) \right) \\ &= -\sum_i R_i \left(\frac{1-\pi_i}{\pi_i} \right) \log f(\mathbf{y}_i|\mathbf{x}_{1i};\boldsymbol{\beta}) \mathbf{z}_i\end{aligned}$$

and the estimating equation for $\boldsymbol{\alpha}$ is now given by setting to zero

$$\begin{aligned}\boldsymbol{S}_1 + \lambda \widetilde{\boldsymbol{S}}_1 &= \sum_i (R_i - \pi_i)\, \mathbf{z}_i - \lambda \sum_i R_i \left(\frac{1-\pi_i}{\pi_i} \right) \log f(\mathbf{y}_i|\mathbf{x}_{1i};\boldsymbol{\beta}) \mathbf{z}_i \\ &= \sum_i \left[(R_i - \pi_i) - R_i \lambda \left(\frac{1-\pi_i}{\pi_i} \right) \log f(\mathbf{y}_i|\mathbf{x}_{1i};\boldsymbol{\beta}) \right] \mathbf{z}_i.\end{aligned}$$

The optimum value of λ is not necessarily the same as the one obtained before, but should be obtained in a similar way: by minimizing the asymptotic variance. To this end, we need equivalent results to results 2.2 - 2.3. These results, however, are not valid for the weighted case. That is, since

$$\mathbb{E}\left(\widetilde{\boldsymbol{S}}_1 | \mathbf{x} \right) = \int \frac{\partial}{\partial \boldsymbol{\alpha}} \left(\frac{\log f}{\pi} \right) \pi f\left(\mathbf{y}|\mathbf{x};\boldsymbol{\beta} \right) dy \neq 0,$$

neither results 2.2 nor 2.3 hold and so the optimum value of λ for the weighted method is not -1. In fact, there is no scalar and constant λ such that $\mathbb{E}\left(\boldsymbol{S}_1 + \lambda \widetilde{\boldsymbol{S}}_1 | \mathbf{x} \right) = 0$ and, for this reason, only the CML+$\widetilde{\boldsymbol{S}}$ method described previously will be considered for simulations in the following chapters.

2.3.3 Asymptotics

The proposed estimator, denoted by CML+$\widetilde{\boldsymbol{S}}$, takes into account, through $\widetilde{S}$ (see equation (2.14)), extra information provided by the data. We now show that this estimator is consistent and asymptotically normal, whether the distribution of $\boldsymbol{Y}$ is discrete or continuous. Let $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ be a vector with dimension p and let Φ be the parameter space. In addition to assuming the independence of R_i and $\pi_i(\mathbf{z};\boldsymbol{\alpha}) > 0$, for $i = 1,\ldots,N$, we also assume the following regularity conditions:

(A) Φ is compact and $\boldsymbol{\phi}_0$, the true value of $\boldsymbol{\phi}$, is an interior point of Φ. The covariate space is a compact subset of $\mathbb{R}^q$, for some integer $q \geq 1$.

(B) $f_c(\mathbf{y}|\mathbf{x};\boldsymbol{\phi})$ is continuous in $\boldsymbol{Y}$ and in $\boldsymbol{\phi}$, and strictly positive for all $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\phi})$. In addition, its first and second partial derivatives with respect to $\boldsymbol{\phi}$ exist and are continuous for all $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\phi})$.

(C) Interchanges of differentiation and integration of $f(\mathbf{y}|\mathbf{x};\phi)$ are valid for both first and second partial derivatives with respect to $\boldsymbol{\phi}$.

(D) $E\left[\dfrac{-\partial^2 \log f_c(\mathbf{y}|\mathbf{x};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T}\right]$ is finite and positive definite at $\boldsymbol{\phi}_0$.

(E) There exists a $\delta > 0$ such that

$$E\left[\sup_{A_\delta}\left|\frac{\partial^2 \log f(\mathbf{y}|\mathbf{x};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T}\right|\right] < \infty$$

for $A_\delta = \{\boldsymbol{\phi} \in \Phi : |\boldsymbol{\phi} - \boldsymbol{\phi}_0| \leq \delta\}$.

(F) The derivatives

$$\frac{\partial}{\partial\boldsymbol{\beta}_j}\int \pi(\mathbf{y},\mathbf{x}) f_j(\mathbf{y}|\mathbf{x}) d\mathbf{y}, \quad \text{for } j = 1,\ldots,p,$$

are linearly independent.

(G) Similarly, the derivatives

$$\frac{\partial}{\partial \boldsymbol{\alpha}_j} \int \pi(\mathbf{y},\mathbf{x}) f_j(\mathbf{y}|\mathbf{x}) d\mathbf{y}, \quad \text{for } j = 1,\ldots,p,$$

are also linearly independent.

Consistency

To show consistency we are going to use the following lemma presented in Lu (2009), which is due to Weaver (2001), a more general restatement of Foutz (1977).

Lemma 2.1. *Let $\{f_N(\boldsymbol{\phi})\}$ be a sequence of continuous random vector-valued functions of $\boldsymbol{\phi} \in \Phi \subset \mathbb{R}^p$. Suppose that, for all N, the partial derivatives of $f_N(\boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$ exist and are continuous in Φ; let $f'_N(\boldsymbol{\phi})$ be the $p \times p$ dimensional matrix containing these partial derivatives. Let $H(\boldsymbol{\phi})$ be a $p \times p$ dimensional matrix whose elements are continuous functions of $\boldsymbol{\phi}$ such that $H^{-1}(\boldsymbol{\phi}^*)$ exists for some $\boldsymbol{\phi}^* \in \Phi$. Suppose that $f'_N(\boldsymbol{\phi}) \xrightarrow{p} H(\boldsymbol{\phi})$ as $N \longrightarrow \infty$ uniformly for $\boldsymbol{\phi}$ in an open neighbourhood around $\boldsymbol{\phi}^*$. Furthermore, assume that $f_N(\boldsymbol{\phi}^*) \xrightarrow{p} 0$. Then, there exists a sequence $\{\hat{\boldsymbol{\phi}}_N\}$ such that*

$$\mathbb{P}\Big(f_N(\hat{\boldsymbol{\phi}}_N) = 0\Big) \longrightarrow 1, \quad as \quad N \longrightarrow \infty, \tag{2.21}$$

and

$$\hat{\boldsymbol{\phi}}_N \longrightarrow \boldsymbol{\phi}^*. \tag{2.22}$$

If another sequence $\bar{\boldsymbol{\phi}}_N$ also satisfies (2.21) *and* (2.22)*, then*

$$\mathbb{P}\Big(\hat{\boldsymbol{\phi}}_N = \bar{\boldsymbol{\phi}}_N\Big) \longrightarrow 1 \quad as \quad N \longrightarrow \infty.$$

By the Law of Large Numbers,

$$\frac{1}{n}\frac{\partial \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \xrightarrow{p} E\left(\frac{\partial \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\middle|\mathbf{x}\right).$$

For CML+$\widetilde{\boldsymbol{S}}$, the loglikelihood can be written as

$$\ell(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_i \left[R_i \log f_{ci} - R_i \log \pi_i - (1-R_i)\log(1-\pi_i)\right]$$

and at the true values, we have that

$$\begin{aligned}
E\left(\frac{\partial \ell}{\partial \boldsymbol{\beta}}\middle|\mathbf{x}\right) &= E\left(\frac{\partial}{\partial \boldsymbol{\beta}}\log f_c\middle|\mathbf{x}\right) \\
&= \int\left(\frac{\partial}{\partial \boldsymbol{\beta}}\log f - \int \frac{\partial}{\partial \boldsymbol{\beta}}\log\left(\int \pi f d\mathbf{y}\right)\right)\pi f d\mathbf{y} \\
&= \int f'\pi d\mathbf{y} - \int \frac{\int \pi f' d\mathbf{y}}{\int \pi f d\mathbf{y}}\pi f d\mathbf{y} \\
&= \int f'\pi d\mathbf{y} - \int f'\pi d\mathbf{y} \\
&= 0.
\end{aligned}$$

and that

$$\begin{aligned}
E\left(\frac{\partial \ell}{\partial \boldsymbol{\alpha}}\middle|\mathbf{x}\right) &= E\left(\frac{\partial}{\partial \boldsymbol{\alpha}}R\log f_c - \frac{\partial}{\partial \boldsymbol{\alpha}}R\log\pi - \frac{\partial}{\partial \boldsymbol{\alpha}}(1-R)\log(1-\pi)\middle|\mathbf{x}\right) \\
&= \int \pi\frac{\partial}{\partial \boldsymbol{\alpha}}\log(\pi) f d\mathbf{y} - \int \frac{\partial}{\partial \boldsymbol{\alpha}}\log\left(\int \pi f d\mathbf{y}\right)\pi f d\mathbf{y} \\
&\quad - \int \pi\frac{\partial}{\partial \boldsymbol{\alpha}}\log(\pi) f d\mathbf{y} - \int (1-\pi)\frac{\partial}{\partial \boldsymbol{\alpha}}\log(1-\pi) f d\mathbf{y} \\
&= -\int \frac{\int \pi' f d\mathbf{y}}{\int \pi f d\mathbf{y}}\pi f d\mathbf{y} + \int \frac{(1-\pi)}{1-\pi}\pi' f d\mathbf{y} \\
&= -\int \pi' f d\mathbf{y} - \int \pi' f d\mathbf{y} \\
&= 0,
\end{aligned}$$

also evaluated at the true value of $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$.

Thus,

$$\frac{1}{n}\frac{\partial \ell(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}} \xrightarrow{p} 0.$$

In addition, by the Law of Large Numbers,

$$\frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^T} \xrightarrow{p} E_{\mathbf{X}}\left(\frac{\partial^2 \ell(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^T}\right).$$

where

$$E_{\mathbf{X}}\left(\frac{\partial^2 \ell(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^T}\right) = E_{\mathbf{X}}\begin{pmatrix} \frac{\partial^2 \ell(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 \ell(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}^T} \\ \frac{\partial^2 \ell(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 \ell(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} \end{pmatrix} = \boldsymbol{\mathcal{I}}$$

and $E_{\mathbf{X}}$ denotes the expected value conditioned on $\boldsymbol{X} = \mathbf{x}$. We can show that $\boldsymbol{\mathcal{I}}$ is positive definite and hence invertible. Since

$$\boldsymbol{\mathcal{I}}_{00} = E_{\mathbf{X}}\left(\frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right) = -E_{\mathbf{X}}\left(\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\right),$$

we have that

$$\begin{aligned}
& \boldsymbol{a}^T\boldsymbol{\mathcal{I}}_{00}\boldsymbol{a} = 0 \\
\iff \quad & \sum_i a_i \frac{\partial_i \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_i} = 0 \\
\iff \quad & \sum_i a_i \left[\frac{\partial}{\partial \boldsymbol{\beta}_i}\log f(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}_0) - \frac{\partial}{\partial \boldsymbol{\beta}_i}\log\left(\int \pi(\mathbf{z};\boldsymbol{\alpha}_0) f(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}_0) d\mathbf{y}\right)\right] = 0 \\
\iff \quad & \sum_i a_i \int \pi(\mathbf{z};\boldsymbol{\alpha}_0) f'(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}_0) d\mathbf{y} = 0
\end{aligned}$$

Thus, since $\int \pi(\mathbf{z}_j;\boldsymbol{\alpha}_0) f'(\mathbf{y}|\mathbf{x}_j;\boldsymbol{\beta}_0) d\mathbf{y}$ are, by assumption (F), linearly independent, $\boldsymbol{a}^T\boldsymbol{\mathcal{I}}_{00}\boldsymbol{a} = 0 \iff \boldsymbol{a} = 0$ and $\boldsymbol{\mathcal{I}}_{00}$ is invertible. Using similar arguments, with assumption

(G) replacing assumption (F) above, we can show that

$$\boldsymbol{\mathcal{I}}_{11} = E_{\mathbf{Z}}\left(\frac{\partial^2 \ell(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}\right)$$

is also invertible and so, $\boldsymbol{\mathcal{I}}$ is nonsingular.

Hence, letting $\boldsymbol{\phi}^* = \boldsymbol{\phi}_0$,

$$f_N = \frac{1}{n}\frac{\partial \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}$$

$$f'_N = \frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}$$

and

$$H(\boldsymbol{\phi}) = E_X\left(\frac{\partial^2 \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}\right)$$

all conditions of lemma 2.1 are satisfied and we conclude that there is an unique $\hat{\boldsymbol{\phi}}$ such that $f_N(\hat{\boldsymbol{\phi}}) = 0$ with probability going one as $N \longrightarrow \infty$ and $\hat{\boldsymbol{\phi}} \longrightarrow \boldsymbol{\phi}_0$.

Normality

Since $\hat{\boldsymbol{\phi}}$ is consistent, using a Taylor expansion for $\partial \ell(\hat{\boldsymbol{\phi}})/\partial \boldsymbol{\phi}$ around the true parameter $\boldsymbol{\phi}_0$,

$$\frac{\partial \ell(\hat{\boldsymbol{\phi}})}{\partial \boldsymbol{\phi}} = \frac{\partial \ell(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}} + \frac{\partial^2 \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0\right)$$

where $\boldsymbol{\phi} = \kappa \boldsymbol{\phi}_0 (1-\kappa)\hat{\boldsymbol{\phi}}$, for some $\kappa \in [0,1]$. The left-hand side is equals to zero because the estimator has been shown to be consistent. Then, rearranging the above equation, we have that

$$\sqrt{n}\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0\right) = -\left[\frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}\right]^{-1}\left[\frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}}\right]. \tag{2.23}$$

And since $\hat{\boldsymbol{\phi}} \xrightarrow{p} \boldsymbol{\phi}_0$, $\boldsymbol{\phi} \xrightarrow{p} \boldsymbol{\phi}_0$. By using the Law of Large Numbers,

$$-\left[\frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}\right] \xrightarrow{p} \boldsymbol{\mathcal{I}},$$

where

$$\boldsymbol{\mathcal{I}} = E_X\left(\frac{\partial^2 \ell(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}\right),$$

and by applying the Central Limit Theorem to the second term of the right-hand side of equation (2.23), we have that

$$\frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\mathcal{C}}),$$

where

$$\boldsymbol{\mathcal{C}} = \text{Var}\left(\frac{\partial \ell(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}}\right).$$

Finally, by combining the asymptotic results for both terms of the right-hand side of equation (2.23) and using the Slutsky's theorem, we have that

$$\sqrt{n}\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0\right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma} = \boldsymbol{\mathcal{I}}\boldsymbol{\mathcal{C}}\boldsymbol{\mathcal{I}}.$$

Therefore, the asymptotic covariance matrix can be written in the same way as in equation (2.10), replacing $\boldsymbol{\mathcal{C}}_{01}$ and $\boldsymbol{\mathcal{I}}_{11}$ by $\widetilde{\boldsymbol{\mathcal{C}}}_{01}$ and $\widetilde{\boldsymbol{\mathcal{I}}}^*_{11}$, respectively, where

$$\widetilde{\boldsymbol{\mathcal{C}}}_{01} = Cov\left(\boldsymbol{S}_0, \boldsymbol{S}_1 - \widetilde{\boldsymbol{S}}_1\right) \quad \text{and} \quad \widetilde{\boldsymbol{\mathcal{I}}}^*_{11} = \boldsymbol{\mathcal{I}}_{11} - \widetilde{\boldsymbol{\mathcal{I}}}_{11}.$$

In addition, using results 2.1-2.3, the asymptotic covariance matrix can also be rewritten as (2.12), where $\text{Cov}_R = \inf_B \boldsymbol{\mathcal{C}}\{\boldsymbol{S}_0 - B\left(\boldsymbol{S}_1 - \widetilde{\boldsymbol{S}_1}\right)\}$. The reduction of adding a new variable into the selection model depends now on the relationship between the score $\boldsymbol{S}_1 - \widetilde{\boldsymbol{S}}_1$ of the added variable and $\boldsymbol{S}_0$.

2.4 Summary

In this chapter we extended Scott and Wild (2011) approach by showing that the optimal value for λ is still -1 whether the outcome Y is discrete or continuous. This new set of estimating equations are more general than those derived in Scott and Wild (2011) and can thus be applied to a wider range of problems. We also extended previous paper by deriving asymptotic properties for the proposed estimator, under mild conditions.

In chapters 3 and 4 we discuss the proposed method (CML+$\widetilde{\boldsymbol{S}}$) through simulations, for both discrete and continuous outcomes. We compare it against well-known methods discussed in chapter 1 and apply it to a real dataset for the first time.

3

Conditional Maximum Likelihood for discrete response

In this chapter we study the performance of the method proposed in chapter 2, for a discrete response. It is applied to a variety of situations that were not considered in Scott and Wild (2011) and is thus a much more detailed analysis of the proposed method. While Scott and Wild (2011) discussed only a few 3-phase scenarios, we consider here 2 and 3-phase sampling schemes, comparing the performance of the proposed method with well-known methods, under both correct and misspecified models. We also discuss, through simulations, how different models for estimating the selection probabilities affect the final estimate. The method is also applied to a real dataset for the first time. The real data was part of the Women Health Initiative program (Rossouw et al., 2002) and the analysis led to the paper Breslow et al. (2013).

3.1 Introduction

Here we study the performance of the conditional likelihood method for logistic binary-response models, with and without adding $\widetilde{\boldsymbol{S}}$, and compare these with the weighted method and multiple imputation in several situations. We will use mean squared error (MSE) for efficiency comparisons.

In most of what follows, data for a population of N individuals were generated as follows. Covariates $X_1, \ldots, X_p$ were generated from independent standard normal distribution. The response variable Y is a Bernoulli variable, with success probability $\text{logit}(\text{pr}(Y = 1|\mathbf{x}; \boldsymbol{\beta})) = \mathbf{x}\boldsymbol{\beta}$. We first work with a 2-phase sampling scheme, which was not considered in Scott and Wild (2011). Here the phase-1 population consists of all N individuals. The response Y, X_1 and a binary coarsening of X_1, $X_{1d} = I(X_1 < 0)$ or 0 otherwise, were considered known for every subject in phase-1. X_2 and thus also its binary coarsening X_{2d} (equals to 1 if $X_2 < .5$ or 0 otherwise) were treated as observed only for a subsample of the phase-1 population, termed the phase-2 sample. The choice of the thresholds are the same as those used in Scott and Wild (2011). Later we also consider a 3-phase sampling scheme, performing a much more detailed analysis than previously discussed in Scott and Wild (2011). The sampling scheme is similar to the 2-phase problem problem, but here the covariate X_2 is only observed for the phase-3 sample, a subsample taken from the phase-2 sample.

In both 2-phase and 3-phase cases, extra information from partially observed variables can be used to obtain more efficient estimates. Even though measuring certain variables for all individuals may be prohibitively expensive, some cheap variable correlated to that partially observed one may be obtainable for all or nearly all subjects and could be used to improve the estimates of interest.

The flexibility of adding variables into the selection models allows us to fit saturated models and large models when saturated models are not possible, thus bringing in extra information from variables that are not necessarily part of the selection process. We have also the flexibility to incorporate continuous or discrete variables, thus using all information available in a very simple manner. This extra information may have a great impact on $\widetilde{\boldsymbol{S}}$, as shown in the following simulated studies, leading to much more efficient estimates obtained via the CML+$\widetilde{\boldsymbol{S}}$ method.

3.2 Simulations

The following methods are compared with respect to their mean square errors (MSE), bias and coverage.

- Ordinary logistic regression based on the full data (*full*), where it is assumed that there were no missing observations so that we can see the amount of information lost;
- Ordinary logistic regression based on the fully observed units only (*Samp*), ignoring the missingness mechanism;
- Multiple imputation (*Imp*), using the R (R Development Core Team, 2011) package *Hmisc* (Harrell, 2002). Here, bootstrap samples are drawn and outcomes are predicted from a "flexible" model. It predicts each variable from each of the others. Since the partially observed variable was continuous, we used the regression method to predict the missing values, and used a linear model with all fully observed variables;
- The weighted method (*Weighted*), where the weights are the inverse of the probability of being selected for the following phase of the study;

$\boldsymbol{\beta}$	cor(Y, X_1)	cor(Y, X_2)	Prop. of cases	Sample size of $(Y=1, X_{1d}=1, R_1=1)$
(-2.5,1,.5)	0.278	0.214	~ 0.112	100
(-3,1,.5)	0.238	0.218	~ 0.075	100
(-3,1,1)	0.241	0.390	~ 0.091	200
(-3,1,2)	0.225	0.578	~ 0.145	400

Table 3.1: Range of $\boldsymbol{\beta}$ and correlation between Y and $\boldsymbol{X}$.

- Conditional maximum likelihood without (*cml*) and with (*cml+s*) $\widetilde{\boldsymbol{S}}$ added.

We worked with a series of simulations varying the parameters of interest as well as the selection mechanisms. We also worked with a misspecified model of interest in order to investigate the robustness of the proposed methods.

3.2.1 2-phase

We start our simulations with a 2-phase sampling design, using different values of $\boldsymbol{\beta}$, the coefficient of interest. We varied the coefficient of the partially observed variable X_2, increasing the effect of X_2, while keeping a moderate effect of X_1 ($\beta_1 = 1$, fixed). If the coefficient associated to X_2 is zero, the partially observed variable has no impact on the outcome and an ordinary logistic regression should be applied. However, as the effect of X_2 increases, the partially observed variable becomes more important in explaining the outcome of interest. Our goal here is thus to investigate the relative performance of the methods as β_2 increases, while keeping the proportion of cases relatively low (see Table 3.1).

Since the model of interest is $\text{logit}(\text{pr}(Y=1|\mathbf{x}, \boldsymbol{\beta})) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, fixing X_1 and increasing X_2 by 1 unit results in the odds of the outcome being multiplied by $\exp\{.5\} \approx 1.65$ when $\beta_2 = .5$ or by $\exp\{2\} \approx 7.39$ when $\beta_2 = 2$.

For all simulations in this chapter, unless stated otherwise, we worked with a total

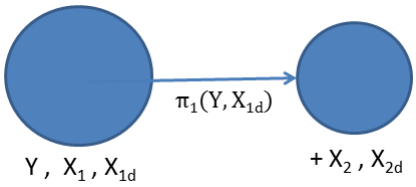


Figure 3.1: Sampling scheme for a 2-phase study, where the response Y, a covariate X_1 and a surrogate variable X_{1d} for X_1 are fully observed at phase-1, for 1000 datasets simulated. An extra covariate X_2 is observed only at the phase-2 of the study

population of $N = 15,000$ and a phase-2 sample taken from each of the four strata defined by the combinations of the categories of (Y, X_{1d}). Four hundred subjects were taken from each stratum, except for the smallest stratum, $(Y = 1, X_{1d} = 1)$. Here we took varying numbers of individuals (see Table 3.1) determined using the typical size of this stratum in the generated population. Fig. 3.1 illustrates this sampling scheme. Both X_1 and X_2 followed a standard normal distribution and the correct model of interest was fitted in all simulations. We fitted two different selection models for the selection probabilities $\pi_i(\mathbf{z}) = \text{pr}(R_i = 1|\mathbf{z}; \boldsymbol{\alpha})$, where $\boldsymbol{Z}$ must be contained in (Y, X_1, X_{1d}) to ensure unbiased estimating equations for $\boldsymbol{\beta}$.

$$\text{Model (i): } \text{logit}(\pi) \sim y * x_{1d}$$

and

$$\text{Model (ii): } \text{logit}(\pi) \sim y * x_{1d} * x_1.$$

Here and in everything that follows, $a * b$ corresponds to a linear regression on a, on b, and on the interaction ab. Notice that model (i) corresponds to the actual selection mechanism while model (ii) is a larger model containing model (i), but also bringing in information on X_1 for all individuals. From equation (2.12), we see that adding X_1 into the selection model will never increase and may even decrease the asymptotic

covariance. Our goal is to see how adding this extra variable might affect the final estimate. The results are presented in Table 3.2.

From our simulations we see that, apart from the simple logistic regression and the multiple imputation method, all methods provide essentially unbiased estimates with a ratio between estimated (Est.SE) and empirical (Emp.SE) standard errors close to 1 and coverage close to the nominal value. As the coefficient of the partially observed variable X_2 increases (and the correlation between Y and X_2 increases), all methods become less efficient when compared to using full data. Multiple imputation shows similar estimates to the weighted method, when the selection model (i) is used. When the selection model (ii) is used, the weighted method becomes considerably more efficient than MI especially for estimating β_1.

As expected from the theoretical results, CML+$\widetilde{\boldsymbol{S}}$ is the most efficient method whether we use model (i) or model (ii) to model the selection probability $\pi(\mathbf{z})$. When model (i) is used, CML+$\widetilde{\boldsymbol{S}}$ gives a much better estimate than other methods, especially with respect to β_1. When model (ii) is used (and so a saturated model for the selection probability is fitted), CML is remarkable close to the fully efficient CML+$\widetilde{\boldsymbol{S}}$ method and both methods provide nearly the same estimates.

The weighted method is less efficient than the CML-based methods. The weighted method can be almost 2 times less efficient than the CML+$\widetilde{\boldsymbol{S}}$ method (MSE 1.85 times higher) for estimating β_1, when the same selection probability is used, or more than 3 times less efficient (MSE 3.23 times higher) for estimating β_1, when the selection probability (i) for the weighted and the selection probability (ii) is used in the CML+$\widetilde{\boldsymbol{S}}$ estimation.

Comparing model (i) with model (ii) for the parameters of interest, we see that by adding the continuous variable X_1 into the selection model, the MSE of β_1, the

Table 3.2: Results for a 2-phase study for different values of $\boldsymbol{\beta}$ and selection models (i) and (ii), for 1000 datasets simulated

		Bias $\times 10^{-3}$			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (-2.5, 1, .5)^T$	Full	-1.26	-0.45	-0.73	0.98	0.98	1.03	1.21	0.93	0.86	0.95	0.96	0.94
	Sample	1586.90	-54.32	-0.13	0.47	0.71	1.01	2519.47	5.62	4.47	0.00	0.94	0.95
	Imp	4.74	-36.19	5.75	0.99	1.01	0.93	2.44	4.63	5.63	0.93	0.86	0.90
	Weighted model(i)	-3.15	2.87	1.20	0.99	1.04	1.02	2.33	4.56	4.95	0.95	0.94	0.94
	cml model(i)	-2.40	1.37	0.25	0.98	1.03	1.01	2.08	3.41	4.47	0.95	0.94	0.94
	cml+s model(i)	-1.70	0.36	0.29	0.97	1.03	1.01	1.86	2.46	4.47	0.96	0.95	0.95
	Weighted model(ii)	-3.01	1.60	1.13	0.96	0.99	1.02	1.75	1.84	4.96	0.96	0.95	0.94
	cml model(ii)	-2.57	0.96	0.13	0.96	0.98	1.01	1.61	1.43	4.47	0.96	0.95	0.94
	cml+s model(ii)	-2.27	0.98	0.24	0.96	0.98	1.01	1.60	1.41	4.47	0.96	0.95	0.95
$\boldsymbol{\beta} = (-3, 1, .5)^T$	Full	-1.59	0.03	-0.04	1.01	0.95	1.00	1.98	1.18	1.12	0.94	0.96	0.95
	Sample	2059.59	-65.76	0.21	0.47	0.65	1.01	4243.15	6.52	4.45	0.00	0.93	0.94
	Imp	5.45	-36.52	6.09	1.02	1.02	1.06	3.14	5.18	5.21	0.94	0.86	0.91
	Weighted model(i)	-3.97	4.34	2.69	1.01	0.99	1.03	3.35	4.89	5.30	0.95	0.94	0.94
	cml model(i)	-3.04	3.10	0.62	1.00	0.97	1.01	2.82	3.24	4.46	0.94	0.95	0.94
	cml+s model(i)	-2.45	2.50	0.64	1.01	1.00	1.01	2.77	2.88	4.46	0.95	0.95	0.94
	Weighted model(ii)	-3.18	3.06	2.85	1.04	1.02	1.03	2.91	2.60	5.31	0.95	0.94	0.94
	cml model(ii)	-2.26	2.16	0.65	1.03	0.99	1.01	2.55	1.85	4.47	0.94	0.96	0.95
	cml+s model(ii)	-1.92	1.93	0.73	1.03	0.99	1.01	2.55	1.81	4.48	0.94	0.96	0.95

Continued on Next Page...

Table 3.2 – Continued

		Bias $\times 10^{-3}$			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (-3, 1, 1)^T$	Full	1.60	-1.50	-0.31	0.98	0.98	0.99	1.95	1.15	1.18	0.95	0.95	0.95
	Sample	1806.43	-22.18	2.06	0.64	0.74	0.96	3266.22	3.81	5.51	0.00	0.99	0.96
	Imp	2.24	-23.85	24.85	0.97	0.94	0.98	4.03	4.66	6.75	0.93	0.89	0.91
	Weighted model(i)	-1.57	3.64	3.95	0.99	1.01	0.99	4.72	5.53	6.86	0.95	0.94	0.95
	cml model(i)	0.63	0.49	2.27	0.95	0.97	0.96	3.85	3.83	5.51	0.96	0.96	0.96
	cml+s model(i)	0.57	1.22	2.30	0.94	0.95	0.96	3.68	3.08	5.51	0.96	0.96	0.96
	Weighted model(ii)	-0.75	2.63	3.85	0.98	1.01	0.99	4.21	3.27	6.85	0.95	0.95	0.95
	cml model(ii)	1.21	0.00	2.28	0.94	0.97	0.96	3.45	2.28	5.50	0.96	0.95	0.96
	cml+s model(ii)	1.23	-0.02	2.40	0.94	0.97	0.96	3.44	2.28	5.48	0.96	0.95	0.96
$\boldsymbol{\beta} = (-3, 1, 2)^T$	Full	-4.32	3.29	3.06	0.99	0.99	0.99	2.43	1.12	1.90	0.94	0.94	0.94
	Sample	1901.40	-440.34	9.99	1.39	1.53	0.94	3620.94	197.28	10.82	0.00	0.00	0.94
	Imp	16.54	5.68	66.94	1.01	0.91	0.89	6.80	4.54	17.11	0.93	0.92	0.86
	Weighted model(i)	-9.06	5.42	11.72	0.98	1.01	0.91	8.78	6.08	15.02	0.95	0.95	0.94
	cml model(i)	-7.89	4.46	10.52	0.97	1.03	0.94	6.37	3.94	10.95	0.95	0.95	0.94
	cml+s model(i)	-7.89	4.26	10.53	0.97	1.02	0.94	6.37	3.76	10.96	0.95	0.96	0.94
	Weighted model(ii)	-8.72	3.26	11.79	0.98	1.01	0.91	8.71	4.83	15.06	0.95	0.95	0.94
	cml model(ii)	-8.10	3.17	10.72	0.97	1.01	0.94	6.28	3.31	10.90	0.95	0.95	0.94
	cml+s model(ii)	-8.13	3.24	10.80	0.97	1.01	0.94	6.29	3.26	10.84	0.95	0.95	0.94

coefficient associated with X_1, is substantially reduced. This is because the residual variance when $\boldsymbol{S}_0$ is regressed on $\boldsymbol{S}_1$ is reduced (see equation (2.12) and comments that follow). The intercept is also affected, but in lower degree. The MSE of $\hat{\beta}_2$ is unchanged, perhaps unsurprisingly, since we are not using any extra information related to this variable.

Assume now that X_{2d} was also fully observed at phase-1. As shown by equation (2.12), adding X_{2d} into the selection model we will never increase, and may even decrease, the mean squared error. Therefore, in order to use the entire information available at phase-1, we fitted another selection model, denoted as *model (iii)*, given by

$$\text{logit}(\pi) \sim y * x_{1d} * x_1 * x_{2d}.$$

Note that model (iii), unlike model (ii), takes information with respect to X_2 into consideration and Table 3.3 shows the impact on the estimates when models (ii) and (iii) are fitted.

From our simulations we see that adding X_{2d} into the selection model improves the estimation of β_2, the coefficient associated to the X_2 variable, reducing its MSE by almost 50%. As discussed before, this happens because adding an extra variable into the selection model will never increase and may even decrease the asymptotic variance of $\hat{\boldsymbol{\beta}}$, even though the real selection mechanism is left unchanged. The reduction depends on the predictive relationship between the score of the added variable and $\boldsymbol{S}_0$.

The CML and the CML+$\widetilde{\boldsymbol{S}}$ methods are also not as similar as before. CML+$\widetilde{\boldsymbol{S}}$ makes use of the extra information in a more efficient way, resulting in much smaller MSE. For example, when $\boldsymbol{\beta} = (-3, 1, 2)^T$, CML produces a MSE for $\hat{\beta}_2$ that is almost 1.30 times larger than the one given by CML+$\widetilde{\boldsymbol{S}}$ and almost 1.81 times larger when $\boldsymbol{\beta} = (-2.5, 1, .5)^T$. When compared to the weighted method, this ratio varies from 1.75

Table 3.3: Results for a 2-phase study when an extra variable is added into the selection model, for 1000 datasets simulated

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (-2.5, 1, .5)^T$	Weighted model(ii)	-0.00	0.00	0.00	0.97	0.95	0.98	2.03	2.08	4.96	0.94	0.94	0.95
	cml model(ii)	-0.00	0.00	0.00	0.99	0.97	0.99	1.82	1.59	4.50	0.94	0.94	0.95
	cml+s model(ii)	-0.00	0.00	0.00	0.98	0.96	0.99	1.82	1.59	4.52	0.94	0.94	0.95
	Weighted model(iii)	-0.00	0.00	0.00	0.97	0.95	0.95	1.71	1.87	2.61	0.93	0.95	0.94
	cml model(iii)	-0.00	0.00	0.00	0.99	0.98	0.97	1.59	1.44	2.42	0.94	0.95	0.95
	cml+s model(iii)	-0.00	0.00	0.00	1.00	0.98	0.98	1.47	1.39	1.34	0.95	0.94	0.94
$\boldsymbol{\beta} = (-3, 1, 1)^T$	Weighted model(ii)	-0.01	0.00	0.01	0.94	0.99	0.98	5.13	3.34	7.37	0.94	0.95	0.95
	cml model(ii)	-0.01	0.00	0.01	0.94	0.99	0.99	4.52	2.52	6.24	0.94	0.95	0.95
	cml+s model(ii)	-0.01	0.00	0.01	0.94	0.98	0.99	4.50	2.52	6.25	0.94	0.95	0.95
	Weighted model(iii)	-0.01	0.00	0.00	0.96	0.99	1.00	3.56	2.49	4.19	0.95	0.96	0.95
	cml model(iii)	-0.00	0.00	0.00	0.97	0.99	1.00	3.15	2.04	3.30	0.95	0.95	0.96
	cml+s model(iii)	-0.00	0.00	0.00	0.97	1.00	0.98	2.84	1.83	2.33	0.94	0.96	0.94
$\boldsymbol{\beta} = (-3, 1, 2)^T$	Weighted model(ii)	-0.01	0.01	0.01	1.01	1.00	1.01	11.20	5.52	15.84	0.95	0.94	0.96
	cml model(ii)	-0.01	0.00	0.01	1.01	1.02	1.02	10.27	4.63	13.74	0.95	0.96	0.95
	cml+s model(ii)	-0.01	0.00	0.01	1.01	1.02	1.02	10.26	4.62	13.74	0.95	0.96	0.95
	Weighted model(iii)	-0.01	0.01	0.01	1.04	1.01	1.02	6.74	3.19	9.81	0.95	0.95	0.95
	cml model(iii)	-0.00	0.00	0.00	1.04	1.04	1.03	5.67	2.79	7.35	0.95	0.96	0.95
	cml+s model(iii)	0.00	-0.00	-0.00	0.95	1.00	0.98	5.64	2.62	5.68	0.95	0.95	0.94

to almost 1.95.

Comparing models (ii) and (iii) in Table 3.3, we see that the ratio between them, which shows the impact of adding X_{2d} into the selection model, is even greater. The MSE for β_2 obtained from the CML+$\widetilde{\boldsymbol{S}}$ method when model (ii) is used is about 3.37 times the one obtained when model (iii) is used and can be almost 4 times more efficient than the weighted method. Even information as minimal as a binary coarsening of X_2 has huge efficiency advantages for estimating β_2.

Model misspecification

We continue to work in the context of missingness by design. In such cases, the "true" selection model for π is controlled by the researcher and so, for studying robustness of the proposed method, we concentrate on a misspecified model of interest. We assume that the true probability of being a case, i.e., $Y=1$, is specified by

$$\text{logit}(\text{pr}(Y=1|\mathbf{x})) = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3$$

where X_1 and X_2 are again independent and identically distributed, following a standard normal distribution, as before. and let X_3 be a quadratic term given by

$$X_3 = X_2^2.$$

We assume that the fitted model is the same as before: $\text{logit}(\text{pr}(Y=1|\mathbf{x})) = \beta_0 + x_1\beta_1 + x_2\beta_2$. We expect to observe bias due to the omitted quadratic term and it is of interest to see, for different values of β_3, how far from the more robust weighted method the efficient estimator CML+$\widetilde{\boldsymbol{S}}$ would be. The results for a 2-phase study are shown in Table 3.4.

Table 3.4: Model misspecification in a 2-phase study, for 1000 datasets simulated

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (-3, 1, .5, .1)^T$	Full	0.09	-0.00	0.07	1.02	1.04	0.95	9.78	1.13	5.98	0.43	0.95	0.43
	Sample	2.04	-0.06	0.05	1.99	1.50	0.95	4174.67	5.68	6.75	0.00	0.95	0.90
	Imp	0.14	0.02	-0.20	0.47	0.23	0.27	27.42	24.29	54.18	0.24	0.37	0.10
	Weighted model(i)	0.09	0.00	0.07	1.00	1.03	0.98	10.72	4.60	10.90	0.67	0.95	0.84
	cml model(i)	0.09	0.00	0.05	1.01	1.06	0.97	11.49	3.06	6.80	0.58	0.96	0.90
	cml+s model(i)	0.09	-0.00	0.05	1.00	1.04	0.97	11.64	2.55	6.80	0.54	0.96	0.90
	Weighted model(ii)	0.09	0.00	0.07	1.00	1.01	0.98	10.34	2.37	10.90	0.59	0.94	0.84
	cml model(ii)	0.09	0.00	0.05	1.01	1.05	0.97	11.20	1.67	6.80	0.51	0.96	0.90
	cml+s model(ii)	0.09	0.00	0.05	1.01	1.05	0.97	11.27	1.64	6.80	0.50	0.96	0.90
$\boldsymbol{\beta} = (-3, 1, .5, 1)^T$	Full	1.33	-0.28	-0.02	0.97	1.02	0.71	1769.00	79.64	1.46	0.00	0.00	0.75
	Sample	2.21	-0.16	-0.14	2.19	1.48	0.82	4902.54	27.10	23.47	0.00	0.36	0.35
	Imp	1.37	-0.23	-0.52	0.48	0.22	0.18	1879.88	67.19	288.18	0.00	0.04	0.00
	Weighted model(i)	1.33	-0.28	-0.02	1.01	1.01	0.99	1766.68	81.42	8.48	0.00	0.00	0.93
	cml model(i)	1.35	-0.28	-0.14	1.03	1.02	0.99	1811.58	83.08	23.71	0.00	0.00	0.47
	cml+s model(i)	1.35	-0.29	-0.14	1.01	1.00	0.99	1824.95	88.19	23.72	0.00	0.00	0.47
	Weighted model(ii)	1.33	-0.28	-0.02	1.01	1.01	0.98	1764.86	79.07	8.60	0.00	0.00	0.94
	cml model(ii)	1.34	-0.28	-0.14	1.04	1.02	0.99	1809.90	80.79	23.76	0.00	0.00	0.47
	cml+s model(ii)	1.35	-0.28	-0.14	1.03	1.02	1.00	1819.78	81.39	23.90	0.00	0.00	0.47

We ran simulations for $\boldsymbol{\beta} = (-3.5, 1, .5, .1)^T$ and $\boldsymbol{\beta} = (-3.5, 1, .5, 1)^T$. For $\beta_3 = .1$, a stepwise regression on the fully observed (phase-2) sample "detects" a quadratic effect in X_2, about 60% of the time, at the 5% level of significance. The intercept and $\hat{\beta}_2$ are slightly biased. CML+$\widetilde{\boldsymbol{S}}$ is still the most efficient method for estimating β_1 and β_2 and shows nearly the same efficiency as the weighted method for estimating β_0.

For the more extreme case in Table 3.4, when $\beta_3 = 1$, all methods are strongly biased but the model misspecification was always detected. The CML+$\widetilde{\boldsymbol{S}}$ heavily depends on the model and is thus greatly affected by model misspecification. The more robust weighted method is the only one that was able to estimate the parameter β_2 correctly, in the sense of giving the same result we would get if we had full data on everything but X_3. Its bias is 7 times lower than the one given by the CML+$\widetilde{\boldsymbol{S}}$ method, resulting in a lower MSE for β_2.

3.2.2 3-phase

We consider now a 3-phase sampling design for different values of $\boldsymbol{\beta}$, performing a more detailed study than discussed in Scott and Wild (2011). We allow the parameters of interest to vary in a wide range of values and consider problems with model misspecifications.

We first start by varying the coefficient of the partially observed variable X_2 while keeping the first two coefficients fixed. Later we vary β_0 in order to simulate a more rare disease. The $\boldsymbol{\beta}-$values used are given in Table 3.5.

For all simulations throughout this chapter, unless stated otherwise, we worked with a total population of $N = 15{,}000$ and the sampling scheme shown in Fig. 3.2. That is, the phase-1 data was considered to be all individuals with (Y, X_{1d}) observed. These subjects were stratified based on the four combinations of (Y, X_{1d}) and a sub-

$\boldsymbol{\beta}$	cor(Y, X_1)	cor(Y, X_2)	Prop. of cases	Sample size of $(Y=1, X_{1d}=1, R_1=1)$
(-2.5,1,.5)	0.388	0.084	~ 0.112	100
(-3,1,.5)	0.391	0.085	~ 0.075	100
(-3,1,1)	0.283	0.158	~ 0.091	200
(-3,1,2)	0.118	0.260	~ 0.145	400

Table 3.5: Range of $\boldsymbol{\beta}$ and correlation between Y and $\boldsymbol{X}$.

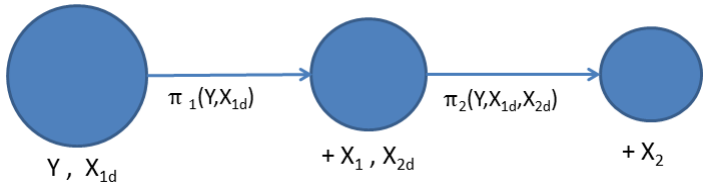


Figure 3.2: Sampling scheme for a 3-phase study, where the response Y and a surrogate variable X_{1d} for X_1 are fully observed at phase-1. At phase-2, X_1 and a surrogate X_{2d} are observed and X_2 observed only for those individuals selected into phase-3.

sample from each stratum was taken. That results in our phase-2 sample, where (Y, X_{1d}, X_1, X_{2d}) have now been observed (400 from each stratum except for the smaller stratum $(Y=1, X_{1d}=1)$, where different sample sizes were taken - see Table 3.5). Subsamples from the phase-2 data were then taken from each stratum defined by (Y, X_{1d}, X_{2d}), resulting in our phase-3 data (50 units from each stratum, except for the smaller strata $(Y=1, X_{1d}=1, X_{2d}=1)$ and $(Y=1, X_{1d}=1, X_{2d}=1)$ where only 25 units were randomly chosen). X_2 is now observed for these individuals. Notice that we have 2 indicator variables R_1 and R_2 denoting which units were selected for phase-2 and phase-3, respectively. Thus, only individuals with $R_1 R_2 = 1$ will have been fully observed.

As before, both X_1 and X_2 variables followed a standard normal distribution and the correct model of interest was fitted in all simulations. We considered the following

selection models. For the first model,

$$\text{logit}(\pi_1) \sim y * x_{1d},$$

which is the actual selection model. For the second model we considered two models

$$\text{Model (i): } \text{logit}(\pi_2) \sim y * x_{1d} * x_{2d}$$

and

$$\text{Model (ii): } \text{logit}(\pi_2) \sim y * x_{1d} * x_{2d} * x_1.$$

Here, model (i) is the actual selection model and model (ii) is included to see the impact of adding X_1 into the selection model for R_2. The results are presented in Table 3.6.

The same pattern observed in the 2-phase study can be seen here. Apart from the ordinary logistic regression, all methods are essentially unbiased, with similar empirical and estimated standard errors and coverage close to the nominal value. CML+$\widetilde{\boldsymbol{S}}$ is still the most efficient method and the weighted method is the least efficient, compared to the other likelihood-based approaches.

Unlike in the 2-phase simulation, we see that CML and CML+$\widetilde{\boldsymbol{S}}$ are no longer of comparable efficiencies. Under the same selection probability model, the latter can be, for example, more than 2 times more efficient than CML when estimating β_1 and about 30% more efficient when estimating β_2.

Comparing the weighted and the CML+$\widetilde{\boldsymbol{S}}$ methods, the improvement is even higher. For the first scenario, where $\boldsymbol{\beta} = (-2.5, 1, .5)^T$, the weighted method produces a MSE that is almost 5 times greater than the one obtained by the CML+$\widetilde{\boldsymbol{S}}$ method, under the same selection model. Comparing different selection models, we see that this ratio

Table 3.6: Results for a 3-phase study for different values of $\boldsymbol{\beta}$ and selection models (i) and (ii), for 1000 datasets simulated

		Bias $\times 10^{-3}$			Est.SE/Emp.SE			MSE			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (-2.5, 1, .5)^T$	Full	-1.60	1.74	1.23	0.99	1.03	0.95	0.93	0.76	0.56	0.96	0.94	0.97
	Sample	1975.02	-357.38	-309.94	0.32	0.62	0.59	3902.40	133.34	100.55	0.00	0.08	0.11
	Imp	20.29	-50.46	-301.70	1.42	1.04	1.85	3.55	5.69	99.12	0.94	0.80	0.31
	Weighted model(i)	-6.81	15.43	4.34	1.04	1.08	0.98	5.76	16.62	9.25	0.94	0.92	0.96
	cml model(i)	-4.15	6.96	4.66	1.01	1.02	0.99	3.79	7.85	8.24	0.94	0.94	0.95
	cml+s model(i)	0.00	3.19	2.23	1.00	0.98	0.99	2.95	3.51	6.79	0.96	0.95	0.95
	Weighted model(ii)	-8.79	15.29	6.33	1.04	1.09	0.98	3.57	7.92	8.76	0.93	0.94	0.96
	cml model(ii)	-4.55	4.84	5.13	1.01	0.97	0.99	3.03	4.00	7.88	0.94	0.96	0.96
	cml+s model(ii)	-0.56	0.84	3.11	0.99	0.98	0.99	2.23	2.54	6.15	0.94	0.95	0.95
$\boldsymbol{\beta} = (-3, 1, .5)^T$	Full	-3.13	1.72	-0.09	0.97	1.02	1.03	1.81	1.35	1.19	0.96	0.94	0.95
	Sample	2461.12	-359.92	-314.94	0.33	0.61	0.61	6058.98	134.92	104.06	0.00	0.07	0.11
	Imp	35.03	-54.46	-300.93	1.37	1.02	1.77	5.11	6.60	98.91	0.92	0.81	0.31
	Weighted model(i)	-14.19	23.46	0.33	1.01	1.03	1.02	7.62	18.33	10.66	0.95	0.94	0.94
	cml model(i)	-7.31	9.44	-0.35	0.99	1.01	0.99	4.61	8.05	8.30	0.95	0.94	0.95
	cml+s model(i)	-2.58	3.15	-2.32	0.99	1.02	0.99	3.96	4.81	6.80	0.96	0.95	0.96
	Weighted model(ii)	-12.45	18.51	1.55	1.02	1.09	1.03	4.94	9.63	9.82	0.94	0.93	0.95
	cml model(ii)	-6.54	6.28	0.76	0.96	0.98	0.98	3.64	4.51	7.79	0.96	0.96	0.95
	cml+s model(ii)	-2.81	2.48	-1.39	0.96	0.98	0.97	3.03	3.39	5.88	0.97	0.95	0.95

Continued on Next Page...

Table 3.6 – Continued

		Bias $\times 10^{-3}$			Est.SE/Emp.SE			MSE			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (-3, 1, 1)^T$	Full	-3.43	1.90	0.85	1.00	1.03	1.00	2.07	1.29	1.20	0.95	0.94	0.95
	Sample	2375.86	-347.39	-633.85	0.38	0.60	0.62	5647.40	126.06	407.27	0.00	0.09	0.00
	Imp	213.45	-96.35	-608.11	1.55	1.06	1.84	49.79	12.50	378.45	0.26	0.59	0.01
	Weighted model(i)	-19.40	20.37	12.12	1.08	1.05	1.07	12.06	17.66	14.31	0.94	0.94	0.94
	cml model(i)	-11.66	11.86	7.89	1.04	1.02	1.04	7.21	9.13	10.55	0.94	0.94	0.94
	cml+s model(i)	-9.27	5.18	8.45	1.03	1.02	1.02	6.30	5.94	8.36	0.95	0.95	0.95
	Weighted model(ii)	-16.15	15.61	11.88	1.08	1.09	1.06	9.24	10.48	13.34	0.92	0.93	0.93
	cml model(ii)	-9.43	7.36	8.42	1.02	1.00	1.02	6.10	5.12	9.94	0.95	0.95	0.94
	cml+s model(ii)	-8.44	5.43	7.87	1.02	1.01	1.00	5.42	4.56	7.47	0.95	0.95	0.95
$\boldsymbol{\beta} = (-3, 1, 2)^T$	Full	-0.22	0.65	-1.32	0.96	1.00	0.97	2.19	1.11	1.77	0.96	0.95	0.96
	Sample	2215.20	-323.74	-1292.81	0.40	0.60	0.58	4910.55	110.59	1677.96	0.00	0.19	0.00
	Imp	703.08	-244.40	-1262.28	1.63	1.13	1.75	499.47	63.01	1604.11	0.00	0.06	0.00
	Weighted model(i)	-12.09	20.40	14.78	1.03	1.04	1.03	22.27	18.27	26.09	0.95	0.94	0.94
	cml model(i)	-10.90	12.76	16.56	1.04	1.01	1.04	13.37	11.87	19.12	0.95	0.95	0.93
	cml+s model(i)	-6.47	8.57	11.80	1.05	1.03	1.04	12.73	9.77	17.11	0.95	0.95	0.93
	Weighted model(ii)	-12.82	20.22	16.14	1.05	1.08	1.04	21.26	13.98	25.95	0.93	0.93	0.94
	cml model(ii)	-9.90	11.04	16.45	1.04	1.02	1.03	12.59	8.57	18.67	0.95	0.95	0.93
	cml+s model(ii)	-5.68	8.44	11.66	1.05	1.02	1.03	12.12	8.16	16.36	0.95	0.95	0.93

can be as large as 6.54, when estimating X_1.

We also see a great reduction in the MSE of all methods when the continuous variable X_1 is added into the second selection model, i.e., when we change from model (i) to model (ii). For the weighted and CML methods, the MSE is reduced by almost a half, while for CML+$\widetilde{\boldsymbol{S}}$ the MSE is reduced by almost 40%. Again, since both models (i) and (ii) carry the same amount of information with respect to X_2, the MSE of β_2 is essentially unchanged.

Next, as in the 2-phase simulation, we assume that X_{2d} was also known at phase-1, but not used for selecting the phase-2 sample. We worked with two set of models for R_1 and R_2:

$$\begin{aligned} \text{Model (ii)} \ : & \text{logit}(\pi_1) \sim y * x_{1d} \\ & \text{logit}(\pi_2) \sim y * x_{1d} * x_{2d} * x_1 \end{aligned}$$

and

$$\begin{aligned} \text{Model (iii)} \ : & \text{logit}(\pi_1) \sim y * x_{1d} * x_{2d} \\ & \text{logit}(\pi_2) \sim y * x_{1d} * x_{2d} * x_1. \end{aligned}$$

Simulation results are shown in Table 3.7.

From our simulations we see that adding X_{2d} into the first selection may lead to estimates that are almost 3 times more efficient than not using this extra information. Comparing CML+$\widetilde{\boldsymbol{S}}$ (under model (iii)) against the weighted method (under model (ii)), we see that the first CML+$\widetilde{\boldsymbol{S}}$ can be almost 6 times more efficient than the latter one, while estimating β_2. Notice also that the CML+$\widetilde{\boldsymbol{S}}$ method makes a better use

Table 3.7: Results for a 3-phase study when an extra variable is added into the selection models, for 1000 datasets simulated

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (-2.5, 1, .5)^T$	Weighted model(ii)	-0.01	0.01	0.00	0.98	0.97	0.93	3.15	6.85	9.89	0.95	0.94	0.93
	cml model(ii)	-0.00	0.00	0.00	1.00	1.04	0.94	2.47	2.98	8.06	0.95	0.96	0.93
	cml+s model(ii)	-0.00	0.00	0.00	1.06	1.10	1.02	1.74	1.95	5.02	0.96	0.96	0.95
	Weighted model(iii)	-0.01	0.01	0.00	0.98	0.97	0.92	2.98	6.77	7.75	0.95	0.95	0.94
	cml model(iii)	-0.00	0.00	0.00	1.00	1.05	0.94	2.38	2.73	6.31	0.95	0.96	0.92
	cml+s model(iii)	-0.00	0.00	0.00	1.05	1.03	1.02	1.60	2.00	1.87	0.95	0.95	0.95
$\boldsymbol{\beta} = (-3, 1, 1)^T$	Weighted model(ii)	-0.02	0.02	0.01	0.98	0.92	0.98	7.83	10.22	12.30	0.94	0.92	0.94
	cml model(ii)	-0.01	0.01	0.01	0.99	0.99	0.98	4.92	4.40	9.41	0.94	0.95	0.94
	cml+s model(ii)	-0.01	0.01	0.01	1.01	1.01	1.00	4.04	3.62	6.80	0.94	0.95	0.95
	Weighted model(iii)	-0.02	0.02	0.01	0.97	0.92	0.97	7.17	9.82	9.80	0.94	0.92	0.94
	cml model(iii)	-0.01	0.01	0.00	1.00	1.00	0.99	4.53	4.09	7.40	0.94	0.94	0.95
	cml+s model(iii)	-0.01	0.01	0.00	1.02	1.02	1.02	3.48	3.18	3.27	0.95	0.95	0.95
$\boldsymbol{\beta} = (-3, 1, 2)^T$	Weighted model(ii)	-0.02	0.02	0.02	0.95	0.92	0.94	20.20	12.70	25.30	0.93	0.91	0.93
	cml model(ii)	-0.01	0.01	0.01	0.97	1.00	0.99	10.93	6.86	16.17	0.95	0.96	0.95
	cml+s model(ii)	-0.01	0.01	0.01	0.97	1.01	0.98	10.00	6.65	13.50	0.94	0.95	0.95
	Weighted model(iii)	-0.02	0.02	0.02	0.94	0.92	0.93	18.81	11.61	22.64	0.93	0.91	0.93
	cml model(iii)	-0.01	0.01	0.01	0.98	1.01	0.99	9.67	6.19	13.75	0.95	0.96	0.95
	cml+s model(iii)	-0.01	0.01	0.01	0.98	1.02	0.97	8.39	5.80	9.60	0.95	0.95	0.95

of the extra information, when compared to CML. While the CML+$\widetilde{\boldsymbol{S}}$ method was slightly more efficient before when X_{2d} was assumed unobserved at phase-1, now this difference is substantially higher.

Model misspecification

As in the 2-phase study, we ran a few simulations in order to investigate robustness of the proposed method against an omitted quadratic term in the model of interest. As in section 3.2.1, we assumed that true model of interest was given by

$$\text{logit}(\text{pr}(Y = 1|\mathbf{x})) = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3$$

where

$$X_3 = X_2^2.$$

As before, we fitted the model: $\text{logit}(\text{pr}(Y = 1|\mathbf{x})) = \beta_0 + x_1\beta_1 + x_2\beta_2$, with the same selection probability models (i) and (ii). Our results are shown in Table 3.8.

For $\beta_3 = .1$, a stepwise regression "detects" a quadratic effect in X_2 about 38% of the time, at the 5% level of significance. In such case, all methods show a small bias for β_2, but the CML+$\widetilde{\boldsymbol{S}}$ method is still the most efficient approach. As the model misspecification increases, the model misspecification is readily detectable. CML+$\widetilde{\boldsymbol{S}}$ shows larger bias for β_2, resulting in estimates with higher MSE than those given by the CML and the weighted methods.

Table 3.8: Model misspecification in a 3-phase study, for 1000 datasets simulated

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (-3, 1, .5, .1)^T$	Full	0.09	-0.00	0.07	0.99	0.97	0.92	9.23	0.97	6.10	0.34	0.94	0.30
	Sample	2.46	-0.36	-0.29	3.02	1.74	1.64	6033.94	133.19	91.05	0.00	0.06	0.13
	Imp	0.14	0.03	-0.21	0.45	0.22	0.26	24.79	20.17	56.54	0.20	0.33	0.07
	Weighted model(i)	0.08	0.02	0.08	0.98	0.98	1.01	13.07	16.76	16.40	0.80	0.93	0.89
	cml model(i)	0.09	0.01	0.07	1.01	1.01	1.03	12.24	7.51	11.81	0.69	0.95	0.89
	cml+s model(i)	0.09	0.00	0.07	0.99	0.97	1.05	11.73	4.39	10.42	0.66	0.95	0.86
	Weighted model(ii)	0.08	0.01	0.08	0.99	0.93	1.01	10.77	8.86	15.81	0.73	0.93	0.87
	cml model(ii)	0.09	0.00	0.07	1.05	1.03	1.03	11.41	3.69	11.53	0.66	0.96	0.89
	cml+s model(ii)	0.09	0.00	0.07	1.05	1.03	1.06	10.97	2.91	9.80	0.58	0.95	0.88
$\boldsymbol{\beta} = (-3, 1, .5, 1)^T$	Full	1.33	-0.28	-0.02	0.91	0.97	0.66	1767.78	80.09	1.23	0.00	0.00	0.70
	Sample	2.60	-0.45	-0.52	4.19	1.67	1.61	6772.51	203.16	276.35	0.00	0.00	0.00
	Imp	1.39	-0.26	-0.50	0.54	0.21	0.17	1921.92	79.64	260.09	0.00	0.01	0.00
	Weighted model(i)	1.32	-0.27	-0.02	1.00	1.02	0.98	1754.30	83.58	11.84	0.00	0.18	0.94
	cml model(i)	1.41	-0.24	-0.06	1.02	1.03	1.12	1980.21	66.76	10.02	0.00	0.19	0.90
	cml+s model(i)	1.36	-0.26	0.11	1.12	1.10	1.10	1847.36	72.14	16.74	0.00	0.01	0.69
	Weighted model(ii)	1.33	-0.28	-0.02	1.02	0.97	0.99	1758.38	80.50	11.22	0.00	0.01	0.95
	cml model(ii)	1.41	-0.25	-0.06	1.04	1.01	1.13	1984.56	64.40	9.85	0.00	0.03	0.89
	cml+s model(ii)	1.36	-0.26	0.11	1.19	1.17	1.03	1844.51	68.51	18.04	0.00	0.01	0.65

3.3 Non-response

Even though we have only discussed cases where data are missing by design, the same idea can be applied to non-response problems. The main difference is that the selection probability, i.e., the probability of observing an individual in the next phase of the study, is no longer controlled by the researcher and must be estimated.

Ghosh and Dewanji (2011) discussed non-response in a 2-phase sampling scheme with a fully observed phase-1 population. Here, the phase-1 subjects have $\boldsymbol{X}_1$ observed and are "asked" to provide information regarding Y and $\boldsymbol{X}_2$. Next, either (1) all phase-1 population have $\boldsymbol{X}_2$ measured or (2) a simple random sample is taken from the phase-1 subjects and have $(\boldsymbol{X}_2, Y, R)$ measured. In case (1), all phase-1 subjects have $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ measured and only the respondents $(R_1 = 1)$ have $(\boldsymbol{X}_1, \boldsymbol{X}_2, Y)$ observed. In case (2), all phase-1 subjects have $\boldsymbol{X}_1$ measured and all respondents $(R_1 = 1)$ and only a sample of non-respondents $(R_1 = 0)$ have $(\boldsymbol{X}_1, \boldsymbol{X}_2, Y)$ observed. Fig. 3.3 represents both designs. Case (1) corresponds to $\boldsymbol{Z}_1 = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, $\boldsymbol{Z}_2 = Y$ and empty $\boldsymbol{Z}_3$ and case (2) corresponds to $\boldsymbol{Z}_1 = \boldsymbol{X}_1$, $\boldsymbol{Z}_2 = \boldsymbol{Z}_3 = (\boldsymbol{X}_2, Y)$.

Here we consider the following sampling scheme, where information on (Y, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$) is of interest. However, only $(Y, \boldsymbol{X}_1)$ have been fully observed for all subjects, known as the phase-1 population, and additional information with respect to $\boldsymbol{X}_2$ is observed only for a sample drawn from the phase-1 subjects. Phase-2 individuals are "asked" for information on other variables of interest. Some of them respond, providing information on $\boldsymbol{X}_3$, while others do not. A follow-up sample is then taken from the non-responding units and the remaining variables, measured.

Suppose now that the probability of responding depends on variables not measured in the original study, say $\boldsymbol{X}_4$. If we do not adjust for this variable, the estimates will be

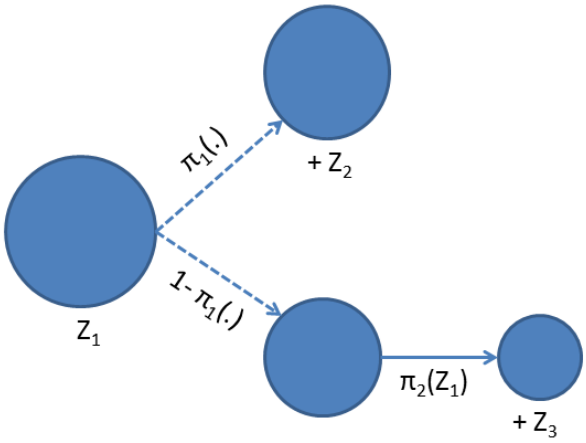


Figure 3.3: Nonresponse sampling scheme used in Ghosh and Dewanji (2011). Here, $\pi_1(.)$ corresponds to the (unknown) probability of responding (dashed line) and $\pi_2(\mathbf{z}_1)$ is the probability of providing extra information regarding $\boldsymbol{Z}_3$, given that this subject did not respond $R_2 = 0$.

biased. So, we must collect information on $\boldsymbol{X}_4$ for both groups: respondents and non-respondents. In other words, we need to perform two follow-up studies and measure the remaining variable of interest. This sampling scheme generalizes that used in Ghosh and Dewanji (2011), explained before, and Jiang et al. (2011), who assumed that all respondents had X_4 observed and so only a sample of non-respondents needed to be selected for follow-up.

In order to deal with this more general sampling scheme, let the indicator variables R_{1i} be equal to 1 if the ith unit was selected into phase-2, $R_{2i} = 1$ if it responded and $R_{3i} = 1$ if it was selected for follow-up. Let also $\pi_1 = \pi_1(\boldsymbol{\alpha}_1)$ be the selection model from phase-1 into phase-2 ($R_1 = 1$). This is known by design and thus controlled by the researcher. Let $\pi_2 = \pi_2(\boldsymbol{\alpha}_2)$ be the (unknown) probability of responding and $\pi_3 = \pi_3(\boldsymbol{\alpha}_3)$ the probability of being selected into the follow-up part of the study. This sampling scheme is illustrated in Fig. 3.4.

Even though this sampling scheme seems to be equivalent to the multi-phase problems discussed so far, they are not quite the same. For instance, in the regular multi-

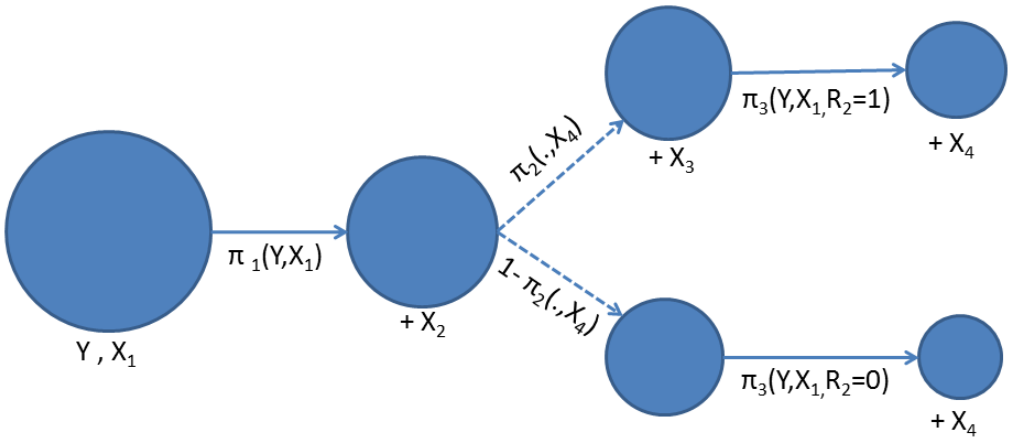


Figure 3.4: Nonresponse sampling scheme used in this chapter. Here, $\pi_2(.,x_4)$ corresponds to the (unknown) probability of responding (dashed line) and $\pi_3(y,x_1,R_2=i)$, for $i=1$ or 2, is the probability of being selected for follow up, given that it responded $R_2=1$ or did not respond $R_2=0$. Only individuals selected for follow-up have been fully observed.

phase problem, only individuals with $R_2=1$ could be selected into phase-3, which is not the case here. Individuals are selected for follow-up here whether $R_2=1$ or 0. That is, R_2 can be seen as a binary outcome of a case-control study, where $R_2=1$ denotes "cases" and $R_2=0$, "controls". In other words, this sampling scheme can be seen as an embedded case-control study: The outcome of the "larger" study is Y, while R_2 can be seen as the response variable for the embedded case-control study. The selection probability π_2 can then be estimated via the conditional likelihood method in the same way we have done so far. From Bayes theorem, the conditional probability of responding given that it was selected for the next phase of the study (here, selected for follow-up) is

$$\text{pr}(R_{2i}=1|R_{1i}=1,R_{3i}=1,\mathbf{x}_i,y_i)=$$
$$\frac{\text{pr}(R_{3i}=1|R_{1i}=1,R_{2i}=1,\mathbf{x}_i,y_i)\,\text{pr}(R_{2i}=1|R_{1i}=1,\mathbf{x}_i,y_i)}{\sum_{j=0,1}\text{pr}(R_{3i}=1|R_{1i}=1,R_{2i}=j,\mathbf{x}_i,y_i)\,\text{pr}(R_{2i}=j|R_{1i}=1,\mathbf{x}_i,y_i)}.$$

The associated estimating equation is

$$\boldsymbol{S}_2^* = \sum_i R_{1i} R_{3i} \left(\frac{(R_{2i} - \pi_{2i}^*)}{\pi_{2i}^*(1-\pi_{2i}^*)} \frac{\partial \pi_{2i}^*}{\partial \boldsymbol{\alpha}_2} \right),$$

where

$$\pi_{2i}^* = \text{pr}(R_{2i} = 1 | R_{1i} = 1, R_{3i} = 1, \mathbf{x}_i, y_i)$$

is the probability of responding among individuals selected for follow-up.

Letting $\text{logit}\{\pi_{2i}\} = \mathbf{z}_2^T \boldsymbol{\alpha}_2$, where $\boldsymbol{Z}_2 = (\boldsymbol{X}_1, Y)$, we have that

$$\pi_{2i} = \frac{e^{(\mathbf{z}_{2i}^T \alpha_2)}}{\pi_{3i} + e^{(\mathbf{z}_{2i}^T \alpha_2)}} = \frac{e^{(\mathbf{z}_{2i}^T \alpha_2 + o_{2i})}}{1 + e^{(\mathbf{z}_{2i}^T \alpha_2 + o_{2i})}}$$

where

$$o_{2i} = \log \pi_{3i}^{(1)} - \log \pi_{3i}^{(0)}.$$

Here,

$$\pi_{3i}^{(0)} = \text{pr}(R_{3i} = 1 | R_{1i} = 1, R_{2i} = 0, \mathbf{x}_i, y_i)$$

and

$$\pi_{3i}^{(1)} = \text{pr}(R_{3i} = 1 | R_{1i} = 1, R_{2i} = 1, \mathbf{x}_i, y_i).$$

are the probabilities of the i*th* subject being chosen for follow-up giving that it did and did not respond, respectively. We wrote an R function for the general sampling scheme showed in Fig. 3.4 that estimates the parameters of interest by setting the enlarged estimating equation system to zero.

3.3.1 Simulation

Following Jiang et al. (2011), the disease incidence was generated from the model

$$\text{logit}(\text{pr}(Y = 1|\mathbf{x}; \boldsymbol{\beta})) = \beta_0 + \beta_1 x_3 + \beta_2 x_4,$$

with $\boldsymbol{\beta} = (-7.9, 1, .5)^T$, X_3 following a standard normal distribution and X_4 following a Bernoulli distribution with success probability 0.2 if $X_3 < 0$ and 0.5 if $X_3 \geq 0$. This results in 0.1% population being cases and we used a population of $N = 1,000,000$ individuals. That is our phase-1 population.

The phase-2 sample was obtained as a sample taken from our phase-1 data, as follows. We first defined an indicator variable X_1 equal to 1 if $X_3 \geq .5$ and 0 otherwise and selected samples of size $n = 300$ from each stratum defined by (X_1, Y). Notice that the selection here is due to design so the "true" selection probability π_1 is controlled by the researcher. The response indicator R_2 was generated from the model

$$\text{logit}(\pi_2(\mathbf{z}_2)) = \alpha_{20} + \alpha_{21} y + \alpha_{22} x_4 + \alpha_{23} y x_4,$$

where $\alpha_2 = (.75, 0, 0, .75)^T$ and $\boldsymbol{Z}_2 = (X_4, Y)$, so that the response rate is independent of X_1 and is approximately 70% for controls and 85% for cases when $X_4 = 1$ and 70% when $X_4 = 0$.

Jiang et al. (2011) assumed that only a fraction of non-respondents ($R_2 = 0$) had provided information regarding X_4 while all respondents had had X_4 observed. Here we allow X_4 to be missing in both groups. Since X_4 is related to non-response, ignoring this variable will lead to a misspecified model for π_2 and thus to biased results. We need, therefore, to select samples for follow-up from both groups (respondents and non-

respondents) and measure X_4 for these sampled individuals. This sampling scheme is illustrated in Fig. 3.4, for empty X_2.

We varied the percentage of subjects in the non-respondents group selected for follow-up (25, 50, 75 and 90%) and kept the number of subjects in the respondents group selected for follow-up constant and equal to 70% of the total of respondents. That is, we assume that only 70% of the respondents ($R_2 = 1$) provided information regarding X_4. The results are shown in Table 3.9.

In our simulations we assumed that all models were correctly specified, resulting in unbiased estimates. The first row of Table 3.9 shows the results when X_4 is observed for all respondents. Here CML+$\widetilde{\boldsymbol{S}}$ is the most efficient method especially when we are interested in estimating β_1. The CML+$\widetilde{\boldsymbol{S}}$ method is about 30% more efficient than CML and almost 100% more efficient than the weighted method. However, if we are interested in estimating β_0 or β_2, the extra term $\widetilde{\boldsymbol{S}}$ does not carry appreciable information and the two methods, CML+$\widetilde{\boldsymbol{S}}$ and CML, show similar results.

For the next three rows we assume that a fixed proportion of 70% of the respondents provide information regarding X_4 and a varying proportion ρ (.90, .75, .50 and .25) of the non-respondents provide information about X_4. We see that as the non-response increases, the MSE of all methods increase. When X_2 is observed for only 75% of the non-respondents, the efficiency of all three methods decrease by about 25% while estimating β_2 and about 50% when $\rho = .25$.

CML+$\widetilde{\boldsymbol{S}}$ is generally the most efficient method among the three. It shows, however, a large loss of efficiency when estimating β_1, with a MSE varying from 1.58 when $\rho = 1$ to 3.18 $\rho = .25$. In fact, at this extreme case where hardly any non-respondents provide extra information regarding X_4, both CML and CML+$\widetilde{\boldsymbol{S}}$ show nearly the same efficiency.

Table 3.9: Effects of varying the probability of providing full information in a non-response study, for 1000 datasets simulated.

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
% of respondents		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
100 %	Weighted	-0.01	0.00	0.01	1.07	0.91	0.98	3.00	2.95	10.15	0.97	0.92	0.94
	cml	-0.00	0.00	0.01	1.06	0.95	0.98	2.93	2.11	9.89	0.96	0.93	0.95
	cml+s	-0.00	-0.00	0.01	1.08	0.96	0.98	2.78	1.58	9.93	0.96	0.94	0.95
90 %	Weighted	0.00	0.00	0.00	1.00	1.03	1.02	5.05	3.95	13.58	0.95	0.95	0.95
	cml	0.00	0.00	0.00	0.99	1.02	1.02	4.74	3.13	13.24	0.95	0.95	0.94
	cml+s	0.00	0.00	0.00	0.99	1.02	1.02	4.69	2.58	13.25	0.96	0.95	0.95
75 %	Weighted	0.00	0.00	0.00	0.97	1.03	0.99	5.10	4.09	13.35	0.96	0.95	0.95
	cml	0.00	0.00	0.00	0.96	0.99	0.99	4.81	3.06	13.09	0.96	0.95	0.95
	cml+s	0.00	0.00	0.00	0.96	0.96	0.99	4.76	2.46	13.03	0.96	0.96	0.96
50 %	Weighted	0.00	0.00	0.00	0.94	0.97	0.98	5.62	3.90	14.66	0.96	0.97	0.96
	cml	0.00	0.00	0.00	0.93	0.96	0.98	5.37	3.11	14.33	0.97	0.96	0.96
	cml+s	0.00	0.00	0.00	0.93	0.89	0.98	5.33	2.44	14.23	0.96	0.97	0.96
25 %	Weighted	-0.01	0.01	0.00	0.82	0.91	1.00	6.41	4.08	20.35	0.98	0.96	0.94
	cml	-0.01	0.00	0.00	0.82	0.87	1.01	6.18	3.19	20.02	0.99	0.97	0.94
	cml+s	0.00	0.01	-0.01	0.80	0.71	0.97	6.01	3.18	20.74	0.99	0.99	0.95

3.4 Application to Women's Health Initiative data

The data we use here come from the Woman's Health Initiative (WHI) study and was collected from September 1993 and December 1998. Our subset consists of the 27,347 postmenopausal women aged between 50 and 79 years old. With the purpose of preventing coronary heart disease (CHD), all individuals were randomized into one of two trials of hormone therapy (HT): 16,608 women with an intact uterus were randomized to a combination of estrogen plus progestin (E+P) versus placebo, while the remaining 10,739 women with prior hysterectomy were randomized to estrogen alone (E) versus placebo. Both trials had to be stopped early because of an increase in CHD, stroke and venous thromboembolism (VTE) among those receiving treatment or lack of CHD benefit and an increase in stroke. Three case-control studies were then conducted to investigate the mechanisms through which HT might increase the risk of each of the cardiovascular events: CHD, stroke and VTE. Here we are only concerned with the CHD case-control study. See Rossouw et al. (2002) for more details regarding the dataset.

The case-control study of CHD involved 359 cases (Rossouw et al., 2008). The 817 controls, matched by date of randomization, trial (E vs E+P), age, and prevalence of the particular cardiovascular disease (CVD) at baseline, were a pool drawn from all three case control studies. Blood samples stored at baseline were assayed for inflammatory, lipid, thrombotic and genetic markers for cases and controls. The data were analysed by ordinary logistic regression, with case-control status as the outcome and each of the biomarkers in turn as the primary risk factor. All analyses were adjusted for baseline age, trial, body mass index, waist-hip ratio, smoking, alcohol, physical activity, diabetes, history of high cholesterol, prevalent CVD, left ventricular hypertro-

phy, systolic blood pressure (SBP), use of anti-hypertensive medications, aspirin and statins.

Among the more interesting findings in Rossouw et al. (2008) was a positive interaction between HT and low density lipoprotein cholesterol (LDL), such that the increased CHD risk for high LDL levels was even greater with treatment, and a negative interaction between HT and high density lipoprotein cholesterol (HDL), such that the decreased CHD risk for high HDL levels was even lower with treatment. The genetic polymorphism GpIIIa leu33pro (C/C+C/T vs T/T) was also associated with increased risk. Preliminary analyses showed that a model with main and interactive effects for the log(LDL/HDL) ratio fit nearly as well as the model with main and interactive effects for both log(LDL) and log(HDL) (deviance difference 0.55, 2 degrees of freedom). Consequently, as a single comprehensive model for our methodological studies, we selected as variables of primary interest the log(LDL/HDL) ratio (centred at log 3), HT, their interaction and the GpIIIa leu33pro polymorphism. For adjustment we used seven of the fifteen variables mentioned above. The other eight variables were retained for use as design or auxiliary variables. Unfortunately, the necessity that values for all of them be known reduced the main cohort size from 27,347 to 23,301, including 276 cases. Twenty-eight more cases and a proportionate number of controls were dropped due to missing values for the lipid levels or genotype, leaving 248 cases and 617 controls in the case-control analysis sample. Results of the standard analysis are shown in Table 3.10.

The data can be viewed as a two-phase sample, where the main cohort is the phase-1 sample, assumed to have been drawn by simple random sampling from a "superpopuluation", i.e., from a logistic regression model for CHD risk. The phase-2 sample is the case-control sample, drawn by outcome dependent stratified sampling from the phase-1 sample. A particular feature of the WHI study, however, invites consideration of a more

complex design. In order to assess adherence to treatment, the investigators selected a 8.6% cohort random sample, stratified on trial and ethnicity, for assay of baseline and subsequent blood samples for selected biomarkers (Anderson et al., 2003). Baseline LDL and HDL levels were thus available for an additional 2,158 controls not sampled for the case-control study. The resulting data may be regarded as having arisen from a three phase design, with the main cohort as phase-1, all subjects with known LDL and HDL levels as phase-2 and the case-control sample, for which genotype is also known, as phase-3.

2 and 3-phase approaches

Table 3.10 shows the results of the standard (STD) case-control analysis in comparison with those for each of the basic 2-phase methods. These were obtained using the *tps* procedure in the R-package *osDesign* that implements the Breslow-Holubkov approach to maximum likelihood (which is, in this setting, equivalent to the CML+$\widetilde{\boldsymbol{S}}$ method, as shown in chapter 7) estimation (Breslow and Holubkov, 1997; Haneuse et al., 1997). The more general program *bin2stg* in Chris Wild's *missreg* package is available as an alternative.

All three 2-phase methods corrected for the serious bias in the control sample with respect to age and prevalent CVD. Consequently, the regression coefficients for these variables were considerably higher, and the standard errors lower, than for the standard method, reflecting the information contained in the phase-1 stratum frequencies N_{ij}. The weighted method produced a noticeably smaller estimate of the interaction between HT and the log(LDL/HDL) ratio. It also produced standard errors that were generally larger than those for the other methods, including even the standard method, reflecting the well-known and often serious loss of efficiency of the weighted method in comparison

Model	STD		Weighted		CML		CML+$\widetilde{\boldsymbol{S}}$	
term	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
log(LDL/HDL)	0.83	0.35	0.98	0.38	0.82	0.35	0.83	0.35
HT	0.16	0.17	0.07	0.18	0.16	0.17	0.15	0.17
HT*log(L/H)	0.97	0.49	0.56	0.50	0.99	0.49	0.99	0.49
GpIIIa	0.43	0.18	0.44	0.19	0.44	0.18	0.43	0.18
Age/10	0.07	0.14	0.56	0.13	0.58	0.12	0.59	0.12
Current smoker	1.45	0.24	1.49	0.25	1.42	0.24	1.42	0.24
Diabetes	1.30	0.28	1.52	0.28	1.35	0.28	1.35	0.28
Prevalent CVD	0.64	0.21	1.19	0.19	1.18	0.17	1.18	0.17
SBP/100	2.30	0.49	2.41	0.56	2.29	0.49	2.35	0.49
Statins	0.67	0.25	0.60	0.27	0.64	0.25	0.65	0.25
Trial E+P	0.24	0.17	0.10	0.16	0.11	0.15	0.13	0.15

Table 3.10: Results of standard case-control and two-phase analysis: phase-1 variables limited to indicators of strata used for sampling, for 1000 datasets simulated.

with ordinary logistic regression for case-control samples. Otherwise, the results for the four analyses were remarkably similar. The 2-phase methods only reduced standard errors for variables used in the phase-1 stratification.

We have also done a 3-phase analysis and the results are shown in Table 3.11. Recall that the phase-2 sample in the 3-phase design consisted of subjects with known LDL and HDL levels. Most of these were from the 8.6% cohort random sample that was stratified on trial and ethnicity. The inclusion model for phase-2 had as predictors: the three-way combination of outcome, trial and ethnicity; the four way combination involving the sampling strata and outcome and used to obtain the results in Table 3.10; and age, treatment, smoking, prevalent CVD, diabetes, SBP, statins plus the product of each with the case-control indicator. The inclusion model for phase-3 added to these terms the three way combination of outcome, treatment and log(LDL/HDL). The idea was to try to bring into the analysis the information available from both phases on marginal associations between outcome and each of the covariates. Incorporation of the additional information on LDL and HDL from the 8.6% sample had a noticeable effect on precision for the three variables of prime interest, namely, log(LDL/HDL), HT and

Model term	Weighted		CML		CML+$\widetilde{\boldsymbol{S}}$	
	Coef.	SE	Coef.	SE	Coef.	SE
log(LDL/HDL)	1.18	0.36	0.99	0.31	1.03	0.31
HT	0.09	0.16	0.21	0.15	0.16	0.15
HT*log(L/H)	0.63	0.47	1.08	0.42	1,00	0.43
GpIIIa	0.45	0.20	0.46	0.18	0.45	0.18
Age/10	0.68	0.13	0.63	0.12	0.65	0.12
Current smoker	1.40	0.21	1.26	0.19	1.26	0.19
Diabetes	1.27	0.22	1.12	0.20	1.02	0.19
Prevalent CVD	1.11	0.19	1.13	0.18	1.11	0.18
SBP/100	1.89	0.51	1.71	0.43	1.85	0.43
Statins	0.61	0.23	0.67	0.21	0.68	0.21
Trial E+P	0.09	0.17	0.08	0.15	0.13	0.15

Table 3.11: Results of three-phase analysis: stratum indicators plus additional covariates at both phases.

their interaction. In comparison with results from Table 3.10, SEs for these variables were reduced for all three methods (Weighted, CML, CML+$\widetilde{\boldsymbol{S}}$). There was stronger evidence from CML and CML+$\widetilde{\boldsymbol{S}}$ for an interaction between HT and log ratio; the results from the weighted method were still inconclusive. The SEs for the adjustment variables decreased for the weighted method whereas several increased slightly for the CML and CML+$\widetilde{\boldsymbol{S}}$ methods in comparison with the 2-phase results. Consequently, the loss of precision from using the weighted method relative to CML/CML+$\widetilde{\boldsymbol{S}}$ was reduced.

3.5 Summary

Here we analyzed the proposed CML+$\widetilde{\boldsymbol{S}}$ method through simulations, for both 2 and 3-phase problems. The 2-phase design was not considered in Scott and Wild (2011) and for the 3-phase design a more detailed analysis was performed. We also discussed robustness against model misspecification for both 2 and 3-phase problems.

In short, both 2 and 3-phase problems produced similar patterns. If all models are

correctly specified, the CML+$\widetilde{\boldsymbol{S}}$ method is the most efficient, with respect to MSE, among those considered here. It can be readily applied to non-response problems, as long the missing at random (MAR) holds. The CML+$\widetilde{\boldsymbol{S}}$ method is even more appealing when there is extra information that was not used in any part of the study, but is available for part of the population of interest. This extra information, maybe a cheap or an easy-to-measure covariate (represented in our simulations by a binary coarsening), is usually ignored by other efficient methods, such as Scott and Wild (1997) or Wang and Zhou (2006). CML+$\widetilde{\boldsymbol{S}}$, on the other hand, makes use of this extra observation in a fairly simple way, reducing the MSE by a significant amount.

4

Conditional Maximum Likelihood for Continuous Response

In chapter 3 we considered cases where the response was discrete. Often, however, the response follows a continuous distribution. One approach is to discretize the response Y into K mutually exclusive intervals and define a discrete variable as a vector of size K and entries equals to 0, 1, 2, ..., $K-1$. The problem is then similar to before and all methods discussed in 3 can be applied for this new discrete response. However, discretizing Y into K intervals may lead to loss of information. In addition, different discretizations may lead to different conclusions and so it is of interest to obtain methods that does not depend on the cut-off points chosen.

We start this chapter by reviewing a few methods that deal with continuous outcome and present a new approach based on results derived on chapter 2. As in previous chapters, the proposed method is shown to be more flexible than those available in the literature, especially when there are extra information that have not been used for selecting the phase-2 sample. These extra information are treated non-parametrically,

improving its robustness against model misspecification. We work with different error distributions and study efficiency and robustness through simulations. Finally, based on Lumley (2013), we discuss the performance of the proposed method under nearly-true models.

4.1 Related research

Discretizing Y into K mutually exclusive intervals has been widely considered in the literature (Zhou et al., 2002; Wang and Zhou, 2006; Song et al., 2009). These authors define $K-1$ cut points, say $c_i^{(y)}$, $i=1,\ldots,K-1$, for Y, to produce a discrete variable Y_d. For example, if $K=2$, we have only 1 cut point and Y_d is the binary coarsening taking values 0 if $Y < c_1^{(y)}$ or 1 otherwise.

Zhou et al. (2002), for example, assumes that the domain of Y is given by the union of K mutually exclusive intervals $C_k = (c_{k-1}^{(y)}, c_k^{(y)})$, $k=1,\ldots,K$, defined by the cut points $c_0^{(y)} = -\infty < c_1^{(y)} < \cdots < c_{k-1}^{(y)} < \infty = c_k^{(y)}$. Their data is obtained via a simple random sample of size n_0 and an additional simple random sample of size n_k from each one of the K intervals C_k, and work with the observed likelihood

$$L(\boldsymbol{\beta}, g) = \left\{ \prod_i^{n_0} f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) g(\mathbf{x}_i) \right\} \times \left\{ \prod_{k=1}^{K} \prod_j^{n_k} f(y_{kj}, x_{kj}|y_{kj} \in C_k;\boldsymbol{\beta}) \right\}. \tag{4.1}$$

The first term relates to the random sample and the remaining term corresponds to the joint density function of $\{(\mathbf{x}_{kj}, y_{kj}),\ j=1,\ldots,n_k\}$, conditional on $y_{kj} \in C_k$, for each k.

Using the fact that

$$f(y_{kj}, x_{kj}|y_{kj} \in C_k;\boldsymbol{\beta}) = \frac{f(y_{kj}|x_{kj};\boldsymbol{\beta}) g(x_{kj})}{\text{pr}(Y \in C_k;\boldsymbol{\beta})}$$

and

$$\text{pr}(\, Y \in C_k; \boldsymbol{\beta}) = F(c_k^{(y)}; \boldsymbol{\beta}) - F(c_{k-1}^{(y)}; \boldsymbol{\beta}),$$

where F is the cumulative distribution of f, they rewrite the likelihood (4.1) as

$$\begin{aligned} L(\boldsymbol{\beta}, g) = & \left\{ \prod_i^{n_0} f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) \times \prod_{k=1}^{K} \prod_j^{n_k} \frac{f(y_{kj}|x_{kj};\boldsymbol{\beta})}{F(c_k^{(y)}|x_{kj};\boldsymbol{\beta}) - F(c_{k-1}^{(y);\boldsymbol{\beta}}|x_{kj};\boldsymbol{\beta})} \right\} \\ & \times \left\{ \prod_i^{n_0} g(\mathbf{x}_i) \times \prod_{k=1}^{K} \prod_j^{n_k} \frac{F(c_k^{(y)}|x_{kj};\boldsymbol{\beta}) - F(c_{k-1}^{(y)}|x_{kj};\boldsymbol{\beta})}{F(c_k^{(y)};\boldsymbol{\beta}) - F(c_{k-1}^{(y)};\boldsymbol{\beta})} \times g(x_{kj}) \right\} \\ & = L_1(\boldsymbol{\beta}) \times L_2(\boldsymbol{\beta}, g) \end{aligned}$$

where $L_1(\boldsymbol{\beta})$ and $L_2(\boldsymbol{\beta}, g)$ denotes the quantities in the first and second brackets, respectively. To estimate $\boldsymbol{\beta}$, the parameter of interest, Zhou et al. propose a semiparametric empirical likelihood approach, maximizing the likelihood $L_2(\boldsymbol{\beta}, g)$ for $\boldsymbol{\beta}$ fixed to obtain the empirical likelihood of g over all distributions whose support contains the observed x-values. The resulting likelihood $L_1(\boldsymbol{\beta}) \times L_2(\boldsymbol{\beta}, \widehat{g})$ is later maximized with respect to $\boldsymbol{\beta}$.

Song et al. (2009) discuss a similar problem but for a 2-phase sampling scheme. Here, the authors assume that the outcome Y was observed for all individuals (considered the phase-1 population), but $\boldsymbol{X}$ was observed only for a sample of subjects taken from the phase-1 population, via a stratified sampling scheme (with strata defined by discretizing Y into mutually exclusive intervals). The full likelihood is

$$L(\boldsymbol{\beta}, G) = \prod_{i:R_i=1} f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) g(\mathbf{x}_i) \prod_{j:R_i=0} \int f(y_j|\mathbf{x};\boldsymbol{\beta}) dG(\mathbf{x}),$$

where $R = 1$ if $\boldsymbol{X}$ has been observed. The loglikelihood is maximized over a discrete g, where the probability is concentrated at each of the observed x-values. This provides a fully efficient approach. Let $\delta_i = \text{pr}(\boldsymbol{X} = \mathbf{x}_i)$. Since the response is continuous, how-

ever, the number of δ_is increases as the sample size increases, and hence the number of parameters is potentially as large as the sample size. In order to avoid this computational problem, the authors suggest to maximize the restricted loglikeliohood (see section 1.5.2) using a mixed Newton method. Except for computational details, this is the same approach as Jiang (2004) and Scott and Wild (2006) who deal with a more general class of 2-phase sampling.

Notice that if there is additional information available for all individuals ($\mathbf{x}$-surrogates, for example), neither Zhou et al. (2002) nor Song et al. (2009) make use of this additional data to get more efficient estimates. A new method that caters for this extra information is then required. Zhou et al. (2011) propose a semiparametric estimator for a 2-phase sampling scheme, with an auxiliary covariate. They assume that the continuous outcome Y as well as a continuous (or discrete) auxiliary variable W were fully observed for all individuals (phase-1 population) and additional information regarding $\boldsymbol{X}$ were obtained only for subjects sampled from each stratum defined by the partition of the domain of $Y \times W$. The authors treat the marginal distribution of W non-parametrically, but assume a parametric model for the conditional distribution $G(\mathbf{x}|w)$. However, since $\boldsymbol{X}$ is usually high dimensional, modelling its distribution may be hard or even infeasible so that less parametric approaches are also of interest, particularly if we want to cope with multidimensional W.

4.2 CML approach

Here we propose a semiparametric estimator that takes into account all extra information obtained for the phase-1 population. Unlike in Zhou et al. (2011), our approach does not require any assumption regarding the conditional distribution of $\boldsymbol{X}|\,W$ and only the model of interest $f(y|\mathbf{x},\boldsymbol{\beta})$ and the selection model $\pi = \text{pr}(R=1|\text{ phase-1 data})$ are

treated parametrically. Recall that under 2-phase designs these probabilities are known, being set by the study designer. The models we use are "supermodels" containing these known probabilities.

We use the conditional maximum likelihood (CML) method, which works for both discrete and binary W, without using a Kernel estimator that is intrinsically dependent on the bandwidth chosen. As will be seen in our simulations (in section 4.3.2), we are able to improve our estimates by a significant amount when extra information is used. Since we do not assume a stratified sampling scheme, the proposed method also works for non-response problems, as long the missing at random (MAR) assumptions holds.

The proposed method is similar to that discussed in chapter 2, but for a continuous response. And since all the methods laid out in that chapter do not make any assumption regarding the distribution of Y, they are all readily extended to this case. We now work with the conditional loglikelihood

$$\ell_c(\boldsymbol{\beta};\boldsymbol{\alpha}) = \sum_i R_i \log f_c(y_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\alpha}), \tag{4.2}$$

where

$$f_c(y_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\alpha}) = \frac{\pi_i(\mathbf{z}_i;\boldsymbol{\alpha}) f(y_i|\mathbf{x}_i;\boldsymbol{\beta})}{\int \pi_i(\mathbf{z}_i;\boldsymbol{\alpha}) f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) dy},$$

$f(y_i|\mathbf{x}_i;\boldsymbol{\beta})$ is the model of interest, $\pi_i(\mathbf{z}_i;\boldsymbol{\alpha})) = \text{pr}(R_i = 1|\mathbf{z}_i;\boldsymbol{\alpha})$ and $\boldsymbol{Z}$ is a function of all variables fully observed at phase-1. We note that $\boldsymbol{Z}$ must contain all variables used for selecting the phase-2 sample to ensure unbiased estimating equations for $\boldsymbol{\beta}$. As discussed in chapters 2 and 3, we can extract more information from the complete observed data $\{i : R_i = 1\}$ by calculating

$$\widetilde{\boldsymbol{S}}_1 = \frac{\partial \ell(\boldsymbol{\beta};\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}.$$

For a 2-phase problem, the parameters $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ are estimated by setting the enlarged estimating equations

$$\boldsymbol{S}(\boldsymbol{\phi}) = \boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{pmatrix} \boldsymbol{S}_0(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ \boldsymbol{S}_1(\boldsymbol{\alpha}) - \widetilde{\boldsymbol{S}}_1(\boldsymbol{\beta}, \boldsymbol{\alpha}) \end{pmatrix}$$

to zero, where

$$\boldsymbol{S}_0 = \frac{\partial \ell(\boldsymbol{\beta}; \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}},$$

$$\boldsymbol{S}_1 = \frac{\partial}{\partial \boldsymbol{\alpha}} \sum_i \Big(R_i \log(\pi_i(\mathbf{z}_i; \boldsymbol{\alpha})) + (1 - R_i) \log(1 - \pi_i(\mathbf{z}_i; \boldsymbol{\alpha})) \Big)$$

as in equation (2.8) and $\widetilde{\boldsymbol{S}}_1$ as given by equation (4.2). As discussed in chapter 2, this is equivalent to working with the pseudo-loglikelihood

$$\ell(\boldsymbol{\beta}; \boldsymbol{\alpha}) = \sum_i R_i \log(f_c(y_i | \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\alpha})) - \sum_i \Big(R_i \log(\pi_i(\mathbf{z}_i; \boldsymbol{\alpha})) + (1 - R_i) \log(1 - \pi_i(\mathbf{z}_i; \boldsymbol{\alpha})) \Big).$$

4.2.1 Distributions covered

Here we work with the linear model

$$Y = \mathbf{x}\boldsymbol{\beta} + \sigma\epsilon,$$

assuming different distributions for the error ϵ and, without loss of generality, $\sigma = 1$. We initially used a normal error distribution, but as discussed by Hampel et al. (1986) and Hill and Dixon (1982), analysis of real data usually rejects the normality assumption and thus models that are not quite normally distributed are also of interest. To this end, we also worked with three more error models: the generalized normal, skew-normal and t-distribution.

Generalized normal distribution

The generalized normal distribution (Subbotin, 1923; Box, 1953) is a symmetric generalization of the normal distribution, obtained by adding a shape parameter to the function. Its density function is given by

$$\frac{\theta}{2\sigma\Gamma(1/\theta)}\exp\left\{-\left(\frac{|y-\mu|}{\sigma}\right)^{\theta}\right\}$$

where μ is the location, σ is the scale and θ is the shape parameter. Note that this family of distribution reduces to

- the normal distribution with mean μ and variance σ^2 when $\theta = 2$;
- the Laplace distribution when $\theta = 1$;
- the uniform distribution $U(\mu-\sigma, \mu+\sigma)$ when $\theta \to \infty$.

Applying the generalized normal to our problem, we have that the expected value μ is equals to $\mathbf{x}\boldsymbol{\beta}$ and the loglikelihood is

$$\ell(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha}) = \sum_i R_i \log(f_c(y_i|\mathbf{x}_i;\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha}) - \sum_i \Big(R_i \log(\pi_i(\mathbf{z}_i;\boldsymbol{\alpha})) + (1-R_i)\log((1-\pi_i(\mathbf{z}_i;\boldsymbol{\alpha})))\Big) \tag{4.3}$$

and the estimates are obtained by setting the enlarged estimating equations

$$\boldsymbol{S}(\boldsymbol{\phi}) = \boldsymbol{S}(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha}) = \begin{pmatrix} \boldsymbol{S}_0(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha}) \\ \boldsymbol{S}_1(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha}) \\ \boldsymbol{S}_2(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha}) \\ \boldsymbol{S}_3(\boldsymbol{\alpha}) - \widetilde{\boldsymbol{S}}_3(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha}) \end{pmatrix}, \tag{4.4}$$

where

$$\boldsymbol{S}_0 = \frac{\partial \ell(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}}, \quad \boldsymbol{S}_1 = \frac{\partial \ell(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha})}{\partial \sigma}, \quad \boldsymbol{S}_2 = \frac{\partial \ell(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha})}{\partial \theta}$$

and

$$\boldsymbol{S}_3(\boldsymbol{\alpha}) - \widetilde{\boldsymbol{S}}_3(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha}) = \frac{\partial \ell(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}.$$

In particular,

$$\boldsymbol{S}_0(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{\alpha}) = \sum_i \left(\mathbf{x}_i \theta \left(\frac{1}{\sigma}\right)^{\theta} \frac{\left((y-\mathbf{x}_i\boldsymbol{\beta})^2\right)^{\theta/2}}{y-\mathbf{x}_i\boldsymbol{\beta}} - \mathbf{x}_i \frac{\int \pi_{1i} g'(y_i|\mathbf{x}_i) dy_i}{\int \pi_{1i} g(y_i|\mathbf{x}_i)\} dy_i} \right), \tag{4.5}$$

where

$$g = \exp\left\{ -\left(\frac{|y-\mathbf{x}\boldsymbol{\beta}|}{\sigma}\right)^{\theta} \right\}$$

and

$$g' = \theta \left(\frac{1}{\sigma}\right)^{\theta} (y_i - \mathbf{x}_i\boldsymbol{\beta})((y_i-\mathbf{x}_i\boldsymbol{\beta})^2)^{(\theta-2)/2} \exp\left\{ -\left(\frac{1}{\sigma}\right)^{\theta} ((y_i-\mathbf{x}_i\boldsymbol{\beta})^2)^{\theta/2} \right\}.$$

To calculate the integral (4.5), we used the Gauss-Hermite quadrature, which is a numerical method for approximating Gaussian integrals. It works as follows:

$$\int_{-\infty}^{\infty} \exp\{-t^2\} f(t) dt \approx \sum_{i=1}^{n} w_i f(r_i),$$

where n is the number of sample points used for the approximation, $r_i s$ are the roots of the Hermite polynomial $H_i(x)$, for $i = 1, \ldots, n$, and the weights $w_i s$ are given by

$$w_i = \frac{2^{n-1} n! \sqrt{\pi}}{n^2 \left(H_{n-1}(\mathbf{x}_i)\right)^2}.$$

Note that this method can be used more generally, for any function $g(t)$. That is, suppose that our goal is to calculate $\int g(t)dt$, for $t \in (-\infty, \infty)$. We can use the Gauss-Hermite quadrature method by simply multiplying the integrand $g(t)$ by $\exp\{-t^2\}\exp\{t^2\}$, resulting in the integral $\int \exp\{-t^2\}h(t)dt$, where $h(t) = g(t)\exp\{t^2\}$. This method produces accurate results if the integrand can be approximated by a polynomial of degree $2n-1$.

In our case, by doing a change of variables $u_i = (y_i - \mathbf{x}_i\boldsymbol{\beta})/\sqrt{2}$, we get

$$\int \pi_{1i}(y_i, \mathbf{x}_i)\exp\{-(y_i-\mathbf{x}_i\boldsymbol{\beta})^2/2\}(y_i-\mathbf{x}_i\boldsymbol{\beta})dy_i = 2\int \pi_{1i}((\sqrt{2}u_i+\mathbf{x}_i\boldsymbol{\beta}), \mathbf{x}_i)\exp\{-u_i^2\}u_i du_i$$

and

$$\int \pi_{1i}(y_i, \mathbf{x}_i)\exp\{-(y_i-\mathbf{x}_i\boldsymbol{\beta})^2/2\}dy_i = \sqrt{2}\int \pi_{1i}((\sqrt{2}u_i+\mathbf{x}_i\boldsymbol{\beta}), \mathbf{x}_i)\exp\{-u_i^2\}du_i$$

for the two integrals of equation (4.5). Using the Gauss-Hermite quadrature, we have that

$$\int \pi_{1i}(y_i, \mathbf{x}_i)\exp\{-(y_i-\mathbf{x}_i\boldsymbol{\beta})^2/2\}(y_i-\mathbf{x}_i\boldsymbol{\beta})dy_i \approx 2\sum_{j=1}^{n} w_j\pi_{1i}((\sqrt{2}r_j+\mathbf{x}_i\boldsymbol{\beta}), \mathbf{x}_i)\exp\{-r_i^2\}r_i$$

and

$$\int \pi_{1i}(y_i, \mathbf{x}_i)\exp\{-(y_i-\mathbf{x}_i\boldsymbol{\beta})^2/2\}dy_i \approx \sqrt{2}\sum_{j=1}^{n} w_j\pi_{1i}((\sqrt{2}r_ij+\mathbf{x}_i\boldsymbol{\beta}), \mathbf{x}_i)\exp\{-r_i^2\},$$

and so the first term of the sum of (4.5) can be approximated. An advantage of working with Gauss-Hermite quadrature in R is that calculations can be vectorized so that all of the integrations can be performed simultaneously with a single summation loop. We can now solve the set of estimating equations equation (2.9), where the Newton-Raphson

method can be used for convergence.

The unknown shape parameter θ was also estimated by maximum likelihood, as shown by the estimating equations above, via the Newton-Raphson method. For the initial value we follow the procedure developed by Domínguez-Molina et al. (2001), which uses the method of moments. The authors first show the existence of the method of moments estimator and, by considering the first two absolute moments and using some properties of the generalized normal distribution, derive the formula

$$M(\widehat{\theta}) = \overline{M}(y), \quad \text{where} \quad \overline{M}(y) = \frac{\left(\frac{1}{N}\sum_{i=1}^{N}|y_i - \mu_y|\right)^2}{\frac{1}{N}\sum_{i=1}^{N}|y_i - \mu_y|^2}$$

and Y follows a generalized normal distribution. Then, $\widehat{\theta} = M^{-1}\left[\overline{M}(y)\right]$, where $M(\theta)$ can be approximated by a proposed function $M^*(\theta)$ that can be inverted in a wide range of values of θ. Using their approximation we are then able to get a first approximation to the parameter θ and use the Newton-Raphson method until convergence.

Skew normal distribution

The Skew normal distribution was first introduced by O'Hagan and Leonard (1976) and has density function $f(y|\mathbf{x})$ given by

$$f(y|\mathbf{x}) = \frac{1}{\sigma\pi}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}\int_{-\infty}^{\kappa\left(\frac{y-\mu}{\sigma}\right)}\exp\left\{-\frac{t^2}{2}\right\}dt,$$

where $\mu, \in \mathbb{R}$ is a location, $\sigma \in \mathbb{R}_+$ a scale and $\kappa \in \mathbb{R}$, a shape parameter. Note that κ controls the skewness: for $\kappa > 0$, the distribution is right skewed and for $\kappa < 0$, left skewed; for $\kappa = 0$, it reduces to the usual normal distribution.

The loglikelihood is given by equation (4.3) and the enlarged estimating equations

system is given by equation (4.4), with θ replaced by κ.

In particular,

$$\boldsymbol{S}_0(\boldsymbol{\beta},\sigma,\kappa,\boldsymbol{\alpha}) = \sum_i \left(\mathbf{x}_i \left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma} - \frac{\kappa}{\sigma} \exp\left\{ -\left(\frac{\kappa(y_i - \mathbf{x}_i\boldsymbol{\beta})}{\sigma} \right)^2 \right\} \left(\int_{-\infty}^{\kappa\left(\frac{y-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)} \exp\left\{ -\frac{t^2}{2} \right\} dt \right)^{-1} \right) - \mathbf{x}_i \frac{\int \pi_{1i} g'(y_i|\mathbf{x}_i) dy_i}{\int \pi_{1i} g(y_i|\mathbf{x}_i)\} dy_i} \right),$$

where

$$g(y_i|\mathbf{x}_i) = \exp\left\{ -\left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma} \right)^2 \right\} \int_{-\infty}^{\kappa\left(\frac{y-\mu}{\sigma}\right)} \exp\left\{ -\frac{t^2}{2} \right\} dt$$

and

$$g'(y_i|\mathbf{x}_i) = \left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma^2} \right) \exp\left\{ -\left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma} \right)^2 \right\} \int_{-\infty}^{\kappa\left(\frac{y-\mu}{\sigma}\right)} \exp\left\{ -\frac{t^2}{2} \right\} dt - \frac{\kappa}{\sigma} \exp\left\{ -\left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma} \right)^2 \right\} \exp\left\{ -\left(\frac{\kappa(y_i - \mathbf{x}_i\boldsymbol{\beta})}{\sigma} \right)^2 \right\}.$$

Gauss-Hermite quadrature was again used to approximate the integrals, for the enlarged system of estimating equations. The estimated coefficients were later computed using the Newton-Raphson algorithm.

T-distribution

We also derived estimating equations for the T-distribution

$$\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)} \left\{ 1 + \frac{1}{\upsilon} \left(\frac{y-\mu}{\sigma} \right)^2 \right\}^{-\frac{\upsilon+1}{2}}.$$

The loglikelihood follows the same structure as (4.3), resulting in an enlarged estimating equations system that is similar to (4.4).

In particular,

$$\boldsymbol{S}_0(\boldsymbol{\beta},\sigma,\upsilon,\boldsymbol{\alpha}) = \sum_i \left[\mathbf{x}_i \left(\frac{\upsilon+1}{\upsilon} \right) \frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})}{\sigma^2} \left\{ 1 + \frac{1}{\upsilon} \left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma} \right)^2 \right\}^{-\frac{\upsilon-1}{2}} - \mathbf{x}_i \frac{\int \pi_{1i} g'(y_i|\mathbf{x}_i) dy_i}{\int \pi_{1i} g(y_i|\mathbf{x}_i) dy_i} \right],$$

where

$$g(y_i|\mathbf{x}_i) = \left\{ 1 + \frac{1}{\upsilon} \left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma} \right)^2 \right\}^{-(\upsilon+1)/2}$$

and

$$g'(y_i|\mathbf{x}_i) = \left(\frac{\upsilon+1}{\upsilon} \right) (y_i - \mathbf{x}_i\boldsymbol{\beta}) \left[1 + \left\{ \frac{1}{\upsilon} \left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma} \right)^2 \right\} \right]^{-(\upsilon+3)/2}.$$

4.3 Simulations

We evaluate the performance of the proposed method against more common approaches in many different scenarios. For comparison, we used the following methods:

- Full data analysis (*full*), where the entire data was used as if there were no missing data. This corresponds to the ideal situation where there are no missing information and is used here as a measure of how much efficiency, with respect to mean square error (MSE), the missing values lead to.

- The complete case analysis (*sample*), where only the fully observed variables were used to fit the model of interest. As discussed in chapter 1, this approach can be seriously biased in case of an outcome-dependent sampling scheme or if there are a large percentage of missing values.

- Multiple imputation (*MI*), where the package *Hmisc* from R was used, as in chapter 3, to generate and analyse the imputed dataset;

- The calibration method (*cal*), as discussed in section 1.6.1. That is, we first used

a linear model to predict the missing values using the fully observed variables. We then fitted the model of interest $f(y|\mathbf{x};\boldsymbol{\beta})$ using the complete data and the influence functions as auxiliary variables to estimate the coefficients of interest $\boldsymbol{\beta}$. For the analysis we used the *survey* package from R.

- The weighted (*wgt*) and the CML (*cml*) methods;
- The CML method with the additional information $\widetilde{\boldsymbol{S}}$ added, considering all (*cml+S*) or just the discrete variables actually used in the selection model (*cml*+S*). This is done so that we can investigate the impact of adding continuous variable and other variables into the selection model for π.

For the simulations, we used a 2-phase sampling scheme in which Y (and its "surrogate" Y_d) and X_1 (and its surrogate X_{1d}) had been fully observed at phase-1. Here, X_{1d} is a discrete version of X_1, being equal to 1 for extreme values of X_1 and zero otherwise. Extreme is defined as values smaller or greater than the $15th$ and $85th$ quantiles, respectively. Y_d, the "surrogate" of Y was defined as an indicator variable equals to if Y less or equal than its $15th$ percentile. The phase-2 sample was taken from each stratum defined by (Y_d, X_{1d}) and the remaining variable X_2, observed. Figure 4.1 illustrates the sampling scheme just described. For the simulations that follow, we analysed using the Generalized-Normal-errors algorithm, which estimates the shape parameter θ. The results for the skew and t-distribution are given in appendix A.

4.3.1 Varying the parameter of interest

In this section we ran a series of simulations varying the parameter of interest $\boldsymbol{\beta}$ in order to investigate how the efficiency of each method is affected by different degrees of association between the response Y and the covariates X_1 and X_2. We discuss situations

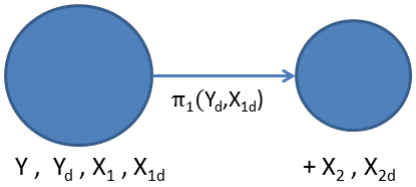


Figure 4.1: Sampling scheme for a 2-phase study, where the response Y, a covariate X_1 and a surrogate variable X_{1d} for X_1 are fully observed at phase-1. An extra covariate X_2 is observed only at the phase-2 of the study

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$	cor(Y, X_1)	cor(Y, X_2)	Phase-2 sample size
(1,.5,.5)	0.414	0.445	800
(1,1,1)	0.589	0.592	800
(1,2,2)	0.670	0.661	800
(1,1,.5)	0.673	0.374	800
(1,1,2)	0.412	0.827	800
(1,1,3)	0.302	0.922	800
(1,.5,1)	0.329	0.688	800
(1,2,1)	0.816	0.389	800

Table 4.1: Range of $\boldsymbol{\beta}$ and correlation between the Y and X_1 and between Y and X_2.

where Y and X_1 or Y and X_1 are weakly or strongly correlated. The range of values covered as well as the phase-2 sample size, are shown in table 4.1.

For all simulations in this chapter, unless stated otherwise, we ran 1000 simulations and compared each method based on their MSE, using a total population of $N = 15,000$ and phase-2 data containing 800 units. Notice that the true selection probability π_1 is a function of (Y_d, X_{1d}) but we can use more information by including the fully observed variables Y and X_1 into the selection model.

For the weighted, CML and CML+$\widetilde{\boldsymbol{S}}$ method, we fitted the selection model

$$\text{logit}(\pi) \sim y_d * x_{1d} + y * x_1.$$

In order to see the impact of adding the continuous variables Y and X_i in π, we used

the CML+$\widetilde{\boldsymbol{S}}$ method with the minimal selection model

$$\text{logit}(\pi) \sim y_d * x_{1d}.$$

Recall that $y_d * x_{1d}$ is equivalent to $y + x_{1d} + yx_{1d}$ and similarly for $y_d * x_1$. We denoted this method by CML*+$\widetilde{\boldsymbol{S}}$. The model of interest is correctly fitted in all cases and the results for 1000 simulations are shown in table 4.2.

The complete case analysis, which does take the biased sampling into consideration, gives biased estimates, large MSE and poor coverage. The remaining methods are all approximately unbiased because all models are correctly specified. The full data method gives much better estimates in the sense that the MSE is about 10 or 15 times lower than the second most efficient approach. That shows that the amount of information lost due to missing observations can significantly impact the estimates of interest.

The most efficient methods are multiple imputation and CML+$\widetilde{\boldsymbol{S}}$. MI gives better estimates for β_1 and β_0 when the effect of and X_2 is small, but become less efficient than the CML+$\widetilde{\boldsymbol{S}}$ method as the X_2-effect increases. Recall that X_2 is the variable that is missing at phase-1. With respect to β_2, both methods give estimates with similar MSE, but as β_2 increases, CML+$\widetilde{\boldsymbol{S}}$ becomes much more efficient than MI.

As expected, CML+$\widetilde{\boldsymbol{S}}$ is significantly more efficient than the weighted method. The CML+$\widetilde{\boldsymbol{S}}$ method is also more efficient than calibration, but only slightly more efficient than CML. That is, the extra term $\widetilde{\boldsymbol{S}}$ is not carrying a large amount of information here, but enough to produce better estimates. The former method is usually 10% more efficient than the latter when β_2 is small, but almost equally efficient for large β_2.

Compared to CML*+$\widetilde{\boldsymbol{S}}$, however, we see that CML+$\widetilde{\boldsymbol{S}}$ is far more efficient, especially when β_2 is small. That is, even though we are fitting the right selection model

Table 4.2: Varying $\boldsymbol{\beta}$ and ϵ following a normal distribution (1000 datasets simulated).

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (1, .5, .5)^T$	full	0.000	0.000	0.000	0.994	0.995	1.020	0.067	0.067	0.064	0.948	0.955	0.955
	sample	-0.460	0.075	0.072	1.205	0.988	1.061	212.999	6.845	6.443	0.000	0.437	0.496
	MI	0.003	0.001	-0.005	0.754	0.978	0.713	0.501	0.319	1.251	0.834	0.918	0.797
	cal	0.002	0.003	0.002	1.022	0.949	1.027	1.195	1.163	1.425	0.957	0.945	0.953
	wgt	0.000	0.001	0.002	0.997	0.954	1.000	0.631	0.861	1.486	0.944	0.940	0.946
	cml	0.003	-0.002	0.001	1.218	1.056	1.118	0.692	0.576	1.195	0.980	0.963	0.968
	cml*+S	0.002	-0.003	0.003	1.151	1.052	1.153	1.117	1.197	1.146	0.977	0.959	0.978
	cml+S	0.005	-0.001	0.004	1.000	0.992	1.135	0.601	0.532	1.078	0.946	0.956	0.972
$\boldsymbol{\beta} = (1, 1, 1)^T$	full	0.000	0.000	0.000	0.988	0.957	1.036	0.068	0.073	0.062	0.936	0.953	0.958
	sample	-0.293	0.061	0.064	1.057	1.001	1.021	87.489	4.765	5.306	0.000	0.536	0.545
	MI	0.005	0.001	-0.001	0.742	0.965	0.707	1.080	0.772	0.915	0.819	0.892	0.807
	cal	0.004	0.001	-0.001	0.994	1.008	1.012	1.553	1.305	1.510	0.947	0.946	0.948
	wgt	0.003	0.002	0.000	1.014	1.015	1.018	1.077	1.162	1.481	0.948	0.946	0.953
	cml	0.004	0.000	0.000	1.135	1.023	1.058	1.019	0.859	1.204	0.966	0.951	0.957
	cml*+S	0.006	-0.001	0.002	1.094	1.005	1.057	1.407	1.169	1.229	0.965	0.951	0.955
	cml+S	0.004	0.001	0.002	1.076	1.020	1.050	0.972	0.829	1.146	0.962	0.954	0.950
$\boldsymbol{\beta} = (1, 2, 2)^T$	full	0.000	0.000	0.000	0.996	0.979	0.990	0.067	0.070	0.068	0.944	0.941	0.950
	sample	-0.194	-0.036	0.113	1.033	1.021	1.001	39.185	2.261	13.839	0.004	0.790	0.064
	MI	0.004	0.000	0.002	0.739	0.917	0.617	1.666	1.559	1.360	0.806	0.880	0.747
	cal	0.002	0.001	-0.001	1.022	0.946	0.952	1.700	1.845	1.817	0.953	0.928	0.941

Continued on Next Page...

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
	wgt	0.001	0.001	0.000	1.023	0.958	0.949	1.502	1.773	1.806	0.958	0.933	0.937
	cml	0.003	0.000	0.000	1.066	1.018	0.964	1.549	0.989	1.272	0.961	0.956	0.940
	cml*+S	0.005	0.000	0.001	1.021	0.986	0.963	1.837	1.090	1.296	0.950	0.948	0.940
	cml+S	0.004	0.000	0.001	1.053	1.017	0.969	1.534	0.986	1.237	0.960	0.951	0.936
$\boldsymbol{\beta} = (1, 1, .5)^T$	full	0.000	0.000	0.000	1.007	0.993	0.988	0.066	0.068	0.069	0.955	0.947	0.941
	sample	-0.357	0.093	0.043	1.136	1.015	1.014	128.594	9.716	3.150	0.000	0.195	0.784
	MI	0.001	0.000	-0.009	0.809	0.938	0.740	0.438	0.358	1.324	0.858	0.904	0.800
	cal	0.001	0.000	-0.002	1.011	0.973	1.015	1.364	1.209	1.556	0.951	0.939	0.948
	wgt	-0.001	0.001	-0.001	1.017	0.979	1.010	0.660	1.012	1.566	0.943	0.942	0.952
	cml	0.001	-0.001	0.000	1.128	0.981	1.117	0.675	0.804	1.202	0.973	0.939	0.969
	cml*+S	0.002	-0.004	0.002	1.118	0.976	1.093	1.208	1.230	1.263	0.974	0.944	0.970
	cml+S	0.002	0.001	0.001	1.049	0.970	1.104	0.545	0.765	1.183	0.950	0.942	0.969
$\boldsymbol{\beta} = (1, 1, 2)^T$	full	0.000	0.000	0.000	0.992	0.989	0.966	0.068	0.068	0.072	0.943	0.951	0.944
	sample	-0.220	-0.006	0.085	1.017	1.033	1.016	49.857	1.074	8.327	0.000	0.957	0.263
	MI	0.002	0.000	0.002	0.652	0.945	0.625	1.827	1.257	1.282	0.748	0.884	0.744
	cal	0.000	0.000	0.000	0.998	0.983	0.975	1.698	1.665	1.566	0.949	0.944	0.937
	wgt	0.000	0.000	0.000	0.989	0.990	0.973	1.562	1.589	1.553	0.944	0.940	0.936
	cml	0.003	0.001	0.000	1.033	1.024	0.965	1.568	1.075	1.204	0.953	0.954	0.937
	cml*+S	0.003	0.000	0.000	1.006	0.989	0.944	1.778	1.252	1.279	0.950	0.951	0.929
	cml+S	0.003	0.000	0.001	1.018	1.021	0.960	1.559	1.064	1.185	0.947	0.958	0.931
$\boldsymbol{\beta} = (1, 1, 3)^T$	full	0.000	0.000	0.000	1.015	0.992	0.959	0.065	0.068	0.072	0.951	0.955	0.940
	sample	-0.158	-0.018	0.067	1.021	0.977	1.023	26.496	1.653	5.480	0.034	0.916	0.444

Continued on Next Page...

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
	MI	0.004	0.000	0.008	0.619	0.906	0.540	2.054	1.500	1.753	0.735	0.873	0.686
	cal	0.001	0.000	0.002	1.022	0.958	0.963	1.685	1.867	1.591	0.955	0.939	0.939
	wgt	0.001	0.000	0.002	1.021	0.970	0.970	1.608	1.794	1.550	0.959	0.941	0.939
	cml	0.003	-0.001	0.002	0.986	0.989	0.976	1.738	1.322	1.096	0.937	0.946	0.948
	cml*+S	0.004	-0.001	0.002	0.994	0.986	0.975	1.785	1.371	1.108	0.947	0.945	0.949
	cml+S	0.004	-0.001	0.003	0.972	0.974	0.971	1.771	1.355	1.099	0.947	0.943	0.948
$\boldsymbol{\beta} = (1, .5, 1)^T$	full	0.001	0.001	0.000	1.016	1.023	1.027	0.065	0.064	0.063	0.952	0.951	0.960
	sample	-0.362	0.042	0.095	1.119	1.026	1.022	132.549	2.860	10.316	0.000	0.777	0.246
	MI	0.004	0.001	0.001	0.736	0.928	0.683	1.066	0.790	1.039	0.808	0.895	0.794
	cal	0.002	0.000	0.002	1.037	0.989	1.009	1.349	1.289	1.455	0.956	0.941	0.949
	wgt	0.002	0.000	0.003	1.033	0.984	1.008	1.006	1.122	1.446	0.966	0.942	0.949
	cml	0.003	-0.002	0.004	1.209	1.145	1.011	1.011	0.701	1.287	0.983	0.972	0.951
	cml*+S	0.003	-0.004	0.006	1.140	1.112	1.007	1.300	1.108	1.352	0.972	0.969	0.948
	cml+S	0.005	-0.001	0.007	1.086	1.065	0.998	1.003	0.716	1.207	0.966	0.965	0.949
$\boldsymbol{\beta} = (1, 2, 1)^T$	full	0.000	0.000	0.001	1.012	0.989	0.983	0.065	0.068	0.069	0.952	0.949	0.942
	sample	-0.166	0.009	0.052	1.032	1.047	1.004	29.164	0.866	3.958	0.013	0.952	0.684
	MI	0.002	0.002	-0.001	0.800	0.928	0.748	1.080	0.964	0.757	0.856	0.880	0.815
	cal	0.000	0.001	0.002	1.014	0.995	0.971	1.615	1.461	1.812	0.955	0.951	0.935
	wgt	0.000	0.001	0.003	1.018	1.001	0.968	1.120	1.407	1.828	0.950	0.962	0.936
	cml	0.004	0.000	0.004	0.977	1.035	1.067	1.320	0.728	1.216	0.941	0.955	0.955
	cml*+S	0.002	0.000	0.004	0.992	1.041	1.066	1.752	0.799	1.231	0.953	0.956	0.957
	cml+S	0.004	0.000	0.005	0.919	1.028	1.069	1.378	0.731	1.193	0.929	0.954	0.952

(which is a function of Y_d and X_{1d} only), by adding the continuous variables Y and X_1 we are able to get much better estimates. When $\beta_2 = .5$, for example CML+$\widetilde{\boldsymbol{S}}$ is about 2 times more efficient for estimating β_0 or β_1 than CML*+$\widetilde{\boldsymbol{S}}$, but only about 10% when β_2 is large.

Apart from the MI method, all others methods show good standard error estimates. While MI seems to underestimate the standard error, for the remaining methods the ratios between the estimated (Est.SE) and empirical (Emp.SE) standard errors are close to 1. This also results in poor coverage for MI and coverage close to the nominal value for the other methods.

Recall that the Generalized Normal Distribution, which was used to fit the model of interest, reduces to the usual Normal distribution when the shape parameter θ is equals to 2. Therefore, in order to quantify the amount of precision lost while using this larger model instead of the simple normal model, we performed a simulation study using the same parameters (population size, sample size and selection model) as above. The results are presented in Table A.1. Thus, for these settings, the estimates obtained through the larger and thus more robust Generalized Normal are almost as efficient as fitting a simple Normal distribution. The differences between them are never bigger than 10%.

Finally, we performed another simulation studied, varying not only the parameter $\boldsymbol{\beta}$ but also the error distribution. Results for the skew and t-distribution are given in appendix A, for the same settings as here, and follow the same pattern as obtained from the Generalized Normal Distribution.

4.3.2 Adding extra information

We will now bring in observation of a binary coarsening of X_2 at phase-1. Let the indicator variable X_{2d}, which is equal to 1 if $X_2 < .5$ and 0 otherwise, be fully observed at phase-1. The selection model is the same as before, i.e., depends only on (Y_d, X_{1d}), and the error is again normally distributed. Our goal here is to use this extra information and see the impact of adding this variable into the selection model even though it was not actually used for selecting the phase-2 data. The selection model used for the weighted, CML and CML+$\widetilde{\boldsymbol{S}}$ methods is then given by

$$\text{logit}(\pi) \sim y_d * x_{1d} * x_{2d} + y * x_1 * x_{2d}$$

while the selection model for CML*+$\widetilde{\boldsymbol{S}}$ is

$$\text{logit}(\pi) \sim y_d * x_{1d} * x_{2d}.$$

We ran 1000 simulations for three different values of β_2, the parameter associated with X_2 so that we could see how would affect the estimation of $\boldsymbol{\beta}$ when Y and x_2 were weakly or strongly correlated. The results are shown in table 4.3.

Consider first estimating the parameter β_1. In the first case where $\beta_2 = .5$, MI is the best method with a MSE that is only 2 times larger than the one obtained if there was no missing data. It is also about 40% lower than the MSE obtained by the calibration method and almost 70% lower than CML+$\widetilde{\boldsymbol{S}}$. As β_2 increases, however, the CML+$\widetilde{\boldsymbol{S}}$ method becomes more efficient than MI. For $\beta_2 \geq 1$, for example, CML+$\widetilde{\boldsymbol{S}}$ is already the most efficient method, being between 1.5 and 2 times more efficient than the weighted method and between 1.3 and 2.5 times more efficient than the CML*+$\widetilde{\boldsymbol{S}}$

Table 4.3: Adding X_{2d} into the selection model (1000 datasets simulated).

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (1, 1, .5)^T$	Imp	0.003	0.002	-0.012	0.811	0.981	0.737	0.282	0.195	1.099	0.881	0.935	0.800
	cal	-0.001	0.000	0.000	0.951	0.980	0.952	0.397	0.310	0.992	0.938	0.955	0.930
	wgt	-0.001	0.002	-0.001	1.009	0.984	1.013	0.463	1.166	1.261	0.947	0.942	0.960
	cml	0.000	-0.001	0.001	1.109	0.983	1.057	0.503	0.861	0.871	0.970	0.942	0.950
	cml*+S	0.000	-0.003	0.004	1.075	1.021	1.123	1.660	1.495	1.607	0.961	0.960	0.963
	cml+S	0.001	0.001	0.001	0.950	0.984	0.967	0.379	0.608	0.496	0.923	0.945	0.942
$\boldsymbol{\beta} = (1, 1, 1)^T$	Imp	0.006	0.016	0.002	0.755	0.968	0.686	0.737	0.836	1.186	0.842	0.859	0.804
	cal	0.000	-0.001	0.001	0.990	0.957	0.942	0.889	0.840	1.269	0.943	0.933	0.934
	wgt	0.000	0.002	0.000	1.020	0.986	0.964	0.805	1.419	1.524	0.955	0.941	0.930
	cml	0.001	0.001	0.001	1.133	1.057	1.060	0.799	0.912	0.979	0.973	0.946	0.953
	cml*+S	0.002	-0.004	0.004	1.090	1.049	1.022	1.748	1.469	1.787	0.962	0.952	0.953
	cml+S	0.001	0.002	0.001	1.020	1.025	0.989	0.664	0.797	0.818	0.954	0.952	0.940
$\boldsymbol{\beta} = (1, 1, 2)^T$	Imp	0.011	0.052	0.014	0.676	0.983	0.625	1.673	3.916	1.943	0.792	0.651	0.732
	cal	0.001	0.001	0.006	0.983	0.953	0.971	1.608	1.689	1.666	0.943	0.931	0.928
	wgt	0.001	0.001	0.003	0.968	0.947	0.960	1.540	2.017	1.777	0.937	0.936	0.932
	cml	0.004	0.002	0.002	1.008	1.004	0.941	1.601	1.334	1.276	0.946	0.942	0.927
	cml*+S	0.005	0.001	0.003	0.972	1.006	0.943	2.363	1.607	1.708	0.932	0.947	0.928
	cml+S	0.004	0.003	0.002	0.971	0.998	0.967	1.504	1.278	1.150	0.939	0.934	0.939

method for estimating β_1.

Regarding the estimation of β_2, we see that CML+$\widetilde{\boldsymbol{S}}$ is the most efficient approach in all cases. For example, for $\beta_2 = .5$, the CML+$\widetilde{\boldsymbol{S}}$ method is almost 3 times more efficient than the weighted method and about 2 times more efficient than MI, CML and calibration. For large β_2, the differences between the MSEs are not as large as before, but CML+$\widetilde{\boldsymbol{S}}$ is still considerably better than the other approaches. For example, for $\beta_2 = 2$, CML+$\widetilde{\boldsymbol{S}}$ is about 1.11 times more efficient than CML and about 1.55 times more efficient than the weighted method.

4.3.3 Model misspecification

For data missing by design, as considered so far in this chapter, the selection probability π is controlled by the researcher and thus, known. Therefore, in order to investigate the robustness of the proposed method under a continuous response, we work with misspecified models of interest. We assumed that the true model was now given by

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma x_3 + \epsilon,$$

where $X_3 = X_1^2 + \epsilon_{x_3}$, for $\epsilon_{x_3} \sim N(0, 10^{-2})$. We work with the same sampling scheme as in section 4.3.1, assuming that the extra variable X_3 have also been fully observed at phase-1. As before, all variables are normally distributed with zero mean and variance 1. Table 4.4 shows the results for $\boldsymbol{\gamma} = (.01, .025, .05, .075)$, indicating small and moderate model misspecifications.

The residuals plots for fitting $y \sim x_1 + x_2$ for each value of γ, based on the phase-2 sample, are shown in Fig. 4.2. For $\gamma = .01$, a stepwise regression detects the model misspecification in about 37% of the time. For $\gamma = .025$, it was detected approximately

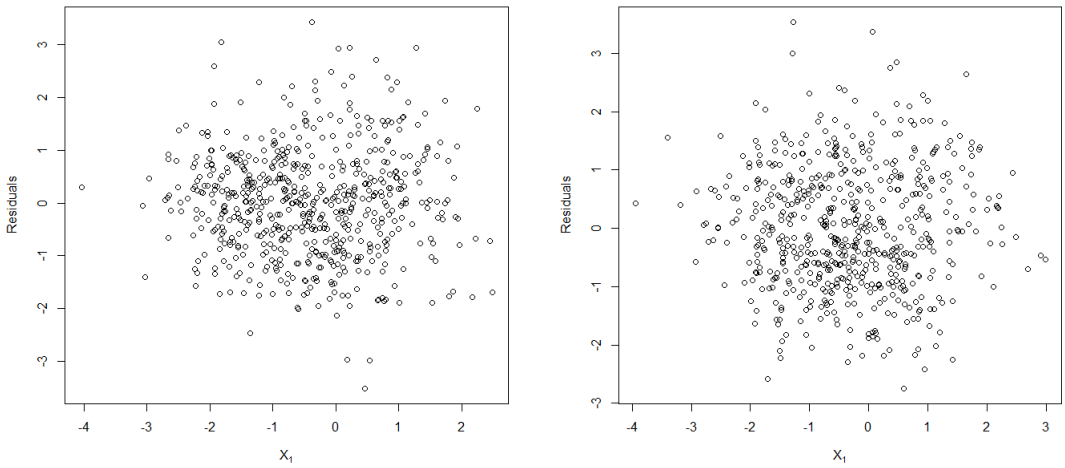
63% of the time. For $\gamma = .05$ and for the more extreme case $\gamma = .075$, the quadratic term was detected in 93% and 99% of the time, respectively.

From table 4.4 we see that as the model misspecification increases (i.e., γ increases), calibration and the weighted methods become the best alternatives to estimate the intercept and β_1. This is because both methods are more robust and show smaller bias when compared to the other methods, especially while estimating β_1. Since the misspecification is not related to β_2, the CML+$\widetilde{\boldsymbol{S}}$ method is still the best approach among all considered here, showing small MSE in all cases.

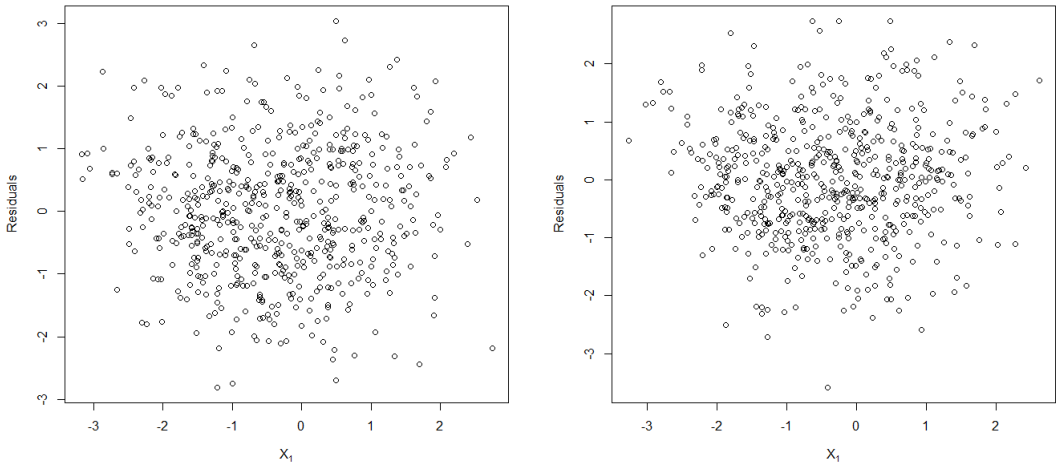
4.4 Nearly true models

It is well-known that the efficiency of maximum likelihood methods is strongly related to how well the model of interest is specified. If the model of interest is correctly specified, under mild conditions, maximum likelihood estimates (MLE) will be unbiased and fully efficient. However, as seen in previous sections, if the model is slightly misspecified these estimates will become biased, resulting in poor estimates that are, sometimes, even worse than the estimates obtained via the more robust weighted method. The question, therefore, is to determine how much misspecification is need before MLE become worse than the weighted ones and understanding this threshold is the goal of this section. Of course, in real life problems, this threshold is impossible to determine since we do not know the true model. The results in this section are, therefore, mainly for theoretical purposes, but serves as an alert for analysts since the true model is almost never known.

Model misspecification and robustness have been objects of study for about 50 years now. Tukey (1952) and Huber (1964) are considered the ones who laid the modern foundations of such a field. Robust methods are usually less efficient than the MLE



(a) (b)



(c) (d)

Figure 4.2: Residual plot for different values of γ. Figure (a) corresponds to $\gamma = .01$, (b) corresponds to $\gamma = .025$, (c) to $\gamma = .05$ and figure (d) to $\gamma = .075$.

Table 4.4: Results for fitting a misspecified model, for 1000 datasets simulated.

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (1,1,1,.01)^T$	MI	0.017	0.017	0.003	0.745	1.037	0.696	1.022	0.782	1.158	0.763	0.861	0.809
	cal	0.012	0.000	0.002	0.950	0.908	0.952	0.871	0.680	1.200	0.915	0.932	0.938
	wgt	0.011	0.002	0.001	1.031	0.976	0.993	0.906	1.446	1.442	0.936	0.948	0.946
	cml	0.015	-0.005	0.002	1.107	1.046	1.054	1.030	0.925	0.964	0.937	0.953	0.961
	cml*+S	0.015	-0.010	0.004	1.040	0.984	1.015	2.034	1.621	1.418	0.937	0.931	0.954
	cml+S	0.013	-0.003	0.001	1.026	1.008	1.027	0.815	0.811	0.735	0.928	0.949	0.955
$\boldsymbol{\beta} = (1,1,1,.025)^T$	MI	0.033	0.018	0.005	0.728	1.014	0.691	1.818	0.829	1.154	0.571	0.863	0.790
	cal	0.027	0.001	0.003	0.917	0.911	0.962	1.496	0.679	1.173	0.785	0.928	0.935
	wgt	0.027	0.003	0.001	0.958	0.976	1.000	1.618	1.443	1.405	0.831	0.939	0.931
	cml	0.031	-0.013	0.002	1.066	1.068	1.050	1.815	1.038	0.974	0.862	0.936	0.956
	cml*+S	0.027	-0.018	0.006	1.044	1.012	0.990	2.534	1.780	1.520	0.914	0.918	0.941
	cml+S	0.029	-0.013	0.002	0.954	1.019	1.017	1.576	0.959	0.747	0.784	0.917	0.946
$\boldsymbol{\beta} = (1,1,1,.05)^T$	MI	0.054	0.019	0.002	0.784	0.936	0.716	3.561	0.978	1.075	0.286	0.851	0.811
	cal	0.049	0.002	-0.001	1.011	0.859	1.013	2.998	0.779	1.052	0.516	0.915	0.939
	wgt	0.046	0.002	-0.002	1.026	0.932	1.037	2.949	1.573	1.310	0.671	0.918	0.959
	cml	0.057	-0.025	-0.001	1.174	1.060	1.098	4.024	1.534	0.903	0.604	0.886	0.962
	cml*+S	0.047	-0.026	0.008	1.370	1.156	1.459	3.949	2.119	1.488	0.825	0.886	0.942
	cml+S	0.053	-0.030	0.003	1.060	1.019	1.032	3.385	1.702	0.743	0.469	0.840	0.950
$\boldsymbol{\beta} = (1,1,1,.075)^T$	MI	0.081	0.017	0.002	0.736	1.037	0.697	7.319	0.774	1.158	0.057	0.875	0.813
	cal	0.075	0.000	0.002	0.964	0.975	0.943	6.379	0.605	1.223	0.168	0.946	0.931

Continued on Next Page...

Table 4.4 – Continued

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$	Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
wgt	0.075	0.003	0.001	0.999	1.051	0.990	6.398	1.245	1.448	0.277	0.954	0.939
cml	0.087	-0.038	0.001	1.106	1.124	1.047	8.341	2.252	0.991	0.221	0.808	0.951
cml*+S	0.070	-0.040	0.010	1.172	1.171	1.248	6.846	2.949	1.537	0.630	0.845	0.936
cml+S	0.079	-0.044	0.005	0.990	1.130	1.017	6.915	2.632	0.783	0.134	0.719	0.939

if all assumptions are satisfied, but become more efficient if small deviations from the model are present. Some methods, however, can also be asymptotically fully efficient as Robins et al. (1994), where the authors derive robust AIPW estimates that achieve the semiparametric efficiency bound asymptotically.

This bias-variance trade-off is discussed in great detail in Claeskens and Hjort (2008) and our approach follows a similar idea, applied to a 2-phase sampling scheme, as done by Lumley (2013). This bias-variance trade-off was analysed by Lumley (2013) for nearly-true models, i.e., models that are close enough to the true one in such a way that their misspecification cannot be reliably detected by available tests or diagnostics. Lumley defines two measures P and Q related to the true and nearly true (misspecified) models and is interested in testing the likelihood ratio $L = dQ/dP$, with the purpose of detecting any departures from the true model. Lumley relies on the theorem known as LeCam's Third Lemma (LeCam, 1960). Let D_n denote some statistic of interest.

Lemma 4.1. *If*

$$\left(D_n, \log \frac{dQ_n}{dP_n}\right) \xrightarrow{d} N\left(\begin{pmatrix} \mu \\ \kappa^2/2 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \tau \\ \tau^T & \kappa^2 \end{pmatrix}\right)$$

under P_n*, then*

$$D_n \xrightarrow{d} N\left(\mu + \tau, \sigma^2\right)$$

under Q_n*.*

That is, going from the true to the nearly-true model, there is a shift in the distribution but no change in scale. Note that τ can be written in terms of the correlation ρ, so that the shift is equals to $\rho\kappa\sigma$. Lumley (2013) rewrites this result in terms of $\boldsymbol{\beta}_{\mathit{eff}}$

and $\boldsymbol{\beta}_{AIPW}$ (estimators obtained via the efficient and AIPW methods), as

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{AIPW}-\boldsymbol{\beta}^*\right) \xrightarrow{d} N(0,\sigma^2+\omega^2)$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\mathit{eff}}-\boldsymbol{\beta}^*\right) \xrightarrow{d} N(\kappa\rho\sigma,\sigma^2)$$

under Q_n and $\boldsymbol{\beta}^*$ is the true parameter under Q_n. The author defines $\boldsymbol{\beta}^*$ as "the value to which the outcome-model point estimator would converge with complete data as $N \to \infty$".

Note that, in terms of MSE we can now obtain a threshold where the AIPW becomes more efficient than the usual efficient approach. That is, if $\kappa^2\rho^2 > 1$, $\boldsymbol{\beta}_{AIPW}$ is more efficient (smaller MSE) than $\boldsymbol{\beta}_{\mathit{eff}}$ and the worst case possible is when $\rho = 1$.

In order to create sequences of probabilities with $\rho \approx 1$, Lumley (2013) suggests to define a parametric family $\widetilde{Q}_\delta$ by

$$\frac{d\widetilde{Q}_\delta}{dP} = C_\delta \exp\{\delta(\hat{V}-\hat{U})\},$$

so that the correlation is 1 asymptotically and where C_δ is a normalizing constant and $\hat{V}$ and $\hat{U}$ are the influence functions for the efficient and AIPW estimators, respectively.

Lumley discusses a 2-phase study with binary response. Here we work with a continuous response. We also consider a 2-phase design, where the phase-1 data consists of N subjects with known response Y and a covariate x is observed at phase-2. For simplicity, we consider only a single x. As before, we divide Y into 3 mutually exclusive intervals: $I_1 = (-\infty, c_1)$; $I_2 = (c_1, c_2)$; $I_3 = (c_2, \infty)$, where c_1 and c_2 are the 15th and 85th percentiles of Y, and select n_1, n_2 and n_3 units from each interval with probabilities p_1,

p_2 and p_3, respectively. This results in a phase-2 data of $n = n_1 + n_2 + n_3$ fully observed units.

The complete likelihood L is given by

$$\prod_{i=1}^{N} (\pi_i(y_i) f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) g(\mathbf{x}_i))^{R_i} \prod_{i=1}^{N} \left((1-\pi_i(y_i)) \int f(y_i|\mathbf{x};\boldsymbol{\beta}) g(\mathbf{x}) d\mathbf{x} \right)^{1-R_i}, \tag{4.6}$$

where $\pi_i(y_i) = \text{pr}(R_i = 1|y_i;\boldsymbol{\alpha})$ is the probability that the ith unit is selected for full observation and g is the density of $\boldsymbol{X}$.

As discussed in section 1.5, the most efficient augmented inverse-probability weighted estimator is obtained by solving

$$\sum_i \frac{R_i}{\pi_i} U_i(\boldsymbol{\beta}) + \sum_i \left(1 - \frac{R_i}{\pi_i}\right) A_i^*(\boldsymbol{\beta}) = 0$$

where

$$A_i^* = E(U_i(\boldsymbol{\beta})|y).$$

Note that, since the term A_i^* differs from zero, the IPW estimator is no longer the best AIPW estimator. Robins et al. (1994) show this best estimator exists, but implementing it while keeping the distribution of $\boldsymbol{X}$ unspecified is not simple and goes beyond the scope of this section. We will assume, instead, that the distribution of $\boldsymbol{X}$ is known so we can calculate A_i^*, keeping in mind that asymptotically this assumption will cause minor interferences in the results as the best AIPW estimator is asymptotically fully efficient. According to Tsiatis (2006), its influence function is

$$\hat{U}(\boldsymbol{\beta}) = E\left(\boldsymbol{S}_{\boldsymbol{\beta}}\right)^{-1} \left[\frac{R_i \boldsymbol{S}_{\boldsymbol{\beta}}}{\pi_i} + \left(1 - \frac{R_i}{\pi_i}\right) E\left(\boldsymbol{S}_{\boldsymbol{\beta}}|y\right) \right],$$

where $\boldsymbol{S}_{\boldsymbol{\beta}} = \partial \log f(y_i|\mathbf{x}_i'\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$.

For the efficient ML estimator, since the selection probabilities π depend only on Y, by Chen (2004) or Zhang and Rockette (2005), the semiparametric estimator (MLE) derived at Scott and Wild (2006) is fully efficient (see chapter 7), with influence function

$$\hat{V}(\boldsymbol{\beta}) = \left(\boldsymbol{\mathcal{I}}_{\boldsymbol{\beta\beta}} - \boldsymbol{\mathcal{I}}_{\boldsymbol{\beta\alpha}}\boldsymbol{\mathcal{I}}_{\boldsymbol{\alpha\alpha}}^{-1}\boldsymbol{\mathcal{I}}_{\boldsymbol{\alpha\beta}}\right)\boldsymbol{S}_{\boldsymbol{\beta}},$$

where $\boldsymbol{\mathcal{I}}$ stands for the expected information matrix and

$$\boldsymbol{S}_{\boldsymbol{\phi}} = \frac{\partial}{\partial \boldsymbol{\phi}} \log\left(\frac{\pi(y_i;\boldsymbol{\alpha}) f(y_i|\mathbf{x}_i;\boldsymbol{\beta})}{\int \pi(y;\boldsymbol{\alpha}) f(y|\mathbf{x}_i;\boldsymbol{\beta}) dy}\right), \quad \boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix}.$$

The misspecified model was constructed as before. Let $b_\delta = C_\delta e^{\delta(\hat{V}-\hat{U})}$ so that

$$\frac{dQ(y|\mathbf{x})}{dF(y|\mathbf{x})} = b_\delta \quad \rightarrow \quad dQ(y|\mathbf{x}) = dF(y|\mathbf{x}) b_\delta,$$

where Q is the misspecified model and F the correct one. When F is normally distributed with mean $\mathbf{x}\boldsymbol{\beta}$ and variance 1, $dQ(y|\mathbf{x})$ is also normally distributed with the same variance but mean $(\mathbf{x}\boldsymbol{\beta})\, b_\delta$. Moreover, as for the binary case, as $n \to \infty$ the correlation between $dQ(y^*|\mathbf{x})/dF(y^*|\mathbf{x})$ and $\hat{V} - \hat{U}$ tends to 1. In this "worst case scenario", $\rho \approx 1$, we focus our attention on k, the variable that defines the threshold where the AIPW estimator becomes more efficient (in terms of the MSE) than the MLE. Our goal is to see if at this point, where both MSE are close, we have enough power to detect the misspecification.

The simulations consist of basically three steps, which are also shown in figure 4.3: (1) generating an iid population of size N with cdf Q and calculate the influence functions; (2) estimating the mean of the true and misspecified models; (3) estimating the power of detecting departures from the correctly specified model.

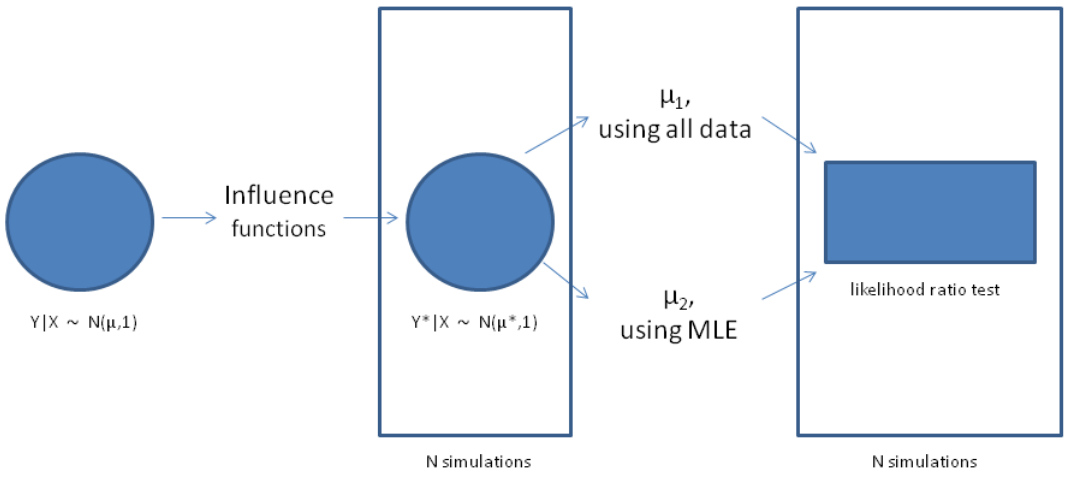


Figure 4.3: Simulation scheme.

Figure 4.4 shows the MSE of the AIPW and the MLE with respect to the power of the Neyman-Person test. The red line corresponds to the best AIPW estimator and the dark line, to the MLE method. Notice that, even when the power for detecting model misspecification is small (30 or 40%), the AIPW is more efficient than the MLE method. When we the power large enough for detecting model misspecification, the MLE method is considerably less efficient than the best AIPW estimator. This raises concern to the use of the MLE method in real problems, where the power of detecting model misspecifications is even smaller than the ones reported here. Unless the model can be considered correctly fitted, the MLE method should be avoided.

4.5 Summary

Here in this chapter we extended the CML+$\widetilde{\boldsymbol{S}}$ method for the continuous case and analyzed its performance through simulations. Unlike other methods discussed (see section 4.1), the CML+$\widetilde{\boldsymbol{S}}$ does not require any parametric model for extra variables fully observed at phase-1 and not used for selecting the phase-2 sample. The method can also be applied to situations in which fitting a saturated model for the selection

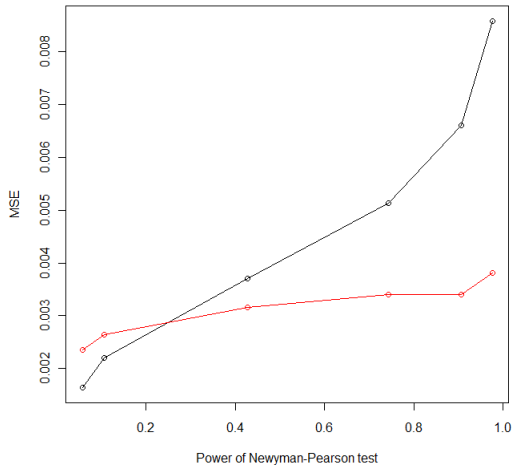


Figure 4.4: MSE of MLE (dark line) and MSE of AIPW (red line).

probability is not possible and makes use of the whole data in a fairly simple way.

Regarding the simulated studies, we note that if all models are correctly specified and the variable that goes missing has little effect than MI and calibration estimate the effects of the remaining variables more efficiently. If it has a reasonable large effect, CML+$\widetilde{\boldsymbol{S}}$ is most efficient for all variables and in particular for β_2. As in the binary case, the CML+$\widetilde{\boldsymbol{S}}$ method makes use of additional variables (denoted in our simulations by a binary coarsening) that was not used in any part of the study in a simple and efficient way. However, as discussed in previous section, small departures from the true model may lead to large bias and, consequently, large mean squared errors. Therefore, the CML+$\widetilde{\boldsymbol{S}}$ method should be used with extreme care in in real problems, otherwise a more robust approach as the weighted or calibration methods are a better choice.

5

Treatment Effects and Propensity Scores

There are strong connections between using estimated weights and many applications of the propensity score concept. Both rely on estimating the probability of a response, often binary, given known covariates and can be considered equivalent under some conditions.

In this chapter we will describe how the propensity score is used to estimate treatment effects and how it is related to the estimated weights approach. We start by reviewing the propensity score theory, discussing its application in the literature, mainly though simulated studies. Later we run a series of simulations with estimators based on both techniques, propensity scores and estimating weights, and see how they behave under a variety of scenarios.

5.1 Treatment effect

The term "treatment effect" comes from the medical literature where interest lies in assessing the efficacy of a new treatment or procedure (Park et al., 2011). This term was later extended to parallel applications in other areas, for example, in economics to assess whether or not using a new training technique will result in an increase of sales. Here training would be the treatment variable and the treatment effect would be given by the comparison between the sales of the treated and untreated groups (Ashenfelter, 1978).

We will generally be discussing a binary treatment (1 for treated and 0 for untreated), but everything that follows can be extended to treatment variables with more than 2 levels. The only requirement here is that different groups receive different treatments. Interest lies in measuring the effect of treatment on the outcome of interest. Letting y_{1i} be the potential outcome of the ith subject if it was treated and y_{0i} be the potential outcome of the same ith subject if it was untreated, the treatment effect TE_i is defined as

$$TE_i = y_{1i} - y_{0i}.$$

Let $T_i = 1$ if the ith individual is treated, 0 otherwise. Normally we observe only one of these outcomes. In other works, we observe $y_i = y_{1i}T_i + y_{0i}(1 - T_i)$. It is impossible, therefore, to obtain the real treatment effect at the individual level. Holland (1986) called this the *Fundamental Problem of Causal Inference*. To address this problem, the author gives two alternatives: the scientific solution and the statistical solution.

In the *scientific solution* the scientist exploits invariance or homogeneity assumptions to overcome the Fundamental Problem of Causal Inference. For example, the researcher may believe that the value of y_{0i} does not vary over time and the value of

y_{1i} is not affected by prior exposure. We can then measure the response y_{0i} for the ith individual before giving him/her the treatment of interest and observing y_{1i} afterwards. Here the invariance assumption was used. The homogeneity assumption is used when y_{1i} is believed to be equal to y_{1j} and $y_{0i} = y_{0j}$, for two units i and j. The treatment effect can be observed as $y_{1i} - y_{0j}$.

For the *statistical solution*, we calculate the average treatment effect ATE by

$$ATE = \mathbb{E}(Y_1) - \mathbb{E}(Y_0). \tag{5.1}$$

These quantities, however, are not observed. The *observed treatment* effect OTE is

$$OTE = \mathbb{E}(Y_1|T = 1) - \mathbb{E}(Y_0|T = 0), \tag{5.2}$$

which takes into consideration only the values that were actually observed. Note that for (5.2) only part of the population is used, while for (5.1) the entire population is taken into consideration. Thus, (5.2) is not necessarily equal to (5.1). They become equal when the individuals are randomly selected into the treated or untreated group. If this assumption holds, $\mathbb{E}(Y_0) = \mathbb{E}(Y_0|T = 0)$ and $\mathbb{E}(Y_1) = \mathbb{E}(Y_1|T = 1)$, so that OTE = ATE.

For randomized experiments, where the study population is randomly assigned to the treatment group, background covariates (measured or unmeasured) will be similar for participants in both groups in such a way that the only systematic difference between them will be the treatment status. We can then establish a causal relationship between the treatment T and response the Y.

On the other hand, if independence between treatment assignment and background covariates does not hold, the covariates will no longer be evenly distributed between the

groups and so the treatment group may be systematically different to the control group in ways other than treatment. This occurs in non-randomized observational studies. Here, if we use equation (5.1) to estimate the average treatment effect, we will usually get biased results because we cannot be sure that T is the only variable affecting Y.

Although randomized experiments are the ideal form of study, they cannot be performed in every situation. Suppose, for example, that we want to study the effect of smoking on developing lung cancer. For ethical reasons, we cannot randomly assign people to the smoking or non-smoking groups and so a randomized experiment cannot be performed. As a result we cannot claim a causal relationship because unobserved covariates such as genetics that are not balanced between the two groups may be related to lung cancer in some way and so direct comparisons in observational studies between treatment and control groups may lead to biased results.

5.2 Propensity score

The propensity score (e) method was developed by Rosenbaum and Rubin (1983) as a way to reduce bias in observational studies in attempting to estimate the average treatment effect ATE. The propensity score is defined as the probability of being treated given a set of observed covariates x_i,

$$e(\mathbf{x}_i) = \text{pr}(T_i = 1 | \boldsymbol{X}_i = \mathbf{x}_i). \tag{5.3}$$

It is usually estimated by a logistic regression model but some other methods have been proposed and will be discussed later in this chapter.

The *strong ignorability assumption* of T given $\boldsymbol{X}$ plays an important role in propensity score theory and is defined by the following two properties:

- $Y(1), Y(0) \perp T \mid \boldsymbol{X}$, that is, potential outcomes under treatment $Y(1)$ and under no treatment $Y(0)$ are independent of the treatment assignment T given $\boldsymbol{X}$. This assumption is also known as *unconfoundedness*, since it assumes that there are no unobserved covariates associated with the potential outcomes and the treatment besides those already measured.

- $0 < \text{pr}(T = 1|\mathbf{x}) < 1$, for all $\mathbf{x}$, which is known as the overlap assumption. It assumes that the conditional distribution of $\boldsymbol{X}|T = 0$ overlaps with the conditional distribution of $\boldsymbol{X}|T = 1$, so that for each treated subject with specific background covariates, an untreated subject with similar characteristics should have also been observed.

We say that treatment is strongly ignorable if both assumptions hold.

Propensity scores are widely used because of the following properties derived in the seminal paper by Rosenbaum and Rubin (1983). They show that the propensity score is the simplest, coarsest function of $\boldsymbol{X}$ that is a balancing score, where a balancing score $b(\mathbf{x})$ is a function of the observed covariates such that $\boldsymbol{X} \perp T|b(\mathbf{x})$. Moreover, it is shown that estimates of a balancing score also behave as a balancing score and if treatment assignment is strongly ignorable given $\boldsymbol{X}$, it is also strongly ignorable given any balancing score. In other words, the use of propensity scores creates a quasi-randomized experiment in the sense that exact matching on $e(\mathbf{x}_i)$ will tend to balance the $\boldsymbol{X}$ distributions in the treated and control groups (Rosenbaum and Rubin, 1985). Propensity scores, however, will only balance the observed covariates. As Rubin stated, "it is important to keep in mind that even propensity score methods can only adjust for observed confounding covariates and not for unobserved ones" (Rubin, 1997). In addition, $e(\mathbf{x}_i)$ is rarely ever known and therefore must be estimated from the available data. This estimated propensity score may not be able to remove all bias if the

covariance adjustment method (see below) is used (Hade and Lu, 2013).

Propensity scores can also be used to improve efficiency, as noticed by Hahn (1998) and Heckman et al. (1999). Rosenbaum (1987) and Rubin and Thomas (1996) show that more efficient estimates can be obtained by using estimated propensity scores instead of the true values and Hirano et al. (2003), based on Robins et al. (1995), show that by using propensity score as weights the semiparametric efficiency bound can be achieved.

Theorem 4 of Rosenbaum and Rubin (1983) shows that a balancing score can be used to obtain unbiased estimates of the average treatment effect and corollaries 4.1-4.3 in particular show that balancing scores can be used in matching, stratification and covariance adjustment in order to provide estimates which are unbiased under the assumption of T being strongly ignorable given $\boldsymbol{X}$. The only exception is for the covariance adjustment method, where it is required that the conditional expectation of T given the balancing score $b(\mathbf{x})$ should be linear in addition to the strong ignorability assumption of T given $\boldsymbol{X}$. We assume throughout that the strong ignorability assumption is met and, additionally the linearity assumption is met when discussing the covariance adjustment method.

Matching

In the context of treatment effects, a matched study consists of matching subjects that received the treatment to a set of untreated subjects, where the matching step is based on the observed covariates. That is, subjects with equal or nearly equal observed covariates are matched together and so the treatment effect can be correctly estimated under the assumption of unconfoundness. A clear drawback occurs when there are a large number of covariates so that finding individuals in the treated and untreated

groups with similar background variables becomes very difficult.

Many matching techniques have been developed (Stuart, 2010) and a common choice is to use the Mahalanobis metric to match treated and untreated subjects (Baser, 2006), as follows. First, all treated subjects are randomly sorted and the Mahalanobis distance between the first treated unit and all controls are calculated. Then, the one with the smallest distance is considered a match for that first treated subject and we move to the second treated individual, and so on, until all treated subjects are matched.

Propensity scores can be used for matching in a much simpler way. It can be seen as a summary of many observed covariates represented in only one number and based on the fact that people with similar propensity score will tend to have similar background variables as discussed earlier. We can use the propensity score for matching instead of using the observed covariates. Thus, a k-dimensional problem, where k is the dimension of the covariates, becomes a 1-dimensional problem.

Rosenbaum and Rubin (1983) also suggest three ways of using propensity scores for matching: nearest available matching on the estimated propensity score, Mahalanobis-metric matching including the propensity score and nearest available Mahalanobis-metric matching within calipers defined by the propensity score.

One drawback of matching is that we may discard a large number of untreated individuals because not all of them will be matched to the treated ones and that imperfect matching may lead to some residual bias (Hill, 2008). Despite this, matching with propensity scores is still a popular method and is commonly applied in medical statistics (Waernbaum, 2011).

Stratification

Stratification (or subclassification) is also commonly used in estimating treatment effects in observational studies (e.g. Rosenbaum and Rubin (1984); Rosenbaum (1991); Hansen (2004)). Here, individuals are divided into strata based on their observed baseline covariates so that treated and control individuals can be compared for each stratum. A large number of covariates can make stratification infeasible since it may result in a large number of strata with some containing just a few, or even no individuals at all or having subjects from only one group making comparison between treated and untreated individuals impossible.

Propensity scores can be used as a way to reduce the number of covariates used to create the strata, as follows. First, the propensity score e_i for each individual is estimated and the vector $e = (e_1, \ldots, e_N)$, where N is the size of the population, is divided into k intervals. The average treatment effect is then estimated within each stratum and the overall treatment effect is estimated as a weighted mean of the previous estimates.

Weighting and covariance adjustment

Propensity scores can also be used as weights to obtain unbiased estimates for the average treatment effect. Under the strong ignorability assumption,

$$E\left\{\frac{YT}{e(\boldsymbol{X})}\right\} = E\left\{E\left[\frac{YT}{e(\boldsymbol{X})}\middle| y, \mathbf{x}\right]\right\} = E\left\{\frac{Y}{e(\boldsymbol{X})}E\left[T|y,\mathbf{x}\right]\right\} = E\left(Y_1\right).$$

and the ATE can be estimated by weighting observations in each group by the inverse of its propensity score (Rosenbaum, 1998). Notice that this estimator is an IPW estimator and falls into the broader class of AIPW discussed in Robins et al. (1994) (see section

1.5).

Propensity scores can also be used in regression as an adjustment variable (covariance adjustment). Thus, propensity scores can be seen as data reduction methods, where a two-step procedure is performed, as discussed in D'Agostino (1998): The first step consists of fitting a very complicated propensity score model and adjust for the propensity score in the model of interest. The covariance adjustment is generally more efficient than simply using propensity scores as weights, but can perform poorly if the sample linear discriminant based on covariates is not a monotone function of the propensity score (Rosenbaum and Rubin, 1983). Moreover, for the linear model

$$E\left(y_i|T_i, x_1, \ldots, x_k\right) = \beta_0 + \beta_1 T_i + f(x_1, \ldots, x_k)\beta_2,$$

where k is the number of covariates and $f(.)$ is how the covariates affect the response (which may not be linear), Hade and Lu (2013) show that replacing $f(x_1,\ldots,x_k)$ by the estimated propensity score $e(x_1,\ldots,x_k)$ may produce a biased treatment-effect estimator. If the true propensity score is used instead, the treatment effect estimator is unbiased.

5.2.1 Estimating the propensity score

Propensity scores are usually estimated by fitting a parametric (usually logistic) model for the treatment assignment given the covariates, i.e., estimating $\text{pr}(T = 1|\mathbf{x})$ by fitting a model $\text{pr}(T = 1|\mathbf{x};\boldsymbol{\eta})$. However, slightly misspecifications may lead to biased estimates of the estimated treatment effect (Kang and Schafer, 2007) and might also result in poor covariate balance. Achieving covariate balance is, in fact, the main criterion to decide whether an estimated propensity is or is not appropriate (Imai et al., 2008).

The balance of the covariates can be assessed in many ways: A simple two sample t-test for the mean difference of each of the covariates is among the most common methods (Imai et al., 2008); the cross-match test (Rosenbaum, 2005); or a permutation-type test (Hansen and Bowers, 2008) are also alternatives for testing if the background covariates are balanced. Rosenbaum and Rubin (1985) give a standardized bias expression to measure the overall covariate imbalance and Rosenbaum and Rubin (1984) and Rubin (1997) suggest a cyclic process of checking for balance and reformulating the propensity score model.

In order to avoid problems with model misspecification, non-parametric models for estimating the propensity score have been studied (McCaffrey et al., 2004), but as the dimension of $\boldsymbol{X}$ increases, the non-parametric approach becomes more challenging. Some work has then been done on double robustness methods (Tan, 2010) or on achieving the balancing property. Imai and Ratkovic (2014) give a good overview of several methods that estimate the propensity score while keeping the balancing property and also propose a new method called the covariate-balancing propensity score (CBPS). Their method is based on the fact that by using the inverse propensity score as weighting we should ideally achieve the balancing property. That is, after weighting, all groups should ideally have similar means

$$E\left\{\frac{T_i\tilde{\boldsymbol{X}}_i}{e(\boldsymbol{X}_i)}-\frac{(1-T_i)\tilde{\boldsymbol{X}}_i}{1-e(\boldsymbol{X}_i)}\right\}=0,$$

where $\tilde{\boldsymbol{X}}_i=f(\boldsymbol{X}_i)$, for any function f specified by the researcher. The authors also add another constraint regarding the equality of second moments and use the generalized method of moments (Hayashi, 2000) or the empirical likelihood approach (Owen, 2001) to obtain the estimated propensity scores. They run simulations in order to examine how different methods of estimating the propensity score affect commonly weighting

estimators (such as the IPW method described earlier). According to their simulations, if the propensity score model is correctly specified, all methods used to estimate the propensity, in particular the CBPS method and the simple logistic regression, produce nearly the same results. Thus, since for everything that follows we assume that the true propensity score model is a submodel of the fitted model, we can use a simple logistic regression to estimate $e(\mathbf{x})$.

5.3 Covariance adjustment

Because of its nice properties and ease of implementation, methods based on propensity scores have become very popular among clinical researchers, with an exponential increase in publications reporting or exploiting propensity scores for multivariate adjustment, reaching nearly 2,000 publications in the period 2005-2009 (Biondi-Zoccai et al., 2011). But as its use increases, concern with its misapplication is also increasing.

Pattanayak et al. (2011) call attention to the dangers of using regression adjustment in observational studies. According to the authors, the most important flaw of regression adjustment is that study design is not separated from outcome analysis. In fact, the design phase of randomized experiments is set prior to seeing any outcome data and since "an observational study should be conceptualized as a broken randomized experiment" (Rubin, 2007), the design phase of an observational study should often be done before seeing the data. More formally, the design of an observational study should follow three steps, known as the Rubin Causal Model (Holland, 1986): the first step consists of defining causal effects as a comparison between the potential outcomes of a treatment and a control on a common set of units; the second step consists of defining an assignment mechanism using a probabilistic model for T given $(\boldsymbol{X}, Y(0), Y(1))$; and the third step, which is optional, consists of specifying a model for $\text{pr}(\boldsymbol{X}, Y(0), Y(1))$.

The propensity score is used at step 2 with the purpose of reconstructing a randomized experiment and so, with no outcome data available at this stage, matching, weighting or stratification on propensity scores are natural ways of eliminating bias due to observed covariates.

Pattanayak et al. (2011) also show concern regarding the covariance adjustment method when there is hardly any overlap between the distributions of covariates in the treated and untreated groups. Note that in this case the strong ignorability assumption of T given $\boldsymbol{X}$ does not hold (see section 5.2).

Based on the fact that all mean bias is along the propensity score, Rubin (2001) discusses three basic distributional conditions that should be used simultaneously for regression adjustment to be trustworthy:

(i) Unless the distributions are nearly symmetric, have nearly the same variance and nearly the same sample sizes, the difference in the means between the propensity scores of each group must be less than half of a standard deviation;

(ii) The ratio between the variance of the propensity score in the two groups must be close to one;

(iii) After adjusting for the propensity score, the ratio of the variances of the residuals of the covariates must also be close to one.

Notice that these conditions implicitly assume that the covariates are normally distributed or following distributions that can be summarized by means and variances. These conditions, however, are usually ignored in many applied fields and if at least one condition does not hold, the difference between the distributions of the covariates in the two groups must be considered as substantial and the regression adjustment should be considered unreliable (Rubin, 2001). Moreover, covariance adjustment is a paramet-

ric approach so that unless the conditional distribution of Y given the covariates can be correctly modelled, non-parametric methods such as propensity scores matching, stratification or weighting should be used.

5.3.1 Covariance adjustment in linear regression

Freedman and Berk (2008) also call attention to the fact that many researchers have used propensity scores without fully understanding the costs of misapplying this technique. For a treatment T and confounders $\boldsymbol{X} = (X_1, X_2)$ correlated with T, the authors use a linear model

$$Y = \beta_0 + \beta_T T + \beta_2 x_1 + \beta_3 x_2 + \epsilon \tag{5.4}$$

or a Bernoulli response with success probability defined by setting

$$\text{logit}(\text{pr}(Y = 1|T, \mathbf{x}; \boldsymbol{\beta})) = \beta_0 + \beta_1 T + \beta_2 x_1 + \beta_3 x_2$$

to compare the performance, with respect to the mean squared error (MSE), of the unweighted and weighted regressions. They ran regressions of Y on T and X, with no adjustment and a weighting adjustment, using the inverse propensity scores as weights. Thus, since the true propensity score is $e(\mathbf{x}) = \text{pr}(T = 1|x_1, x_2)$, subjects with $T = 1$ receive weight $1/\hat{e}$ and subjects with $T = 0$ receive weight $1/(1-\hat{e})$. The weighted regression minimizes the weighted sum of squares. The authors assume that the propensity score model is correctly specified and run both weighted and unweighted regressions of Y on T and (X_1, X_2); Y on T and X_1; Y on T alone, in order to see the trade-off between variance and bias incurred when making the weighting adjustment.

Their simulations show that when the model of interest is correctly specified, there

is no bias to reduce and unnecessary weighting by propensity scores increases the MSE error as might be expected. However, it may provide more efficient estimates when covariates are not included in the model of interest, but are included in the propensity score model. Still regarding the weighted regression, their simulations show some small-sample biases and some substantially underestimated standard errors. Worryingly, these were about 3 times smaller than the empirical standard errors (Emp.SE) obtained (through 250 simulations). It is of interest, then, to see if the covariance adjustment method, not considered by Freedman and Berk, also underestimates the standard error while estimating the treatment effect.

To this end, we simulated the same scenario as in Freedman and Berk (2008). That is, consider the linear continuous response given by equation (5.4) with coefficients $\boldsymbol{\beta} = (1,1,1,2)^T$ and let

$$T = \begin{cases} 1, & \text{if } \eta_0 + \eta_1 x_1 + \eta_2 x_1 + \nu > 0 \\ 0, & \text{otherwise} \end{cases}$$

with $\boldsymbol{\eta} = (.5,.25,.75)^T$. In addition let $\epsilon \sim N(0,1)$ and $\nu \sim N(0,1)$ and $\boldsymbol{X} = (X_1, X_2)$ be bivariate normal, with $E(X_1) = .5$, $E(X_2) = 1$, $\text{Var}(X_1) = 2$, $\text{Var}(X_2) = 1$ and correlation $\rho = .5$. We estimated the propensity score using a probit model, correctly specified, and ran the same regressions as before, but including the estimated propensity score as an extra covariate.

The results are presented in table 5.1. Regression adjustment for the propensity score removes bias in β_T, which is expected since we are fitting the correct model for $e(x)$. With respect to the standard errors, the situation is reversed. The nominal standard errors of β_T are somewhat overestimated by the covariance adjustment method by a factor of about 1.3 (i.e., the estimated standard errors (Est.SE) were conservative).

Table 5.1: Results for the covariance adjustment, for 500 datasets simulated.

	Bias				Emp.SE/Est.SE			
	β_0	$\boldsymbol{\beta_T}$	β_2	β_3	β_0	$\boldsymbol{\beta_T}$	β_2	β_3
Y on (T, X_1, X_2)	-0.030	**0.002**	-0.004	0.002	1.047	**1.007**	1.037	0.975
Y on (T, X_1)	-4.409	**0.008**	0.174		2.257	**0.781**	1.653	
Y on (T)	-8.984	**0.012**			3.426	**0.724**		

Notice that in the simulation above as well as in both scenarios studied by Freedman and Berk (2008), each variable was either fully observed or not observed at all; no partially observed variables were considered. Our goal here is to fill this gap, by allowing some variables to be missing by happenstance so that only a sample of the population will have been fully observed. We ran a small simulation to understand how this partial information affects the results previously obtained.

Consider the following scenario. Let $\text{logit}(\pi(\mathbf{x})) = 1 + x_1$ be the logit of the probability of providing information regarding X_2 while the remaining variables are known for all subjects in the study. We work with a similar model to (5.4), but with the extra variable $Z = X_1^2 + \epsilon_z$ added into the model, i.e.,

$$Y = \beta_0 + \beta_1 T + \beta_2 x_1 + \beta_3 x_2 + \beta_4 z + \epsilon$$

where $\epsilon_z \sim N(0,1)$. The reason for adding Z is that a missed quadratic term is perhaps more likely than an extra unmeasured covariate.

For the remaining variables, let the treatment indicator variable T take the value 1 with probability

$$e = \text{pr}(T = 1|\mathbf{x};\boldsymbol{\eta}) = \frac{\exp(\mathbf{x}\boldsymbol{\eta})}{1+\exp(\mathbf{x}\boldsymbol{\eta})}$$

and let $\boldsymbol{X} = (X_1, X_2)$ be multivariate normal centred around 0, with variances 1 and correlation $\rho = .5$. Following Freedman and Berk (2008), we ran regressions of Y on

$(T, \boldsymbol{X}, Z)$, Y on $(T, \boldsymbol{X})$ and Y on (T), comparing three methods: unweighted least squares with (Adj) and without (Lin) adding the propensity score as an extra covariate and the propensity score weighted adjustment (wgt). The regression models are:

- Model 1: $Y = \beta_0 + \beta_T T + \beta_2 x_1 + \beta_3 x_2 + \beta_4 z$;
- Model 2: $Y = \beta_0 + \beta_T T + \beta_2 x_1 + \beta_3 x_2$;
- Model 3: $Y = \beta_0 + \beta_T T + \beta_2 x_1$;
- Model 4: $Y = \beta_0 + \beta_T T$,

with $\boldsymbol{\eta} = (1, -2, -2)^T$ and $\boldsymbol{\beta} = (5, 5, 1, 1, .3)^T$. Model 1 is the true model that generates the data and so the remaining models are also used to fit the data. We estimated the propensity score using a logistic model correctly specified, expecting that it would correct the omitted-variable bias of the three last misspecified models. We ran 500 simulations and the results for bias, empirical (Emp.SE) and estimated (Est.SE) standard errors and coverage are shown in Table 5.2.

Our interest is in the treatment effect β_T. If the complete model is fitted, there is no bias to be adjusted, and the simple linear regression, which makes no use of propensity scores, has the lowest variance. In the other three scenarios, a similar pattern is observed to that obtained by Freedman and Berk. Methods based on propensity scores result in much less biased results than the simple linear regression, resulting in substantially smaller MSEs. The weighted regression again has a small-sample bias while estimating β_T and the nominal standard error is again substantially underestimated. Covariance adjustment leads to smaller standard errors while keeping the estimates nearly unbiased in all scenarios. For the last two cases where (X_2, Z) and (X_1, X_2, Z), respectively, are not included in the model of interest, the nominal standard error is

Table 5.2: Results for β_T only, for 500 datasets simulated

Regressions	Method	Bias	Emp.SE	Est.SE	MSE	% Coverage
	wgt	-0.006	0.352	0.165	0.352	0.645
Y on (T, X_1, X_2, Z)	Adj	-0.007	0.252	0.255	0.252	0.947
	Lin	-0.006	0.229	0.232	0.229	0.959
	wgt	0.002	0.375	0.179	0.375	0.654
Y on (T, X_1, X_2)	Adj	-0.002	0.267	0.284	0.267	0.966
	Lin	0.079	0.258	0.256	0.264	0.940
	wgt	-0.257	0.445	0.222	0.511	0.575
Y on (T, X_1)	Adj	-0.018	0.274	0.320	0.274	0.974
	Lin	-0.939	0.273	0.276	1.155	0.075
	wgt	-0.698	0.714	0.312	1.201	0.305
Y on T	Adj	-0.004	0.280	0.358	0.280	0.988
	Lin	-2.614	0.282	0.293	7.113	0.000

slightly overestimated, but this error is not nearly as large as in the weighting adjustment.

Hade and Lu (2013) discuss using the estimated propensity score as a regression covariate to remove bias in observational studies and express strong concern with its use in specific situations regarding the overlap between the distributions of $\boldsymbol{X}|T=1$ and $\boldsymbol{X}|T=0$. Notice, however, that this overlap must be controlled in order to satisfy the strong ignorability assumption of T given $\boldsymbol{X}$. They worked with different response functions and different overlaps through simulations. The method recommended by these authors depends on the type of overlap between the covariates. Their simulation study works as follows: they first generate the treatment and control groups and assign $\boldsymbol{X}_t$ and $\boldsymbol{X}_c$ for each group, respectively. For the linear model and a univariate X, they generate data from

$$Y = \beta_0 + \beta_T T + \beta_2 f(x) + \epsilon$$

where $f(x)$ can be linear, quadratic, exponential or a step function, ϵ follows a standard

normal distribution and $X = (X_t T, X_c(1-T))$, where X_t and X_c are two independent normals with means (μ_t, μ_c) varying from 20-30 and variances (σ_t^2, σ_c^2) varying from 3-5. They used $\boldsymbol{\beta} = (1, 10, 1)^T$ and used a logistic regression of T on X to estimate the propensity score. The model $Y \sim T$ is fitted to the data so that the unadjusted linear regression will have omitted-variable bias and methods that use propensity scores are expected to perform better. The goal here is to compare all methods with respect to the relative bias of the estimated treatment effect.

Their simulations show that biases are related to the distributional overlap of the covariates and also to the relationship between Y and $f(X)$. When $f(X) = X$, adding the propensity score as covariate is enough to remove nearly all bias, but as the relationship between $f(X)$ and Y departs from linearity, the relative bias increases. When the distribution of X_t is contained in the distribution of the control group and $f(X)$ is quadratic or defined as a step function, the relative bias is approximately 500% and 100%, respectively. The model misspecifications they were using, however, were so strong that they would be detected essentially 100% of the time, at the 5% level, with a standard regression. In fact, as used in Hade and Lu, with 400 (or 2,000) individuals allocated into the treatment group and 100 (or 500) allocated into the control group, a quick analysis is enough to strongly detect all model misspecifications used in their paper. As X is used to estimate the propensity score, we ran a regression of Y on (T, X) and the residual plots for N = 500 are shown in figure 5.1, for $f(X)$ quadratic,

exponential and equals to the step function

$$f(X) = \begin{cases} (X-25)^2+1 & \text{if } X \in (\infty, 25] \\ 14(X-25)+1 & \text{if } X \in (25, 26) \\ -5(X-26)+15 & \text{if } X \in [26, 27) \\ 10 & \text{if } X \in [27, 29) \\ -5(X-29)+10 & \text{if } X \in [29, 30) \\ 15(X-30)+5 & \text{if } X \in [30, 31) \\ 20 & \text{otherwise.} \end{cases}$$

To see whether these effects were still evident with moderate misspecification, we weakened the relationship between Y and $f(X)$ by decreasing the coefficient β_2 so that the residual plot (Fig. 5.2) does not show such severe model misspecification and thus better reflecting reality. Also, the same argument used by Freedman and Berk should be applied in this setting: if X is used to estimate the propensity score, X is also likely to be included in the model of interest. Then, for linear $f(X)$, all methods including the unadjusted linear regression, are able to remove all bias. Table 5.3 shows the result for quadratic $f(X)$ and for the same scenario discussed above: 400 subjects allocated into the control group, 100 allocated into the treatment group and (X_c, X_t) are both normally distributed with mean $(20, 25)$, variance $(5, 3)$ and correlation $\rho = 0$. Data were generated from the true model

$$Y \sim \beta_0 + \beta_T T + \beta_2 x_1^2$$

and the propensity score was estimated by a logistic regression. The estimated and true propensity scores are shown in figure 5.3.

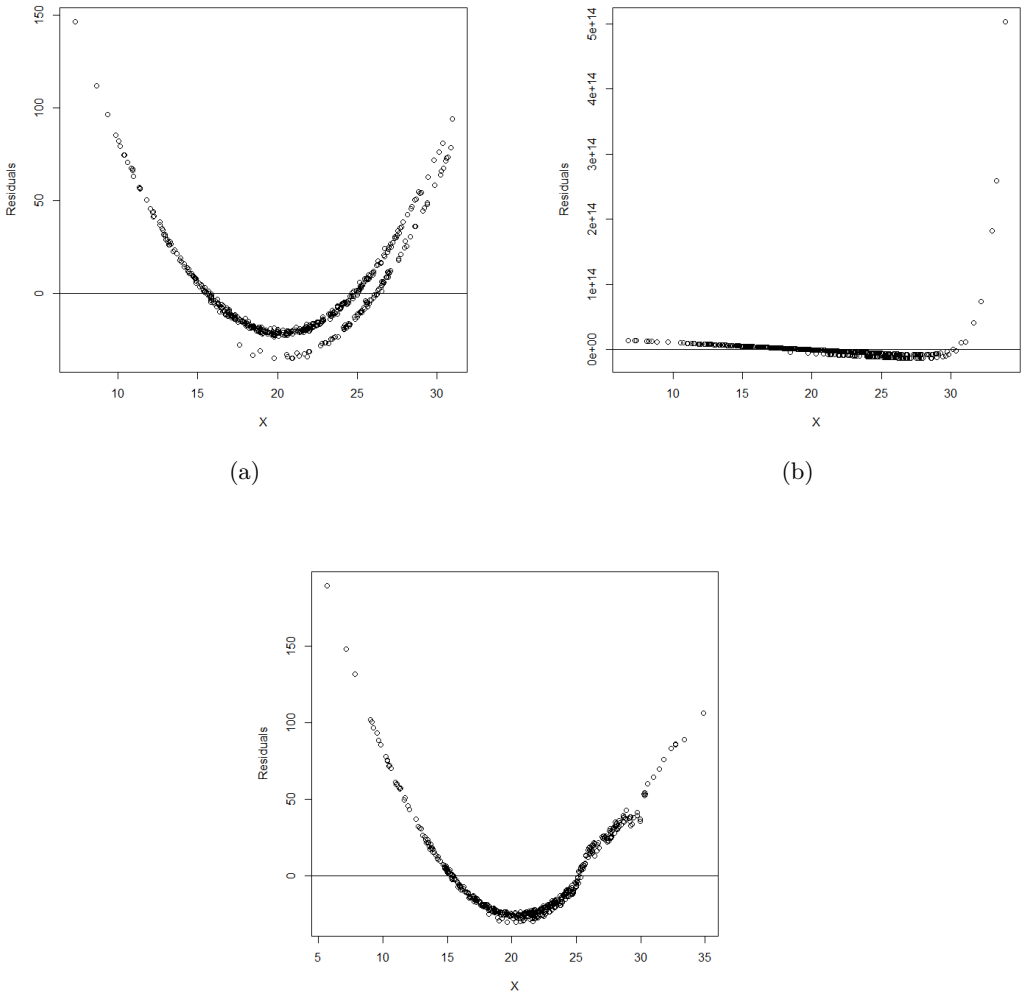


Figure 5.1: Residual plot for $f(X)$ equals to a (a) quadratic, (b) exponential and (c) step function.

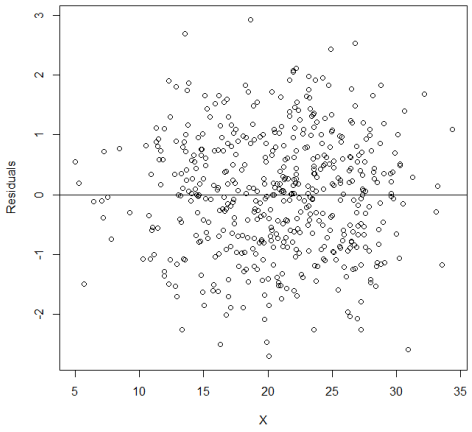


Figure 5.2: Residual plot for quadratic $f(X)$ and $\beta_2 = .004$.

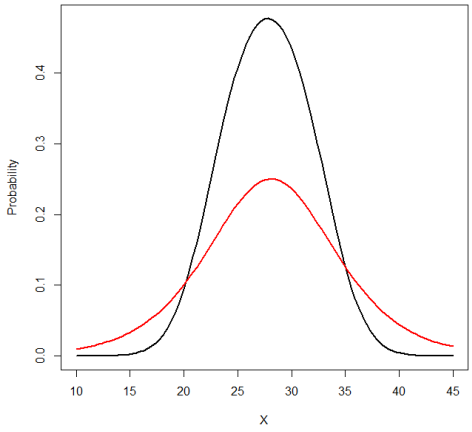


Figure 5.3: True (black line) and estimated (red line) propensity scores.

We fitted the model

$$Y \sim \beta_0 + \beta_T T + \beta_2 x_1 \tag{5.5}$$

to obtain the results in Table 5.3 and ran 1,000 simulations with $\boldsymbol{\beta} = (1, 10, .004)^T$. Note that (5.5) is misspecified and the quadratic term was detectable in about 80% of the time. The residual plot is shown in Fig. 5.2.

Even with a weak relationship between Y and $f(X)$, the unadjusted linear regression shows a bias that is about 20 times bigger than that obtained by the propensity score covariate adjustment, but nearly the same MSE. That is, adding the estimated propensity score into the regression model does not help estimating the treatment effect. Moreover, as in previous simulations, the propensity score weighted adjustment underestimates the standard error while for the covariance adjustment method the ratio between the empirical and nominal standard error was close 1. Most of the bias was removed by the covariance adjustment method and the coverage was close to the nominal value. Clearly, omitting X_1 from the fitted model (5.5) will increase the bias in all cases, especially for the unadjusted linear regression, but the covariance adjustment

Table 5.3: Results for the covariance adjustment for quadratic $f(X)$, for 1000 datasets simulated.

Regressions	Method	Bias			Emp.SE/Est.SE		
		β_0	β_T	β_2	β_0	β_T	β_2
Y on (T, X)	wgt	-1.792	**-0.074**	0.171	1.412	**1.519**	1.531
	adj	-0.636	**-0.001**	0.095	0.993	**0.964**	1.017
	Linear	-1.561	**0.023**	0.159	1.052	**0.972**	1.041
		MSE			% Coverage		
		β_0	β_T	β_2	β_0	β_T	β_2
Y on (T, X)	wgt	4.623	**1.524**	1.560	0.001	**0.730**	0.000
	adj	1.397	**0.964**	1.026	0.607	**0.958**	0.030
	Linear	3.489	**0.973**	1.066	0.000	**0.952**	0.000

method continues as the best estimator.

As in Hade and Lu (2013), we also ran simulations when the covariates (X_t, X_c) are normally distributed with mean $(30, 20)$ and variances $(3, 5)$ so that there is some, but not too much overlap between the two distributions. In this scenario, however, the ratio between the propensity scores of the treated and untreated groups is around 3.1 which goes against the suggestions of Rubin (2001), discussed in the beginning of this section. Moreover, there is a very weak overlap between the estimated propensity scores of the two groups, with values being close 1 (maximum value equals to 0.9999988) and 0 (minimum value equals to 0.0000000), and thus failing to satisfy the second property of propensity scores (as discussed in section 5.2). Following the Rubin (2001) suggestions, regression adjustment should not be trusted and so these results are not reported here.

5.4 Multiphase studies

Propensity scores and the estimated weights method are closely related, especially for the propensity score weighting adjustment method. In both cases each unit is

weighted by a function in order to create balance between the sampled (treated) and not sampled (untreated) groups. In both cases, the weights are usually estimated by a logistic regression and the resulting estimating equations and asymptotic variance have the same structure. Also, if they are correctly applied and all models are correctly specified, both methods lead to unbiased estimation.

However, they do differ in their definition. While estimated weights were originally mainly used to estimate population totals (Horvitz and Thompson, 1952), the propensity score was created as a way to approximate or replicate a randomized experiment when designing an observational study. And since in randomized experiments the outcome data is not available at the design phase, propensity scores are a function of the observed covariates only and not of the outcome data. The method was created to match individuals on a set of particular covariates in order to remove confounder-caused bias (under the unconfoundedness assumption) and not intended to increase precision (Rubin, 2007).

Throughout this thesis we have been working with a multi-phase sampling scheme where selection into the next phase was a function of the response and some covariates of interest. Propensity scores cannot be used alone in this setting because it does not take biased sampling into account and must be used along with a way of catering for biased sampling such as the estimated weights method in order to provide unbiased estimates. In this section we discuss methods that combine both approaches while estimating treatment effects in an outcome-dependent sampling scheme and using propensity scores as protection against model misspecification. We will discuss situations that are more likely to be found in real problems and emphasize likelihood based methods to produce efficient estimates.

5.4.1 Estimating equations

Apart from the methods discussed in previous chapters, we consider the following methods in our simulations:

(i) *Replacing the weights.* The first method we considered replaces the weights $1/\pi_1$ by $e(\mathbf{x}_i)/\pi(\mathbf{x}_i, y_i)$, where $\pi(\mathbf{x}_i, y_i) = \text{pr}(R_i = 1|\mathbf{x}_i, y_i)$. Notice that since e is a function of X only, when applied to the CML method, these new weights do not make an impact since the numerator cancels out and the weights becomes $1/\pi_1$, as before.

(ii) *Weighted CML.* Propensity scores can also be used to produce a weighted version of the CML method. That is, we maximize the following pseudo-loglikelihood

$$\ell(\mathbf{x}_i, e(\mathbf{x}_i); \boldsymbol{\phi}) = \sum_{i=1}^{N} \frac{1}{e(\mathbf{x}_i)} \log f_c(y_i|\mathbf{x}_i, T_i).$$

where $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\eta}^T)^T$.

(iii) *Propensity score covariate adjustment.* Finally, we also considered adjusting for the propensity score for both weighted and CML methods, where the fitted model $\mu(\mathbf{x}; \boldsymbol{\beta})$ is now replaced by $\mu = \mathbf{x}\beta + e\boldsymbol{\beta}_{ps} + \epsilon$.

In all cases, under correctly specified models, all estimators produce unbiased estimates and the same idea used in chapter 2 can be used here to derive estimating equations and the asymptotic covariance matrix. Since we are now estimating both $\pi(\mathbf{x}, y) = \text{pr}(R = 1|\mathbf{x}, y; \boldsymbol{\alpha})$ and $e_i(\mathbf{x}_i) = \text{pr}(T = 1|\mathbf{x}; \boldsymbol{\eta})$, the enlarged estimating equa-

tions system is

$$\boldsymbol{S}(\boldsymbol{\phi}) = \boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \begin{pmatrix} \boldsymbol{S}_0(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\eta}) \\ \boldsymbol{S}_1(\boldsymbol{\alpha}) \\ \boldsymbol{S}_2(\boldsymbol{\eta}) \end{pmatrix}$$

where $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$ are given by equations (2.7) and (2.8), respectively, and

$$\boldsymbol{S}_2 = \frac{\partial}{\partial \boldsymbol{\eta}} \sum_i \Big(T_i \log(e_i(\mathbf{x}_i)) + (1 - T_i) \log(1 - e_i(\mathbf{x})) \Big).$$

Estimates $\hat{\boldsymbol{\phi}}$ are obtained by setting $\boldsymbol{S}(\hat{\boldsymbol{\phi}}) = 0$. Reasoning as in section 2.2, the asymptotic variance is given by

$$ACov(\boldsymbol{\beta}) = \boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{C}}_{00} \boldsymbol{\mathcal{I}}_{00}^{-1} - \boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{C}}_{01} \boldsymbol{\mathcal{I}}_{11}^{-1} \boldsymbol{\mathcal{C}}_{01}^{T} \boldsymbol{\mathcal{I}}_{00}^{-1} - \boldsymbol{\mathcal{I}}_{00}^{-1} \boldsymbol{\mathcal{C}}_{02} \boldsymbol{\mathcal{I}}_{22}^{-1} \boldsymbol{\mathcal{C}}_{02}^{T} \boldsymbol{\mathcal{I}}_{00}^{-1}, \tag{5.6}$$

where the first and second terms are, as in section 2.2.1, due to estimating β and π_1 and so the third term is what we get by estimating $\tilde{e}$. We can still show that $\boldsymbol{\mathcal{C}}_{01} = \boldsymbol{\mathcal{I}}_{01}$, but we can no longer guarantee that $\boldsymbol{\mathcal{C}}_{02} = \boldsymbol{\mathcal{I}}_{02}$. All methods were implemented in R and the maximization is carried out by the Newton-Raphson method.

5.4.2 Simulations

For the following simulations we considered the 2-phase design shown in Fig. 5.4. Here the goal was to estimate the treatment effect β_T. To this end, we assumed that information on (Y, T, X_1) was collected for all individuals, known as the phase-1 data, and that the remaining information was observed for only a sample of them. The sampling was based on (Y, T), where Y can be continuous or discrete and T is the

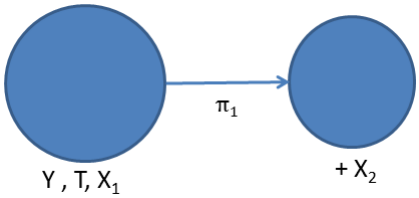


Figure 5.4: Sampling scheme for a 2-phase study, where the Y, X_1 and the treatment T are fully observed at phase-1 and the remaining variable X_2 is observed only at phase-2.

treatment indicator, and the true model of interest used to generate the data is

$$\text{logit}(\text{pr}(Y = 1|T, \mathbf{x}; \boldsymbol{\beta})) = \beta_0 + \beta_T T + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \tag{5.7}$$

for a binary response or

$$y = \beta_0 + \beta_T T + \beta_1 X_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon, \tag{5.8}$$

for a continuous response. In both cases, X_1 and X_2 are independent random variables following a standard normal distribution, $X_3 = X_1^2$ and T is a binary variable taking the value 1 with probability e, where

$$\text{logit}(e(\mathbf{x}; \boldsymbol{\eta})) = \eta_0 + \eta_1 x_1 + \eta_3 x_3. \tag{5.9}$$

We ran 1000 simulations for each scenario and compared all methods described in previous chapters as well as those methods described in section 5.4.1 which make use of propensity scores, with respect to their bias and efficiency for different treatment effects and also under slightly misspecified models. These methods are:

- Complete data analysis (*complete data*), which uses only data from phase-2;

- Multiple imputation alone (*MI*);
- Weighted method used throughout this thesis, with corrected weights $1/\pi$ and with (*wgt+ps*) and without (*wgt*) the propensity score included as an extra covariate into the regression model of interest;
- Conditional likelihood method with (*cml+ps*) and without (*cml*) the propensity score included as an extra covariate into the regression model of interest;
- CML weighted by propensity score (*cml reg*);
- CML with the extra information $\widetilde{\boldsymbol{S}}$ added (*cml+*$\widetilde{\boldsymbol{S}}$).

For the following simulations, let $\boldsymbol{X} = (X_1, X_2)$ be normally distributed with mean $\mathbf{0}$ and variance $\Sigma = \mathbb{I}$, where $\mathbb{I}$ is the 3×3 identity matrix, and let $\epsilon \sim N(0,1)$. We considered a fixed treatment effect and ran 1,000 simulations varying the coefficient β_3, the coefficient of the missing variable.

Binary response

Table 5.4 shows the results for a discrete response with parameters $\boldsymbol{\eta} = (\eta_0, \eta_1, \eta_3)^T = (1, -1, 1)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_T, \beta_1, \beta_2, \beta_3)^T = (-4.5, \log(2), 2, 2, \beta_3)^T$. Here both covariates X_1 and X_2 have strong effects and the treatment effect has an odds ratio of 2. From a total population of $N = 15{,}000$ individuals with observed (Y, T, X_1), $n = 100$ were sampled from each stratum generated by (Y, T) and X_2 observed. The true model of interest and the true propensity score are given by equations (5.7) and (5.9), respectively. We assumed that the quadratic term X_3 included in the fitted models, that is, we fitted the models

$$\text{logit}(\text{pr}(Y = 1 | T, \mathbf{x}, \boldsymbol{\beta})) = \beta_0 + \beta_T T + \beta_1 x_1 + \beta_2 x_2$$

and

$$\text{logit}(e(\mathbf{x})) = \eta_0 + \eta_1 x_1.$$

to the data. Both models are then misspecified and the degree of misspecification is controlled by the coefficients β_3 and η_3, from equations (5.7) and (5.9), respectively.

Table 5.4 shows the results for four scenarios based on different values of β_3. We started with a reasonably strong correlation between X_3 and Y and later decreased the coefficient β_3 so the misspecification is unlikely to be detected. For $\beta_3 = .05$ we detected departures from $\beta_3 = 0$ 34% of the time, while for $\beta_3 = .1$, the detection rate was 69%, at the 5% level.

From Table 5.4 we see that by using propensity scores as covariates we are able to remove most of the bias due to the omitted quadratic term X_3, even in situations where the model misspecification is unlikely to be detected. When the model of interest is correctly specified, i.e., $\beta_3 = 0$, all methods, with the exception of the complete data and MI methods, result in unbiased or nearly unbiased estimates. The most efficient, as expected, is the CML+$\widetilde{\boldsymbol{S}}$ method, followed by the CML method.

As the misspecification increases, the CML+ps and the wgt+ps methods are still able to remove almost entirely the bias, while the other methods show a relative bias, which is defined as

$$\Delta = \frac{\beta_T - \hat{\beta}_T}{\beta_T},$$

that varies from 0 to almost 10%. When $\beta_3 = .1$, for example, the CML+$\widetilde{\boldsymbol{S}}$ method shows a relative bias of about 9%, but at the same time still has the lowest MSE among all methods analyzed. It is about 45% lower than the wgt+ps method and about 9% lower than CML+ps.

Table 5.4: Results for β_T with discrete Y, for 1000 datasets simulated

Coefficients	Method	Estimates	Bias	Relative Bias	Est. SE	Emp. SE	MSE	% Coverage
$\boldsymbol{\beta} = (-4.5, \log 2, 2, 2, 0)$	mi	0.69	-0.01	-0.01	0.14	0.17	29.85	0.85
	wgt	0.70	0.01	0.01	0.20	0.21	43.42	0.94
	wgt+ps	0.70	0.01	0.01	0.22	0.23	53.57	0.93
	cml	0.70	0.00	0.01	0.16	0.16	25.49	0.95
	cml+wgt	0.70	0.00	0.00	0.46	0.17	28.44	1.00
	cml+ps	0.70	0.01	0.01	0.17	0.18	33.06	0.95
	cml+s	0.70	0.00	0.01	0.16	0.16	25.49	0.95
$\boldsymbol{\beta} = (-4.5, \log 2, 2, 2, .05)$	mi	0.71	0.02	0.03	0.14	0.16	25.80	0.86
	wgt	0.74	0.05	0.07	0.20	0.20	43.63	0.94
	wgt+ps	0.70	0.01	0.01	0.22	0.23	51.78	0.94
	cml	0.73	0.04	0.06	0.16	0.15	24.12	0.95
	cml+wgt	0.74	0.04	0.06	0.46	0.16	28.22	1.00
	cml+ps	0.70	0.00	0.01	0.18	0.17	28.67	0.96
	cml+s	0.73	0.04	0.06	0.16	0.15	24.19	0.95
$\boldsymbol{\beta} = (-4.5, \log 2, 2, 2, .1)$	mi	0.73	0.04	0.06	0.14	0.16	28.37	0.86
	wgt	0.75	0.06	0.09	0.20	0.20	42.06	0.94
	wgt+ps	0.69	0.00	0.00	0.22	0.22	48.46	0.95
	cml	0.76	0.06	0.09	0.16	0.15	27.54	0.94
	cml+wgt	0.76	0.06	0.09	0.46	0.16	30.10	1.00
	cml+ps	0.70	0.00	0.00	0.18	0.17	29.87	0.96
	cml+s	0.75	0.06	0.09	0.16	0.15	27.28	0.95

Continuous response

Table 5.5 shows the results for a continuous response with parameters $\boldsymbol{\beta} = (1,5,5,1,\beta_3)^T$ and $\boldsymbol{\eta} = (-1,1,1)^T$. We used $N = 15{,}000$ units with observed (Y,T,X_1) from which $n = 100$ subjects were sampled from each stratum generated by (Y_d,T), where Y_d is equals to 1 if $Y < c$ (c is 30th percentile) and 0 otherwise. X_2 was observed only for the sampled units. The true model of interest and the true propensity score are given by equations (5.8) and (5.9), respectively, but, as before, we assumed that the quadratic term X_3 was not included in the fitted model and thus fitted the models

$$y = \beta_0 + \beta_T T + \beta_1 x_1 + \beta_2 x_2$$

and

$$\text{logit}(e(\mathbf{x})) = \eta_0 + \eta_1 x_1.$$

to the data.

Table 5.5 shows the results for $\beta_3 = 0, .05$ and $.1$. For $\beta_3 = .05$ we detected departures from $\beta_3 = 0$ 59% of the time, while for $\beta_3 = .1$, the detection rate was 94%, at the 5% level. That is, for $\beta_3 = .1$, we are very likely to detect the model misspecification.

As expected, when there is no model misspecification ($\beta_3 = 0$), all methods are essentially unbiased and CML+$\widetilde{\boldsymbol{S}}$ is the most efficient method. As the model misspecification increases, all methods become slightly biased and the CML+$\widetilde{\boldsymbol{S}}$ method turns out to be almost twice as biased as CML+ps and about 3 times as biased as wgt+ps. The relative biases are still small for all methods, but enough to affect the MSE. For example, when $\beta_3 = .05$, MI is the most efficient method, followed closely by the wgt method. When $\beta_3 = .1$, all biases are even greater and the wgt+ps method, which shows the smallest bias, is the most efficient method. CML+ps is also able to remove most bias and, as a result, its MSE is 43% lower than the one obtained from the CML+$\widetilde{\boldsymbol{S}}$

Table 5.5: Results for β_T with continuous Y, for 1000 datasets simulated

Coefficients	Method	Estimates	Bias	Relative Bias	Est. SE	Emp. SE	MSE	% Coverage
$\boldsymbol{\beta} = (1,5,5,1,0)$	mi	4.99	-0.01	-0.00	0.08	0.08	6.27	0.90
	wgt	4.99	-0.01	-0.00	0.08	0.08	6.95	0.94
	wgt+ps	4.99	-0.01	-0.00	0.10	0.10	9.66	0.94
	cml	4.99	-0.01	-0.00	0.08	0.07	5.03	0.96
	cml+wgt	5.00	-0.00	-0.00	0.08	0.08	7.15	0.96
	cml+ps	5.00	-0.00	-0.00	0.09	0.09	8.62	0.93
	cml+s	5.00	-0.00	-0.00	0.07	0.07	5.33	0.95
$\boldsymbol{\beta} = (1,5,5,1,.05)$	mi	5.06	0.06	0.01	0.08	0.08	10.27	0.81
	wgt	5.06	0.06	0.01	0.08	0.08	10.62	0.88
	wgt+ps	5.03	0.03	0.01	0.10	0.10	10.98	0.94
	cml	5.09	0.09	0.02	0.08	0.07	14.17	0.79
	cml+wgt	5.10	0.10	0.02	0.08	0.10	19.35	0.77
	cml+ps	5.05	0.05	0.01	0.09	0.10	13.64	0.88
	cml+s	5.09	0.09	0.02	0.07	0.07	13.65	0.79
$\boldsymbol{\beta} = (1,5,5,1,.1)$	mi	5.11	0.11	0.02	0.08	0.08	18.37	0.67
	wgt	5.11	0.11	0.02	0.08	0.09	18.74	0.75
	wgt+ps	5.05	0.05	0.01	0.10	0.10	12.60	0.93
	cml	5.17	0.17	0.03	0.08	0.08	35.43	0.44
	cml+wgt	5.17	0.17	0.03	0.08	0.09	36.01	0.49
	cml+ps	5.08	0.08	0.02	0.09	0.11	19.26	0.84
	cml+s	5.16	0.16	0.03	0.08	0.08	33.25	0.43

method (without propensity scores corrections).

5.5 Generalized propensity score

Note that the propensity score in section 5.3.1 was a function of (X_1, X_2), where X_2 had been partially observed. Thus, it could not be estimated for the entire data, but only for the fully observed sample. Unless we use the extra information provided by the partially observed data, we are likely to get inefficient estimates. Some research has been done on how to estimate propensity scores with missing data

This problem was first noticed by Rosenbaum and Rubin (1985), where they introduced the generalized propensity score method. Here, the probability of being treated is now conditioned on all observed covariates as well as on the response indicator R_i. In other words, if we denote by $\boldsymbol{X}_{\text{obs}}$ and $\boldsymbol{X}_{\text{mis}}$ the observed and missing covariates, the generalized propensity score e_c is defined as

$$e_c = \text{pr}\left(T_i = 1 | \mathbf{x}_{obs}, R_i\right).$$

The authors suggest using a "pattern mixture model" (Little, 1993) to estimate the propensity score with missing covariates. They suggest that e_c can be estimated by a separate logit model for each pattern of missing data, using the fully observed covariates in each case. However, as the number of patterns of missing data increases, some of them might have very few observations, which would then make this technique infeasible.

D'Agostino and Rubin (2000) suggest modelling the joint distribution of $(T, \boldsymbol{X}, R)$ using the EM algorithm. This approach can be hard to implement and simpler and efficient techniques are still of interest. A common approach is to use multiple imputation (MI), as done by Song et al. (1999) and Crowe et al. (2010).

Crowe et al. (2010) compared four imputation techniques to obtain $\hat{p}$: treatment mean imputation, where the missing value is replaced by the mean of the observed values, within treatment; a MI model using only the covariates; a MI model using the covariates and the treatment; and another MI model including all observed variables, including treatment and outcome. Through simulations the authors showed that the best model appears to be the last one, which is also in agreement with Song et al. (1999).

Qu and Lipkovich (2009) developed the Multiple Imputation Missingness Pattern (MIMP) method, which also uses MI to handle missing data while estimating the propensity score. Here, the authors first use MI to obtain the imputed data and later regress T on the covariates and on a new variable containing the missingness pattern, obtaining $\hat{p}$.

Stumer et al. (2007) considered a 2-phase design with a non-monotone missing pattern. In their design Y, T and X_1 are observed for a population of size N while an extra confounder X_2, say, is observed for another population together with Y and T. T is correlated with $\boldsymbol{X} = (X_1, X_2)$ and, in order to get unbiased estimates for the propensity score, the authors define two propensity scores based on each population and use MI to create a complete propensity score and regress Y on T and both propensity scores.

We will return to the generalized propensity score in the next chapter, where it will be studied from a more general point of view and will be estimated using the CML method.

5.6 Summary

Here we have discussed regression adjustment with propensity scores and derived new estimating equations for estimating treatment effects in a multiphase study.

Unlike in most approaches discussed in the literature, we take into account variability in estimating the propensity score. From our simulations, we see that if the regression model of interest is correctly specified, adding propensity scores as an extra covariate leads to losses in efficiency. It can, however, be used as a protection against model misspecification in a few situations. For example, suppose that there are measured confounders that for some reason cannot be included into the model of interest, but could be included into the propensity score model. Then, including the estimated propensity score as an extra covariate into the model of interest will reduce, or even avoid, the omitted-variable bias. Moreover, if the true regression model is not linear on $\boldsymbol{X}$, but the fitted model is, propensity scores may be used as a way to reduce bias. In our simulations, however, we noticed that for a quadratic $\boldsymbol{X}$ and moderate misspecifications (deletion rate lower than 50%), propensity scores did not improve the MSE. Including all variables into the regression of interest, when possible, seemed to be the best choice. Finally, even though it looks appealing, to add all variables and their interactions into the propensity score model in order to minimize omitted-variable bias might not be the optimum solution, especially if $\boldsymbol{X}$ is of high-dimension. Deciding which variables should and should not be included into the model is, however, outside the scope of this study and the interested reader is referred to Clarke et al. (2011).

6

Beyond the Simple 2-phase Design

So far we have only worked with multiphase designs where interest was centred on fitting a regression model for $\boldsymbol{Y}$, known for all at phase-1, on a set of covariates $\boldsymbol{X}$ that could potentially be correlated with the missingness as long the MAR assumption was valid. We now discuss more general designs, where an auxiliary variable $\boldsymbol{V}$ is fully observed at phase-1, but is not part of the model of interest. Such designs generalize those discussed in previous chapters and also incorporate the secondary analysis problem discussed in Neuhaus et al. (2006) and Jiang et al. (2006). We also present a new and more general semiparametric estimator that copes with these different designs, whether $\boldsymbol{V}$ and Y are continuous or discrete, discuss its performance under different scenarios, and apply it to a real dataset.

6.1 More general designs

In chapter 5 we were interested in estimating the treatment effect in a 2-phase design, where units were selected for further observation based on the response $\boldsymbol{Y}$ and the treatment indicator T. Here we consider a more general framework: suppose that $\boldsymbol{V}$ is an additional (discrete or continuous) variable observed for all at phase-1. We are concerned with two cases:

(i) The response $\boldsymbol{Y}$ is also observed for all at phase-1 and the selection into phase-2 is a function of $(\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{X}_1)$, where $\boldsymbol{X}_1$, a component of $\boldsymbol{X}$, is also observed for all at phase-1 (see figure 6.1);

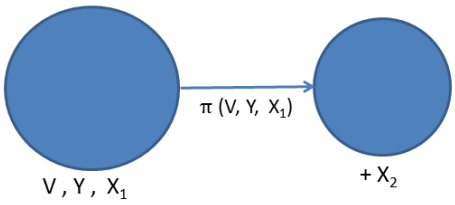


Figure 6.1: Sampling schemes for case (i). $(\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{X}_1)$ are fully observed while $\boldsymbol{X}_2$ is observed only at phase-2.

(ii) $\boldsymbol{Y}$, which could be correlated to $\boldsymbol{V}$, is only observed at the phase-2 and the selection model depends on $(\boldsymbol{V}, \boldsymbol{X}_1)$ (see figure 6.2).

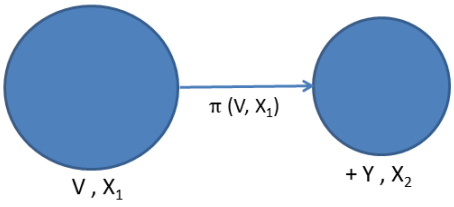


Figure 6.2: Sampling schemes for case (ii). Here the response of interest $\boldsymbol{Y}$ and $\boldsymbol{X}_2$ are only observed at phase-2 while a design variable $\boldsymbol{V}$ as well as $\boldsymbol{X}_1$ are fully observed at phase-1.

Notice that when $\boldsymbol{V}$ is null, case (i) is reduced to the simple 2-phase problem with a fully observed response $\boldsymbol{Y}$ and partially observed covariate $\boldsymbol{X}$. This problem has been discussed throughout this thesis and the Woman's Health Initiative program (see section 3.4) is an illustration of such a design.

In addition to the simple 2-phase problem, case (i) also generalizes a number of other situations as follows.

- *Propensity scores:* The propensity score and the generalized propensity score approaches discussed in chapter 5 are special cases of (i), with $\boldsymbol{V}$ replacing T. Under the additional assumption of $\boldsymbol{V}$ being strongly ignorable given $\boldsymbol{X}$, case (i) is formally equivalent to the 2-phase treatment effect problem discussed in chapter 5, with $\boldsymbol{V}$ replacing T.

- *The expensive covariate problem:* In large studies, collecting full information for the entire population or cohort is often expensive or even unfeasible. Some characteristics are then only partially observed, being measured for a small percentage of the total cohort. Let $\boldsymbol{X}_2$ be this expensive or hard to observe variable. Even though $\boldsymbol{X}_2$ is not available at phase-1, an approximate surrogate $\boldsymbol{V}$ for $\boldsymbol{X}_2$ may be available and this extra information can be used to produce more efficient estimates. This problem was also discussed in Zhou et al. (2011), where the authors use the empirical likelihood method to derive efficient estimators that require a parametric model for the conditional distribution of $\boldsymbol{X}|\boldsymbol{V}$.

- *Missing by design problem with extra variable included into the selection model:* Sometimes $\boldsymbol{V}$ can also be used to select the phase-2 data, without being used as a predictor in the model of interest. For example, if $\boldsymbol{V}$ is an approximate surrogate for $\boldsymbol{X}_2$ (as in the expensive covariate problem), it may be used just to select the phase-2 data. Wang et al. (2009), motived by a lung cancer biomarker

study, also discuss this problem. Here, a patient is selected into the subset based on the outcome of interest $\boldsymbol{Y}$ (tumour response) and on an surrogate variable $\boldsymbol{V}$ (the likelihood score of epidermal growth factor receptor inhibitors mutations) for an expensive covariate $\boldsymbol{X}$ (related to genotyping the epidermal growth factor receptor genes). This surrogate variable, however, is only used for selecting the phase-2 data and is not included in the model of interest.

- *Missing by happenstance problem with extra variable included into the selection model:* Note that, in the previous case, including $\boldsymbol{V}$ into the selection model was a design decision. A similar idea is used in the missing data context, more specifically when data is missing not by design, but by happenstance. This missing data problem can also be seen as multi-phase problem as long as the missingness follows a monotonic pattern (see chapter 1). And if $\boldsymbol{V}$ is thought to be associated with the non-response, it must be included into the selection model whether or not it is part of the model of interest.

- *Adding variables into the selection model to increase efficiency:* As discussed in chapter 3, including more variables into the selection model will never decrease and may even increase asymptotic efficiency, even if these extra variables are not predictors of the true selection model. This was also discussed in Scott and Wild (2011) and in chapters 3 through simulations and while analysing the WHI data (see section 3.4).

Even though case (ii) can also be re-formulated as a 2-phase problem, it is slightly different from the previous ones considered so far. Here, the response of interest $\boldsymbol{Y}$ is observed only at phase-2 and extra information regarding $\boldsymbol{V}$ and $\boldsymbol{X}_1$, say, are available for all data. Note that all examples discussed above are special cases of this design as well as some important problems that are now discussed.

- *Approximate surrogate for* $\boldsymbol{Y}$*:* Suppose that the response $\boldsymbol{Y}$ is expensive or hard to measure for the entire study population. If we can measure some extra variable $\boldsymbol{V}$ that is closely related to $\boldsymbol{Y}$, we can use it to select the phase-2 data where the response of interest $\boldsymbol{Y}$ is then observed. An example (which is due to Clayton et al. (1998)) can also be seen in Zhao et al. (2009), where the goal is to estimate the prevalence of dementia in the elderly. Here, a definitive diagnosis of dementia (response of interest, $\boldsymbol{Y}$) is difficult and expensive to make so that it is available for only a subsample of the total population. However, a Mini-Mental State Examination, which provides a score that is an imperfect measure of dementia status is administered to each person. The authors use the profile likelihood method with the EM algorithm to estimate the parameters of interest. This problem is also discussed in Neuhaus et al. (2006) and Neuhaus et al. (2013), where the authors deal with a longitudinal binary response and discrete $\boldsymbol{V}$ and suggest working with the full or with the conditional likelihood

- *Secondary analysis:* Another important example of case (ii) is the secondary analysis problem. Here, we are interested in performing a secondary analysis, where $\boldsymbol{V}$ is a design variable that was used for data collection (e.g., $\boldsymbol{V}$ was the response variable collected for the primary analysis) and is correlated to the outcome $\boldsymbol{Y}$. This has been discussed by Jiang et al. (2006), where a variety of methods are compared while performing secondary analysis of case-control data. The authors work with a semiparametric approach based on Scott and Wild (2006) with $\boldsymbol{Y}^* = (\boldsymbol{Y}, \boldsymbol{V})$ as well as fully non-parametric approach and parametric modelling of the conditional distribution of $\boldsymbol{V}|(\boldsymbol{Y}, \boldsymbol{X})$. It is also discussed by Lee et al. (1997), where data from a case-control study on Sudden Infant Death Syndrome (SIDS) is analysed. The main design variable here $\boldsymbol{V}$ is an indicator of being or not being a SIDS victim but interest in the secondary analysis is in modelling the

chance that a child would receive the standard childhood inoculations ($\boldsymbol{Y}$).

In this chapter we develop estimating equations that cope with both cases (i) and (ii). The proposed method is semiparametric in the sense that the distribution of the covariates $\boldsymbol{X}$ is treated non-parametrically. That is, unlike the approach discussed in Zhou et al. (2011) with respect to the expensive covariate problem, we do not assume any model for the conditional distribution of $\boldsymbol{X}|\,\boldsymbol{V}$ and information regarding $\boldsymbol{V}$ is still taken into account. We also allow the outcome of interest $\boldsymbol{Y}$ and the additional variable $\boldsymbol{V}$ to be correlated and to be either continuous or discrete. This extends the work done by Lee et al. (1997), Neuhaus et al. (2006) and Jiang et al. (2006), where only discrete $\boldsymbol{Y}$ and $\boldsymbol{V}$ was considered.

As we are now dealing with a more general design that is different from those discussed in previous chapters, we need to derive new estimating equations that cope with both cases (i) and (ii) presented earlier in this chapter.

6.2 General approach

Let $f(\mathbf{y}|\mathbf{x};\boldsymbol{\beta})$ be the model of interest, where $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, and $h(\boldsymbol{v}|\mathbf{y},\mathbf{x},\mathbf{w};\boldsymbol{\zeta})$ the conditional probability density function of $\boldsymbol{V}$ given ($\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}$). Here $\boldsymbol{W}$ is a variable used only to predict $\boldsymbol{V}$, but not $\boldsymbol{Y}$. For example, in the secondary analysis problem discussed in previous section, $\boldsymbol{W}$ can be an extra covariate for the primary analysis (when $\boldsymbol{V}$ was the outcome), but it is not included into the model of interest $f(\mathbf{y}|\mathbf{x};\boldsymbol{\beta})$.

Let, in addition, ($\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{X}_1$) be fully observed at phase-1 and $\boldsymbol{Z}$ be a function of the variables fully observed at phase-1. Suppose that $\boldsymbol{X}_2$ is observed only at phase-2 and let, as before, R_i be an indicator variable denoting which unit was selected into phase-2.

Under the MAR assumption, the full likelihood L_F for case (i) is

$$L_F \propto \prod_{i:R_1=1} f(\mathbf{y}_i|\mathbf{x}_i,\mathbf{w}_i;\boldsymbol{\beta})h(\boldsymbol{v}_i|\mathbf{y}_i,\mathbf{x}_i,\mathbf{w}_i;\boldsymbol{\zeta})g(\mathbf{x}_i,\mathbf{w}_i) \prod_{i:R_1=0} f(\mathbf{y}_i,\boldsymbol{v}_i,\mathbf{w}_i,\mathbf{x}_{1i}),$$

where

$$f(\mathbf{y}_i,\boldsymbol{v}_i,\mathbf{w}_i,\mathbf{x}_{1i}) = \int f(\mathbf{y}_i|\mathbf{w}_i,x;\boldsymbol{\beta})h(\boldsymbol{v}_i|\mathbf{y}_i,\mathbf{w}_i,\mathbf{x};\boldsymbol{\zeta})g(\mathbf{w}_i,\mathbf{x}_i)d\mathbf{x}_2. \tag{6.1}$$

A similar expression can be derived for case (ii), noticing that the integral (6.1) is also calculated with respect to $\boldsymbol{Y}$.

Unless $\boldsymbol{W}$ and $\boldsymbol{Y}$ are conditionally independent given $\boldsymbol{X}$, we would not be working with the model of interest $f(\mathbf{y}_i|\mathbf{x}_i)$, but with $f(\mathbf{y}_i|\mathbf{w}_i,\mathbf{x}_i)$. Thus, for everything that follows we consider that $f(\mathbf{y}_i|\mathbf{w}_i,\mathbf{x}_i) = f(\mathbf{y}_i|\mathbf{x}_i)$, that is, $\boldsymbol{W}$ and $\boldsymbol{Y}$ are conditionally independent given $\boldsymbol{X}$. If we are not prepared to make such assumption, this variable should be included in $\boldsymbol{V}$.

Finally, notice that the full or complete likelihood depends on the distribution of $(\boldsymbol{W},\boldsymbol{X})$ and, as discussed in previous chapters, modelling its joint distribution may be hard or even infeasible since these variables are usually high dimensional. We consider then the semiparametric conditional maximum likelihood method (CML), which was already discussed in chapters 3 and 4 and will be used here to develop estimating equations for the parameter of interest $\boldsymbol{\beta}$.

6.2.1 Conditional maximum likelihood

In addition to the models defined in previous section, let $\pi(\mathbf{z},;\boldsymbol{\alpha}) = E(R_i|\ \mathbf{z};\boldsymbol{\alpha})$ be a parametric model for the probability of being selected into phase-2 and assume that the missing at random (MAR, see section 1.2.2) assumption holds. The conditional

likelihood is given by

$$\begin{aligned} L_c &= \prod_i f(\mathbf{y}_i, \boldsymbol{v}_i | \mathbf{x}_i, \mathbf{w}_i, R_{1i} = 1; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\zeta}) \\ &= \prod_i h(\boldsymbol{v}_i | \mathbf{y}_i, \mathbf{x}_i, \mathbf{w}_i, R_{1i} = 1; \boldsymbol{\alpha}, \boldsymbol{\zeta}) f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{w}_i, R_{1i} = 1; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &= \prod_i h(\boldsymbol{v}_i | \mathbf{y}_i, \mathbf{x}_i, \mathbf{w}_i, R_{1i} = 1; \boldsymbol{\alpha}, \boldsymbol{\zeta}) f(\mathbf{y}_i | \mathbf{x}_i, R_{1i} = 1; \boldsymbol{\beta}, \boldsymbol{\alpha}), \end{aligned}$$

since $\boldsymbol{W}$ and $\boldsymbol{Y}$ are conditionally independent given $\boldsymbol{X}$. Thus, we have that

$$\begin{aligned} L_c &= \prod_i \frac{\pi(\boldsymbol{v}_i, \mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha}) h(\boldsymbol{v}_i | \mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\zeta})}{\int \pi(\boldsymbol{v}, \mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha}) h(\boldsymbol{v} | y_i, w_i, \mathbf{x}_i; \boldsymbol{\zeta}) d\boldsymbol{v}} \frac{\pi_v(\mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha}, \boldsymbol{\zeta}) f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\beta})}{\int \pi_v(\boldsymbol{y}, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha}, \boldsymbol{\zeta}) f(\mathbf{y} | \mathbf{x}_i; \boldsymbol{\beta}) d\mathbf{y}} \\ &= L_1(\boldsymbol{\alpha}, \boldsymbol{\zeta}) L_2(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\zeta}), . \end{aligned} \tag{6.2}$$

Writing the likelihood as above is convenient for programming. L_1 and L_2 have a similar structure, which also resembles those obtained in chapter 2. The selection probability $\pi = \text{pr}(R = 1 | \mathbf{y}, \boldsymbol{v}, \mathbf{w}, \mathbf{x}; \boldsymbol{\alpha})$ and

$$\pi_v(\mathbf{y}, \mathbf{w}, \mathbf{x}, ; \boldsymbol{\alpha}, \boldsymbol{\zeta}) = \mathbb{E}\left[\pi(\mathbf{Y}, \boldsymbol{V}, \mathbf{W}, \mathbf{X}; \boldsymbol{\alpha}) | \mathbf{y}, \mathbf{w}, \mathbf{x}\right] = \int \pi(\mathbf{y}, \boldsymbol{v}, \mathbf{w}, \mathbf{x}; \boldsymbol{\alpha}) h(\boldsymbol{v} | \mathbf{y}, \mathbf{w}, \mathbf{x}; \boldsymbol{\zeta}) d\boldsymbol{v}.$$

Later in simulations we will investigate the effect of using $\pi(\mathbf{y}, \boldsymbol{v}, \mathbf{w}, \mathbf{x}; \boldsymbol{\alpha})$ without taking the expectation over $\boldsymbol{V}$ given ($\boldsymbol{Y}$, $\boldsymbol{W}$, $\boldsymbol{X}$). We will call this naive CML.

Note that if $\boldsymbol{V}$ is not used for selecting the phase-2 data and is not used as part of the selection model (e.g., to gain efficiency), that is, if $\pi(\mathbf{y}, \boldsymbol{v}, \mathbf{w}, \mathbf{x}; \boldsymbol{\alpha}) = \pi(\mathbf{y}, \mathbf{w}, \mathbf{x}; \boldsymbol{\alpha})$, we have that $\pi_{\boldsymbol{v}}(\mathbf{y}, \mathbf{w}, \mathbf{x}; \boldsymbol{\alpha}, \boldsymbol{\zeta}) = \pi(\mathbf{y}, \mathbf{w}, \mathbf{x}; \boldsymbol{\alpha})$ and the conditional likelihood L_c can be factorized into $L_1(\boldsymbol{\zeta}) L_2(\boldsymbol{\beta}, \boldsymbol{\alpha})$. The parameter of interest $\boldsymbol{\beta}$ can then be estimated without assuming a model for $h(\boldsymbol{v} | \mathbf{y}, \mathbf{w}, \mathbf{x})$.

Consider now the case where $\boldsymbol{V}$ is fully observed at phase-1 and used to select the

phase-2 data. The outcome of interest $\boldsymbol{Y}$ can be fully observed, which corresponds to case (i) discussed earlier in this chapter, or observed only at phase-2, corresponding to case (ii). One advantage of working with the CML method over the full likelihood is that, unlike in the latter where each case results in different likelihoods and thus different estimating equations, CML handles these both cases (i) and (ii) in a similar way. The only difference between case (i) and (ii), with respect to the CML method, is related to the selection probability π. The estimating equations have the same structure whether $\boldsymbol{Y}$ has been fully observed at phase-1 or not. That is, in both cases the coefficients of interest can be estimated by solving the enlarged estimating equations system

$$\boldsymbol{S}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\zeta}) = \begin{pmatrix} \boldsymbol{S}_0(\boldsymbol{\beta},\boldsymbol{\zeta},\boldsymbol{\alpha}) \\ \boldsymbol{S}_1(\boldsymbol{\alpha}) \\ \boldsymbol{S}_2(\boldsymbol{\beta},\boldsymbol{\zeta},\boldsymbol{\alpha}) \end{pmatrix}$$

where

$$\begin{aligned} \boldsymbol{S}_0(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\zeta}) &= \frac{\partial \log L_c(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\zeta})}{\partial \boldsymbol{\beta}}, \\ \boldsymbol{S}_2(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\zeta}) &= \frac{\partial \log L_c(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}}. \end{aligned} \tag{6.3}$$

As before (see equation (2.8), for instance), $\boldsymbol{S}_1$ is the estimating equation associated with the selection probability $\pi(\mathbf{z};\boldsymbol{\alpha})$ and is given by

$$\boldsymbol{S}_1(\boldsymbol{\alpha}) = \sum_i \left(\frac{R_1 - \pi(\mathbf{z})}{\pi(\mathbf{z})(1-\pi(\mathbf{z})} \frac{\partial \pi}{\partial \boldsymbol{\alpha}} \right),$$

where $\boldsymbol{Z}$ contains the phase-1 variables used in the selection model. Thus, (i) and (ii) only differ on $\boldsymbol{Z}$ and so both cases can then be treated together.

Finally, the information matrix is given by

$$\boldsymbol{\mathcal{I}} = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{00} & \boldsymbol{\mathcal{I}}_{01} & \boldsymbol{\mathcal{I}}_{02} \\ 0 & \boldsymbol{\mathcal{I}}_{11} & 0 \\ \boldsymbol{\mathcal{I}}_{20} & \boldsymbol{\mathcal{I}}_{21} & \boldsymbol{\mathcal{I}}_{22} \end{pmatrix},$$

where

$$\boldsymbol{\mathcal{I}}_{kl} = \frac{\partial \boldsymbol{S}_k}{\partial \boldsymbol{\phi}_k \partial \boldsymbol{\phi}_l^T},$$

for $k = 0, 1, 2$, $l = 0, 1, 2$ and $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\zeta})^T$. The asymptotic covariance and the estimates of interest can then be obtained following the same steps as in chapters 2 and 4.

In particular, for discrete V, the ascertainment-corrected model of interest can be written as

$$f(\mathbf{y}_i, v_i | \mathbf{w}_i, \mathbf{x}_i, R_{1i} = 1; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\zeta}) = \frac{\pi(v_i, \mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha}) h(v_i | \mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\zeta})}{\sum_j \pi(v_j, \mathbf{y}_i, \mathbf{w}_i, \boldsymbol{X}_i; \boldsymbol{\alpha}) h(v_j | \mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\zeta})} \times \frac{\pi_v(\mathbf{y}_i, \mathbf{x}_i, \mathbf{w}_i; \boldsymbol{\alpha}, \boldsymbol{\zeta}) f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\beta})}{\int \pi_v(\mathbf{y}_j, \mathbf{x}_i, \mathbf{w}_i; \boldsymbol{\alpha}, \boldsymbol{\zeta}) f(\mathbf{y} | \mathbf{x}_i \boldsymbol{\beta}) d\mathbf{y}}.$$

If we further assume that Y and V are both binary and logistic regression models

$$\text{pr}(V = 1 | Y = j, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\zeta}) = \text{logit}(h(v_1 | y_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\zeta})) = \mathbf{k}_i^T \boldsymbol{\zeta}$$

and

$$\text{pr}(Y = 1 | \mathbf{x}_i; \boldsymbol{\beta}) = \text{logit}(f(y_1 | \mathbf{x}_i; \boldsymbol{\beta})) = \mathbf{x}_i^T \boldsymbol{\beta},$$

for $j = \{0, 1\}$ and $\boldsymbol{K} = (Y, \boldsymbol{W}, \boldsymbol{X})$, we have that

$$f(y_i, v_i | \mathbf{w}_i, \mathbf{x}_i, R_{1i} = 1; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\zeta}) = \frac{e^{o_{1i} + \mathbf{k}_i^T \boldsymbol{\zeta}}}{1 + e^{o_{1i} + \mathbf{k}_i^T \boldsymbol{\zeta}}} \frac{e^{o_{2i} + \mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{o_{2i} + \mathbf{x}_i^T \boldsymbol{\beta}}}.$$

The two offsets are given by

$$o_{1i} = \log\left(\frac{\pi(v_i^{(1)}, y_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha})}{\pi(v_i^{(0)}, y_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha})}\right) \quad \text{and} \quad o_{2i} = \log\left(\frac{\pi_v(y_i^{(1)}, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha}, \boldsymbol{\zeta})}{\pi_v(y_i^{(0)}, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha}, \boldsymbol{\zeta})}\right),$$

where $\pi(v_i^{(l)}, y_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha}) = \text{pr}(R_{1i} = 1 | v_i = l, y_i = 1, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha})$. The second offset, however, is still a function of v_i and since it is a binary variable, o_{2i} can be calculated as

$$o_{2i} = \log\left(\frac{\pi(v_i^{(1)}, y_i^{(1)}, \mathbf{w}_i, \mathbf{x}_i)h^{(1)} + \pi(v_i^{(0)}, y_i^{(1)}, \mathbf{w}_i, \mathbf{x}_i)(1 - h^{(1)})}{\pi(v_i^{(1)}, y_i^{(0)}, \mathbf{w}_i, \mathbf{x}_i)h^{(0)} + \pi(v_i^{(0)}, y_i^{(0)}, \mathbf{w}_i, \mathbf{x}_i)(1 - h^{(0)})}\right)$$

where $h^{(l)} = h(v_i | Y_i = l, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\zeta})$, for $l = \{0, 1\}$. The coefficients of interest can then be estimated by solving the enlarged estimating equations system given above.

The same idea is applied for continuous V and all results are directly extended. Now,

$$f(y_i, v_i | \mathbf{w}_i, \mathbf{x}_i, R_{1i} = 1; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\zeta}) = \frac{\pi(v_i, y_i, \mathbf{w}_i, \mathbf{x}_i)h(v_i | y_i, \mathbf{w}_i, \mathbf{x}_i)}{\int \pi(v_j, y_i, \mathbf{x}_i)h(v | y_i, \mathbf{x}_i)dv} \frac{\pi_v(y_i, \mathbf{x}_i, \mathbf{w}_i)f(y_i | \mathbf{x}_i)}{\int \pi(y_j, \mathbf{x}_i, \mathbf{w}_i)_v f(y | \mathbf{x}_i)dy},$$

and for binary Y, the offset is

$$o_{2i} = log\left(\frac{\int \pi(y_i^{(1)}, v, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha})h(v | y_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\zeta})dv}{\int \pi(y_i^{(0)}, v, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha})h(v | y_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\zeta})dv}\right).$$

Finally, as in previous chapters, we can extract more information from the fully observed data by obtaining $\widetilde{\boldsymbol{S}}_1$, so that the problem is equivalent to maximizing the

pseudo-likelihood

$$L(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\zeta}) = \prod_i \frac{\pi(v_i,y_i,\mathbf{w}_i,\mathbf{x}_i;\boldsymbol{\alpha})h(v_i|y_i,\mathbf{x}_i,\mathbf{w}_i;\boldsymbol{\zeta})}{\int \pi(v,y_i,\mathbf{w}_i,\mathbf{x}_i;\boldsymbol{\alpha})h(v|y_i,\mathbf{w}_i,\mathbf{x}_i;\boldsymbol{\zeta})dv} \frac{\pi_v(y_i,\mathbf{w}_i,\mathbf{x}_i;\boldsymbol{\alpha},\boldsymbol{\zeta})f(y_i|\mathbf{x}_i;\boldsymbol{\beta})}{\int \pi_v(y,\mathbf{w}_i,\mathbf{x}_i;\boldsymbol{\alpha},\boldsymbol{\zeta})f(y|\mathbf{x}_i;\boldsymbol{\beta})dy}$$
$$\times \left(\pi(v_i,y_i,\mathbf{w}_i,\mathbf{x}_i)^{R_{1i}}\left(1-\pi(v_i,y_i,\mathbf{w}_i,\mathbf{x}_i)\right)^{1-R_{1i}}\right)^{-1}.$$

In general, the likelihood above cannot be written as $L_1(\boldsymbol{\alpha},\boldsymbol{\zeta})L_2(\boldsymbol{\beta},\boldsymbol{\alpha})$ and so the coefficients must be jointly estimated. However, if we add restrictions on the relationship between $\boldsymbol{V}$ and $\boldsymbol{Y}$ we can get unbiased estimates without modelling the conditional distribution of $\boldsymbol{V}$ given $\boldsymbol{W}$ as we now show.

Conditional independence

Suppose first that $\boldsymbol{V}$ is a surrogate variable for $\boldsymbol{X}$ that brings no additional information to the outcome once $\boldsymbol{X}$ has been provided, i.e., $\boldsymbol{Y}$ and $\boldsymbol{V}$ are conditionally independent given $\boldsymbol{X}$. Then, from equation (6.3), at the true values of $\boldsymbol{\phi} = (\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\eta})$,

$$\begin{aligned}
E(\boldsymbol{S}_0|\mathbf{x}) &= \int \left(\frac{\partial}{\partial\boldsymbol{\beta}}\log f(\mathbf{y}|\mathbf{x}) - \frac{\partial}{\partial\boldsymbol{\beta}}\log\left(\int \pi f(\mathbf{y}|\mathbf{x})h(\boldsymbol{v}|\mathbf{x})d\mathbf{y}d\boldsymbol{v}\right)\right)\pi_v f(\mathbf{y}|\mathbf{x})d\mathbf{y} \\
&= \int \left(\frac{\partial}{\partial\boldsymbol{\beta}}\log f(\mathbf{y}|\mathbf{x}) - \frac{\partial}{\partial\boldsymbol{\beta}}\log\left(\int \pi_v f(\mathbf{y}|\mathbf{x})d\mathbf{y}\right)\right)\pi_v f(\mathbf{y}|\mathbf{x})dy \\
&= \int \pi_v f'(\mathbf{y}|\mathbf{x})d\mathbf{y} - \int \pi_v f'(\mathbf{y}|\mathbf{x})d\mathbf{y} \\
&= 0.
\end{aligned}$$

That is, the $\boldsymbol{S}_0$ estimating equation is unbiased regardless of the distribution of $h(\boldsymbol{v}|\mathbf{x})$, subject to regularity conditions that allow us to interchange the order of integration and differentiation.

Correlated outcomes

Consider now the case where $\boldsymbol{V}$ and $\boldsymbol{Y}$ are not conditionally independent, but $\boldsymbol{V}$ is still observed for everyone in phase-1. Consider also that the phase-2 data was selected based on $(\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{X}_1)$. Then, the result shown in previous section, which corresponds to the estimating equation based on the conditional likelihood method, does not hold anymore and fitting a model for the conditional distribution $f(\mathbf{y}|\mathbf{x})$ ignoring the extra variable $\boldsymbol{V}$ leads to biased estimates. This will show up strongly in the simulations that follow.

However, there are still two alternative methods by which we can get unbiased estimates. We can work with a new response $\boldsymbol{Y}^* = (\boldsymbol{Y}, \boldsymbol{V})$, as done by Lee et al. (1997), or work with $\pi(\mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_{1i}; \boldsymbol{\alpha}, \boldsymbol{\zeta}) = \int \text{pr}(R_{1i} = 1|\mathbf{y}_i, \mathbf{x}_{1i}, \boldsymbol{v}_i) h(\boldsymbol{v}_i|\mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i) d\boldsymbol{v}_i$ instead. In either case we must take $\boldsymbol{V}$ into account in order to get unbiased estimates. Here we consider the second alternative and work with $\boldsymbol{Y}^* = (\boldsymbol{Y}, \boldsymbol{V})$. We assume parametric models for the two conditional distributions $f(\mathbf{y}|\mathbf{x})$ and $h(\boldsymbol{v}|\mathbf{y}, \mathbf{w}, \mathbf{x})$.

6.2.2 Weighted likelihood

We can always get unbiased estimates whether $\boldsymbol{V}$ and $\boldsymbol{Y}$ are correlated or independent given $\boldsymbol{X}$ without modelling the conditional distribution of $\boldsymbol{V}|(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{X})$, if we used a weighted method similar to the weighted approach discussed in previous chapters. Here, the weights are the inverse of the probability of being selected into phase-2, i.e., $1/\pi(\mathbf{y}, \boldsymbol{v}, \mathbf{w}, \mathbf{x})$, and the weighted loglikelihood ℓ_w is given by

$$\ell_w(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_i \frac{R}{\pi(\mathbf{y}_i, \boldsymbol{v}_i, \mathbf{w}_i, \mathbf{x}_i; \boldsymbol{\alpha})} \log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\beta}).$$

At the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$,

$$E(\boldsymbol{S}_0|\mathbf{x}) = \int\int \frac{\partial \log f(\mathbf{y}_i|\mathbf{x}_i)}{\partial \boldsymbol{\beta}} \frac{1}{\pi_i(\mathbf{y}_i, \boldsymbol{v}_i, \mathbf{w}_i, \mathbf{x}_{1i})} \text{pr}(R_i = 1, \mathbf{y}_i, \boldsymbol{v}_i|\mathbf{w}_i, \mathbf{x}_i) d\mathbf{y}_i d\boldsymbol{v}_i,$$

where $\pi_{1i} = \text{pr}(R_i = 1|\mathbf{y}_i, \boldsymbol{v}_i, \mathbf{x}_{1i}, \mathbf{w}_i)$ and

$$\begin{aligned} E(\boldsymbol{S}_0|\mathbf{x}) &= \int\int \frac{\partial \log f(\mathbf{y}_i|\mathbf{x}_i)}{\partial \boldsymbol{\beta}} \frac{1}{\pi_i(\mathbf{y}_i, \boldsymbol{v}_i, \mathbf{w}_i, \mathbf{x}_{1i})} \pi_i(\mathbf{y}_i, \boldsymbol{v}_i, \mathbf{w}_i, \mathbf{x}_{1i}) h(\boldsymbol{v}_i|\mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i) f(\mathbf{y}_i|\mathbf{x}_i) d\mathbf{y}_i d\boldsymbol{v}_i \\ &= \int\int \frac{\partial f(\mathbf{y}_i|\mathbf{x}_i)}{\partial \boldsymbol{\beta}} h(\boldsymbol{v}_i|\mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i) d\mathbf{y}_i d\boldsymbol{v}_i \\ &= \frac{\partial}{\partial \boldsymbol{\beta}} \int\int f(y_i|\mathbf{x}_i) h(\boldsymbol{v}_i|\mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i) d\mathbf{y}_i d\boldsymbol{v}_i \\ &= 0 \end{aligned}$$

for any function $h(\boldsymbol{v}_i|\mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i)$, assuming regularity conditions that allow us to interchange the order of integration and differentiation. Notice that we get unbiased estimates without taking $\mathbb{E}[\pi|\mathbf{y}, \mathbf{w}, \mathbf{x}]$, so we do not need to model $h(\boldsymbol{v}_i|\mathbf{y}_i, \mathbf{w}_i, \mathbf{x}_i)$ and the method is then not exposed to risks of model misspecification.

6.3 Simulations

Here we consider a 2-phase sampling scheme, focusing on the secondary analysis problem. The sampling scheme is defined as follows. At phase-1, complete information regarding the auxiliary variable V as well as X_1 is collected while information with respect to Y and X_2 is observed only for a sample of the phase-1 data (see figure 6.3). This sample consists of n individuals taken from each stratum defined by V (for discrete V) or V_d (for continuous V), where $V_d = 1$ if $V < c$ and 0 otherwise and c is the 15th percentile of $\boldsymbol{V}$. Interest lies in estimating the parameters $\boldsymbol{\beta}$ of $f(y|\mathbf{x}, \boldsymbol{\beta})$.

We will work with three cases:

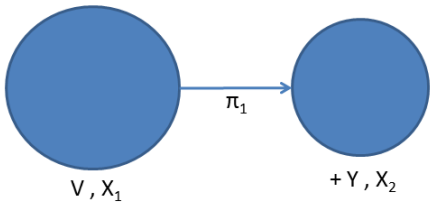


Figure 6.3: Sampling scheme for the secondary analysis problem: the response of interest $\boldsymbol{Y}$ is observed only at phase-2 while a design variable $\boldsymbol{V}$ associated to $\boldsymbol{Y}$ is fully observed.

- *Case (i): continuous V and binary Y*

  Y is binary and was generated from a Bernoulli distribution with success probability

  $$\text{logit}(\text{pr}(Y=1|\mathbf{x})) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

  The auxiliary variable now follows the linear model

  $$V = \eta_0 + \eta_1 y + \eta_2 y x_1 + \eta_3 y x_2 + \epsilon.$$

- *Case (ii): binary V and continuous Y*

  Here the response Y is continuous and we assumed the linear model

  $$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

  The auxiliary variable is discrete and generated from a Bernoulli distribution with success probability

  $$\text{logit}(\text{pr}(V=1|\mathbf{x}, y)) = \eta_0 + \eta_1 y + \eta_2 y x_1 + \eta_3 y x_2$$

- *Case (iii): continuous V and continuous Y*

  Finally, in the third case both variables are continuous and were generated from

the linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad \text{and} \quad V = \eta_0 + \eta_1 y + \eta_2 y x_1 + \eta_3 y x_2 + \epsilon_v.$$

where ϵ and ϵ_v are independent.

In all cases, X_1, X_2, ϵ and ϵ_v followed a standard normal distribution, and the total population size was N = 5000 for cases (i) and (ii) and N = 3000 for case (iii). Phase-2 data consisted of n = 300 subjects for the first two cases and n = 200 for the third case. We ran 1000 simulations for different values of $\boldsymbol{\eta}$ in order to see how the resulting bias is affected by which variables are included in the conditional model of $h(v|y, \mathbf{x})$. For example, by varying η_2 from 0 to 1, we expect to see how much bias is introduced while using the naive CML method and its impact on the MSE. Results for cases (i), (ii) and (iii) are shown in tables 6.1, 6.2 and 6.3, respectively. We use the weighted method (*wgt*) and two versions of the conditional maximum likelihood method for comparisons: *cml**, which fits a parametric model for the conditional distribution of $V|(Y, \boldsymbol{X})$, and naive *cml*, which does not take the expectation over V when calculating π into account.

Overall, the estimated standard errors (Est.SE) are close to the empirical (Emp.SE) ones and all unbiased estimating equation methods show coverage close to the nominal value. CML* is the most efficient method, producing smaller MSE when compared to the weighted and the naive CML methods. Even though in some cases the CML* and the weighted methods show similar MSEs (first row, table 6.2), CML* is usually 10% more efficient than the weighted method, and in some cases about 20% (fifth row, table 6.1) more efficient.

With respect to bias, in all cases the weighted method, which does not require modelling the conditional distribution of $V|(\boldsymbol{X}, Y)$, gives essentially unbiased estimates,

Table 6.1: Results for weighted *(wgt)*, naive CML *(cml)* and CML* *(cml*)*, for continuous Y and discrete V and 1000 datasets simualted

		Bias			Est.SE/Emp.SE			MSE $\times 10^3$			Coverage %		
$\boldsymbol{\beta} = (-3, 1, .5)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\eta} = (1,1,1,1)^T$	wgt	-0.027	0.025	0.014	1.034	0.958	1.054	54.374	51.176	40.503	0.951	0.948	0.964
	cml	-0.252	-0.083	-0.092	1.006	0.936	1.025	118.750	57.532	49.881	0.864	0.896	0.909
	cml*	-0.007	0.011	0.006	1.046	0.938	1.051	50.462	48.008	35.743	0.961	0.935	0.971
$\boldsymbol{\eta} = (1,1,0,0)^T$	wgt	-0.034	0.023	0.020	0.995	1.016	1.003	64.314	47.851	48.312	0.945	0.953	0.953
	cml	-0.308	0.019	0.013	1.026	0.998	1.018	151.072	45.428	43.769	0.825	0.958	0.958
	cml*	-0.005	0.016	0.013	1.037	1.006	1.025	57.865	45.586	42.170	0.955	0.962	0.958
$\boldsymbol{\eta} = (1,1,1,0)^T$	wgt	-0.062	0.018	0.029	0.996	0.964	0.983	64.196	49.487	51.330	0.952	0.939	0.940
	cml	-0.277	-0.123	0.025	0.996	0.959	0.992	132.903	63.066	46.811	0.851	0.877	0.952
	cml*	-0.032	0.007	0.024	1.037	0.967	1.004	55.350	44.383	44.193	0.963	0.940	0.955
$\boldsymbol{\eta} = (1,1,0,1)^T$	wgt	-0.033	0.003	0.022	1.037	0.926	0.961	54.651	55.398	48.728	0.957	0.922	0.925
	cml	-0.245	0.002	-0.121	1.049	0.933	0.986	109.035	49.826	59.551	0.875	0.937	0.882
	cml*	-0.001	-0.001	0.008	1.059	0.930	0.992	50.328	50.889	39.820	0.953	0.925	0.933
$\boldsymbol{\eta} = (1,.5,.5,.5)^T$	wgt	-0.047	0.027	0.032	0.985	0.944	0.930	61.324	52.635	54.375	0.959	0.945	0.938
	cml	-0.190	-0.080	-0.077	0.998	0.958	0.959	88.530	52.561	52.472	0.911	0.899	0.915
	cml*	-0.013	0.017	0.022	1.063	0.969	0.974	52.830	46.453	45.160	0.967	0.940	0.943

Table 6.2: Results for weighted *(wgt)*, naive CML *(cml)* and CML* *(cml*)*, for discrete Y and continuous V and 1000 datasets simulated

		Bias			Est.SE/Emp.SE			MSE $\times 10^3$			Coverage %		
$\boldsymbol{\beta} = (1, 1.5, .5)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\eta} = (-2.5, 1, 1, 1)^T$	wgt	-0.001	0.000	0.002	0.977	0.932	0.951	3.347	3.633	3.510	0.944	0.929	0.933
	cml	0.039	0.024	0.033	1.005	0.980	1.002	4.972	3.920	4.490	0.892	0.906	0.908
	cml*	0.000	-0.001	0.003	1.029	0.992	1.016	3.375	3.404	3.391	0.937	0.933	0.950
$\boldsymbol{\eta} = (-2.5, 0, 1, 1)^T$	wgt	0.004	0.003	0.000	0.978	0.979	0.980	3.831	3.355	3.560	0.947	0.952	0.945
	cml	-0.003	0.072	0.072	1.013	1.031	1.008	3.459	7.891	8.327	0.956	0.739	0.748
	cml*	0.008	0.002	-0.001	1.018	1.069	1.057	3.639	2.843	3.125	0.948	0.957	0.948
$\boldsymbol{\eta} = (-2.5, 1, 0, 1)^T$	wgt	-0.004	0.002	0.000	0.984	0.981	0.989	3.607	3.696	3.362	0.945	0.945	0.944
	cml	0.093	-0.006	0.063	1.063	1.022	1.028	0.895	3.418	7.068	0.669	0.937	0.801
	cml*	-0.006	0.004	0.002	1.072	1.045	1.073	3.381	3.353	3.062	0.960	0.945	0.956
$\boldsymbol{\eta} = (-2.5, 1, 1, 0)^T$	wgt	0.002	-0.002	-0.001	0.943	0.994	0.982	3.698	3.234	3.554	0.944	0.937	0.940
	cml	0.069	0.037	-0.005	1.009	1.058	1.007	8.302	4.234	3.474	0.785	0.903	0.940
	cml*	0.002	-0.003	-0.001	1.016	1.060	1.035	3.563	2.991	3.319	0.957	0.937	0.947
$\boldsymbol{\eta} = (-2.5, .5, .5, .5)^T$	wgt	0.003	0.000	0.001	0.971	0.968	0.984	4.121	3.759	3.787	0.940	0.929	0.944
	cml	0.074	0.053	0.057	1.031	1.035	1.021	8.885	5.605	6.313	0.753	0.840	0.817
	cml*	0.004	0.002	0.002	0.975	1.056	1.045	3.972	2.920	3.141	0.931	0.949	0.951

Table 6.3: Results for weighted *(wgt)*, naive CML *(cml)* and CML* *(cml*)*, for continuous Y and continuous V and 1000 datasets simulated

		Bias			Est.SE/Emp.SE			MSE $\times 10^3$			Coverage %		
$\boldsymbol{\beta} = (1,1,1)^T$		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\eta} = (1,1,1,1)^T$	wgt	-0.005	0.006	0.011	1.045	0.973	0.959	6.083	7.221	7.365	0.936	0.959	0.924
	cml	-0.158	-0.126 -	0.117	1.119	0.989	0.958	29.731	22.511	20.919	0.523	0.605	0.674
	cml*	0.033	0.025	0.033	1.170	1.040	1.069	5.131	6.222	6.350	0.965	0.953	0.948
$\boldsymbol{\eta} = (1,1,1,0)^T$	wgt	0.002	0.006	0.015	0.969	0.836	1.026	6.880	9.225	6.348	0.943	0.862	0.931
	cml	-0.208	-0.140	0.098	1.027	0.931	1.029	48.982	27.360	14.031	0.259	0.575	0.724
	cml*	0.022	0.026	0.004	1.013	0.994	1.085	5.808	6.462	4.341	0.943	0.943	0.971
$\boldsymbol{\eta} = (1,1,0,1)^T$	wgt	0.009	-0.007	0.002	0.924	0.951	0.965	7.426	6.914	6.742	0.944	0.944	0.933
	cml	-0.196	0.088	-0.133	0.970	0.961	0.969	44.784	12.616	24.562	0.279	0.710	0.609
	cml*	0.027	-0.006	0.020	0.974	1.023	1.036	6.397	4.837	5.637	0.944	0.944	0.939
$\boldsymbol{\eta} = (1,0,1,1)^T$	wgt	-0.004	-0.008	0.004	0.989	0.947	1.015	6.792	7.295	6.246	0.932	0.944	0.955
	cml	0.133	-0.178	-0.168	1.009	0.993	1.064	23.583	39.067	34.437	0.559	0.429	0.441
	cml*	0.004	0.011	0.018	0.982	1.013	1.052	5.583	5.943	5.638	0.944	0.972	0.955
$\boldsymbol{\eta} = (1,1,0,0)^T$	wgt	0.010	0.006	-0.013	1.024	0.995	0.980	6.520	5.794	7.117	0.940	0.900	0.980
	cml	0.088	-0.229	0.046	0.929	1.124	0.898	14.876	58.053	8.735	0.800	0.220	0.860
	cml*	0.012	0.019	-0.005	1.013	1.129	0.973	5.640	5.114	5.711	0.980	0.960	0.960

as expected from theoretical results. CML*, which assumes a parametric model for $\mathrm{h}(v|\mathbf{x},y)$, also give unbiased estimates as long all models are correctly specified. Naive CML, on the other hand, which does not take $\mathbb{E}_V[\pi|\mathbf{y},\mathbf{w},\mathbf{x}]$ and therefore does not assume a parametric model for $h(V|\boldsymbol{X},Y)$, gives biased estimates for all coefficients whose corresponding variables are correlated to V. That is, if both η_2 and η_3 are different from 0, $\hat{\beta}_1$ and $\hat{\beta}_2$ are biased and, in addition, if $\eta_1 \neq 0$, so is the intercept.

Naive CML shows even more severe bias when both variables are continuous. In this case, if either η_1, η_2 or η_3 are different from zero, the estimates are biased, resulting in poor coverage and large MSE. In fact, the resulting biases from the CML method are usually large enough to produce MSEs that are higher than the ones obtained via the weighted method, making naive CML (using $\pi(\mathbf{y},\boldsymbol{v},\mathbf{w},\mathbf{x};\boldsymbol{\alpha})$ without taking the expectation over $\boldsymbol{V}$) the least efficient method.

Model misspecification

Since the CML* method depends on the joint modelling of (V,Y), it is also of interest to see how misspecified models affect these estimates. We first fitted a misspecified linear model for the conditional distribution of $V|(\boldsymbol{X},Y)$ and later a misspecified error distribution. The results are shown in table 6.4. We compared the weighted, naive CML and CML* methods with respect to bias, standard error, MSE and coverage, for five model misspecification, assuming a discrete V and continuous Y. The true models are the same as before, i.e., we generated data from the following models

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

and

$$\text{logit}(\text{pr}(V=1|\mathbf{x},y)) = \eta_0 + \eta_1 y + \eta_2 y x_1 + \eta_3 y x_2.$$

In the first three cases shown in table 6.4 we assumed that ϵ was normally distributed and fitted a misspecified linear model for $V|(\boldsymbol{X},Y)$. The conditional distribution of $Y|\boldsymbol{X}$ was correctly modelled. For the last two cases, we generated data from the same models given above, but for ϵ following a T-distribution with 5 and 10 degrees of freedom. Then we fitted the right model for $V|(Y,\boldsymbol{X})$, but a normally distributed model for $Y|\boldsymbol{X}$. The model misspecifications considered here are not easily detectable by a simple plot of residuals and we see that they may lead to very biased estimates.

In general, we see from table 6.4 that the weighted method is unbiased because it does not assume a model for V and is robust under V-model misspecification. It also shows good estimated standard errors (Est.SE) and good coverage. The naive CML method, which does not take $E[\pi|y,\mathbf{x}]$ and therefore does not assume a parametric model for $h(V|\boldsymbol{X},Y)$, is biased in all cases, as also seen in previous results.

CML*, on the other hand, assumes a model for the conditional distribution of $V|\boldsymbol{X},Y$ and thus lead to unbiased estimates as long all models are correctly specified. In the first case where the interaction term $Y \times X_2$ is not included in the model for $V|\boldsymbol{X},Y$, the parameter β_2 is strongly biased. When both interaction terms $Y \times X_1$ and $Y \times X_2$ are missing, β_1 and β_2 are biased, and finally all estimates are biased if Y is not included in the model for $V|\boldsymbol{X},Y$. Finally, CML*, not surprisingly, can perform poorly when one of the models is misspecified.

For the last two cases, the correct model for $V|\boldsymbol{X},Y$ was fitted, apart from a misspecified error ϵ distribution assumed for the model of interest $Y|\boldsymbol{X}$. If ϵ is heavy-tailed, CML* is slightly affected, resulting in approximately unbiased estimates but with poor

Table 6.4: Results for weighted *(wgt)*, naive CML *(cml)* and CML* *(cml*)*, for a model misspecification applied to case (ii).

Method	Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Fitted $\text{logit}(\text{pr}(v=1\|y,\mathbf{x})) = -2.5+y+x_1y+x_2$ and true $\epsilon \sim N(0,1)$												
wgt	-0.002	0.002	-0.004	0.958	0.952	0.952	3.509	3.501	3.468	0.932	0.932	0.937
cml	0.039	0.026	0.026	1.004	0.969	0.994	5.045	4.162	4.066	0.888	0.900	0.912
cml*	0.003	-0.006	0.029	1.001	0.972	1.005	3.599	3.571	4.199	0.944	0.935	0.913
Fitted $\text{logit}(\text{pr}(V=1\|y,\mathbf{x})) = -2.5+y+x_1+x_2$ and true $\epsilon \sim N(0,1)$												
wgt	0.001	0.003	-0.001	0.986	0.975	0.994	3.276	3.299	3.179	0.949	0.941	0.937
cml	0.042	0.026	0.030	1.030	1.000	1.039	4.998	3.852	3.981	0.881	0.928	0.919
cml*	0.002	0.024	0.027	1.020	1.002	1.061	3.409	3.762	3.721	0.947	0.929	0.927
Fitted $\text{logit}(\text{pr}(V=1\|y,\mathbf{x})) = -2.5+x_1+x_2+x_1x_2$ and true $\epsilon \sim N(0,1)$												
wgt	0.001	0.002	0.001	0.969	0.986	0.956	3.406	3.265	3.391	0.944	0.948	0.941
cml	0.040	0.027	0.031	0.976	1.022	0.968	5.269	3.857	4.487	0.899	0.921	0.883
cml*	0.040	0.027	0.031	0.979	1.026	0.970	5.270	3.858	4.487	0.896	0.927	0.883

Continued on Next Page...

Table 6.4 – Continued

Method	Bias $\times 10^{-3}$			Est.SE/Emp.SE			MSE $\times 10^{-3}$			MSE Ratio		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Fitted $\text{logit}(\text{pr}(V=1\|y,\mathbf{x})) = -2.5+y+x_1y+x_2y$ and true $\epsilon \sim T(5)$												
wgt	0.003	-0.003	0.004	1.018	0.971	0.974	4.959	5.353	5.341	0.954	0.945	0.943
cml	0.051	0.027	0.049	1.019	0.962	1.219	9.384	6.924	8.127	0.806	0.857	0.817
cml*	0.003	-0.005	0.004	1.049	1.065	1.080	5.451	4.951	5.062	0.885	0.890	0.885
Fitted $\text{logit}(\text{pr}(V=1\|y,\mathbf{x})) = -2.5+y+x_1y+x_2y$, and true $\epsilon \sim T(10)$												
wgt	0.003	0.003	0.004	0.989	0.937	0.960	4.029	4.441	4.226	0.941	0.928	0.939
cml	0.048	0.032	0.038	1.206	1.185	1.186	6.453	5.081	5.648	0.925	0.941	0.943
cml*	0.004	0.004	0.003	1.240	1.213	1.215	3.987	3.946	4.027	0.967	0.972	0.959

coverage. It is still more efficient than the weighted method for both β_1 and β_2. When the true distribution becomes more similar to the normal distribution, the fitted and true models are more similar and the CML* method produces better estimates. They are still unbiased, with coverage close to the nominal value and lower MSE when compared to the other two approaches. It shows, however, slightly overestimated standard errors.

We performed another simulation study, but for *case (i)* discussed earlier. We worked with misspecified error distributions for the conditional distribution of $V|(\mathbf{X}, Y)$. Results are similar to the previous one obtained for *case (ii)*, with CML* more efficient than the other methods, and are shown in Table A.5.

6.4 Application to the Auckland Collaborative Birthweight Study

Here we work with the Auckland Collaborative Birthweight Study, discussed in Jiang (2004) and Jiang et al. (2006). This study was conducted between October 1995 and November 1997 and the goal was to find potential risk factors for the condition of low birthweight in newborn babies, where low birthweight was defined as birthweight equal to or below the sex-specific $10th$ percentile for gestational age in the New Zealand population (Thompson and Michell, 1994). It was designed as a case-control study, where newborns with low birthweights (cases) were termed as *small for gestational age* (SGA) and the remaining (controls) *appropriate for gestational age* (AGA). The data consists of 1714 completed interviews (844 SGA and 870 AGA).

Another response of interest was the *ponderal index* variable, which is available for only a sample of the total study population. Ponderal index, due to Rohrer (1921),

Table 6.5: Relevant covariates used in the Auckland Collaborative Birthweight Study.

Variable	Definition
occ	Mother's social-economic status (high, medium, low)
mstrat	Mother's marital status (married, defacto, never married and separated, divorced or widowed)
ethnic	Mother's ethnicity (european, pacific, maori, indian, chinese, other asian, others)
hyper	Mother's hypertension status (yes, no)
smoked	Mother's smoking status during pregnancy (yes, no)
smokemar	Mother's marijuana use (yes, no)
primi	Mother's parity status (primi, multi)
mumwt	Mother's weight (in kg)
mumht	Mother's height (in cm)
agepreg	Mother's age at this pregnancy
gest	Gestational age (in weeks)

is defined as weight in grams divided by the cube of height in centimetres. The two outcomes ponderal index and SGA are associated and so standard logistic regression is not valid.

Instead of working with the continuous ponderal index, Jiang (2004) and Jiang et al. (2006) used an indicator variable where the cut-off point was the $10th$ percentile and conducted a secondary analysis fitting two logistic regression models. The authors derived fully efficient estimating equations and compared their method to several others, such as the ordinary logistic regression, the weighted approach, among others. However, since these approaches consider only two discrete variables, they may lead to loss of information. Moreover, inference may be sensitive to the cut-off chosen and so we decided to work with the continuous variable using then all of the information available. Note that this study falls into the case (ii) discussed before: discrete phase-1 data and continuous phase-2 data. We use the weighted and the naive CML methods to estimate

the parameters of interest. Recall that none of these methods take into account a model for the ponderal index variable. The CML* method, as presented in previous section, fits a parametric model for the distribution of SGA (V-variable) and the ponderal index (Y-variable) and is also used here to analyse the data.

In our analysis we first modelled the conditional distribution $h(v|y,\mathbf{x})$, where $\boldsymbol{X}$ is a set of covariates shown in table 6.5, by fitting an ordinary logistic regression. We used the same model as Jiang et al. (2006), which includes occupational class, ethnicity, smoking, hypertension, primi, a quadratic in mother‘s weight and mother‘s height as covariates. For CML*, we assumed that the conditional distribution $Y|\boldsymbol{X}$ was normally distributed and the fitted model included marital status, ethnicity, smoke marijuana and age of pregnancy as covariates.

We fitted a logistic model for the selection probability $\pi(v,\mathbf{x})$, including sex, gest and SGA (and their interactions) as covariates, and compared four methods: an ordinary linear regression for $Y|\boldsymbol{X}$, which ignores the biased sampling, and the weighted, naive CML and CML* methods. The estimated coefficients and their standard errors are shown in table 6.6.

The ordinary linear regression and the naive CML method do not use a model for the conditional distribution $h(v|y,\mathbf{x})$ and show similar results. Both methods are known to provide biased results if the two responses Y and V are associated, which is the case here. When compared to the robust weighted approach, the naive CML method and the ordinary linear regression show, in most cases, biased estimates.

The CML* method, on the other hand, which does use a parametric model for the conditional model $h(v|y,\mathbf{x})$, shows estimates that are much more similar to the ones obtained from the weighted method than those obtained by the other two methods. For example, while estimating the parameter associated to “Ethnicity Pacifican”, naive

Table 6.6: Coefficients (std. errors) for the disproportionate growth data using an ordinary linear regression and the weighted *(wgt)*, naive CML *(cml)*, CML* *(cml*)* and CML** *(cml**)* methods.

Parameters	Linear regression	Weighted	CML	CML*	CML**
Intercept	2.368(0.144)	2.843(0.229)	2.378(0.157)	2.7705(0.169)	2.7205(0.168)
Marital status (baseline is “married")					
defacto	-0.014(0.063)	0.042(0.086)	-0.003(0.061)	0.0377(0.061)	0.0336(0.062)
never married	0.066(0.082)	0.123(0.128)	0.083(0.087)	0.1217(0.087)	0.1199(0.088)
separated, divorced or widowed	0.247(0.166)	0.638(0.528)	0.5382(0.315)	0.5288(0.324)	0.4899(0.314)
Ethnicity(baseline is “European")					
pacifican	0.137(0.062)	0.071(0.086)	0.148(0.078)	0.0782(0.079)	0.0845(0.079)
maori	0.040(0.083)	0.012(0.144)	0.027(0.088)	0.0286(0.088)	0.0339(0.087)
other asian	0.003(0.108)	-0.021(0.145)	0.024(0.114)	0.0173(0.114)	0.0300(0.115)
chinese	-0.087(0.087)	-0.098(0.066)	-0.078(0.048)	-0.0723(0.048)	-0.0633(0.048)
indian	-0.062(0.086)	-0.074(0.195)	-0.093(0.091)	-0.0255(0.091)	-0.0180(0.096)
other	-0.195(0.150)	-0.230(0.121)	-0.177(0.095)	-0.1964(0.097)	-0.1828(0.096)
smokemar	-0.111(0.091)	-0.290(0.092)	-0.119(0.056)	-0.2297(0.068)	-0.2009(0.065)
agepreg	0.007(0.004)	-0.004(0.007)	0.007(0.005)	-0.0019(0.005)	-0.0003(0.005)

CML gives an estimate that is about twice larger than the one given by the weighted method (0.071 against 0.148), but the estimate obtained via the CML* method is nearly the same as the weighted one. Differences are observed for the "Marital status defacto", "never married", "smokemar" and for the intercept.

CML* shows a much smaller standard error than the weighted method. This reduction, for example, can be as small as 10%, when "Ethnicity Pacifican" is estimated, or even more than 50%, as obtained while estimating the parameter associated with the "Ethnicity Indian" variable.

As discussed before in section 3.4, the flexibility of the conditional maximum likelihood method allows us to use more information from the dataset by including more variables into the selection model. So far only the discretized version (SGA) of the newborns birthweight has been used for analysis. And since the birthweight has been fully observed at phase-1, we can use this extra information to hopefully get even better estimates. We thus fitted the old selection model plus birthweights and its interaction with SGA and sex, and kept the same models for $Y|\boldsymbol{X}$ and $V|(\boldsymbol{X}, Y)$. This new method is denoted by CML**.

However, even though CML** brings in extra information regarding the continuous variable birthweight, this was not sufficiently large to improve our estimates. We see that all standard errors and estimates are nearly the same whether this extra information was used or not.

6.5 Summary

In this chapter we generalized all 2-phase designs previously discussed in this thesis, which also covers a wide range of problems (see section 6.1). We also generalized the

method discussed in previous chapters to cope with this more general design and the new set of estimating equations are more general and can be applied to a much broader class of problems.

More specifically, we notice that, if there is an extra variable which is uncorrelated to the response of interest, all different sampling schemes were shown to be mathematically equivalent under the CML approach. Moreover, unlike other methods, CML does not require fitting a parametric model for this extra variable, making it robust against model misspecification. If, however, this extra variable is correlated to the response, modelling the selection probability naively by only using variables fully observed at phase-1 will sometimes lead to strongly biased results, as seen from our simulations. Instead, we should use the expected value of this selection probability given a set of variables fully observed at phase-1, which requires assuming a parametric model for this extra variable in order to get unbiased estimates. This increases the risks of model misspecifications and in cases where it is only slightly more efficient than the more robust weighted method, the latter one should be preferred.

7

Semiparametric Efficiency

In this chapter we investigate the efficiency of the CML+$\widetilde{\boldsymbol{S}}$ method with discrete and continuous responses. As in previous chapters, our interest centres on the 2-phase design in which the response and possibly some covariates are fully observed at phase-1 and the remaining information is collected at phase-2. We start with the discrete case, showing the equivalence between the new approach with the Scott and Wild's method (Scott and Wild, 1997), under a number of scenarios where it is known to be fully efficient. Next we derive the semiparametric lower bound for the variance and investigate closeness of the variance of CML+$\widetilde{\boldsymbol{S}}$ to the lower bound in several scenarios by simulation.

7.1 Binary response

An important property of the conditional likelihood method is the fact that it can achieve full efficiency if some conditions are satisfied. This result makes the method very appealing because it is easy to use and implement, unlike most other equally fully

efficient methods discussed in chapter 1.

To show that the CML+$\widetilde{\boldsymbol{S}}$ method is semiparametric fully efficient, we will use the result presented by Scott and Wild (1997), which is known to be fully efficient when the phase-1 variables are discrete (Lee and Hirose, 2010).

Let $\boldsymbol{Y}$ be a discrete variable with I levels and $\boldsymbol{X}_1$ be also discrete taking values $\{\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1J}\}$. Let also N_{il} and n_{il} denote the number of units with $\{\boldsymbol{Y} = i, \boldsymbol{X}_1 = \mathbf{X}_{1l}\}$ in the population and sample, respectively. Scott and Wild (1997) proves the following theorem, which gives a method for calculating a semiparametric efficient estimator for $\boldsymbol{\beta}$.

Theorem 7.1 (Scott and Wild (1997))**.** *Under supplemented case-control sampling, the maximum likelihood estimate satisfies*

$$\frac{\partial L^*}{\partial \theta} \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \sum_i \sum_j \log \mathrm{pr}^* (\boldsymbol{Y} = i | \boldsymbol{x}_{ij}) = 0,$$

where

$$\mathrm{pr}^* (\boldsymbol{Y} = i | \boldsymbol{x}) = \frac{\mu_i P_i(\boldsymbol{x}; \boldsymbol{\beta})}{\sum_{l=1}^{I} \mu_l P_l(\boldsymbol{x}; \boldsymbol{\beta})},$$

$$\mu_i = \frac{n_i - \xi_i}{N_i - \xi_i},$$

$$\xi_i = n_i - \sum_{l=1}^{I} \sum_{j=1}^{n_l} \mathrm{pr}^* (\boldsymbol{Y} = i | \boldsymbol{x}_{ij})$$

for $i = 1, \ldots, I$.

In the special case where the response is binary, using our notation Theorem 7.1 can be rewritten as follows: Since f_c is a function of $\boldsymbol{\omega} = \log(\pi_{1l}/\pi_{0l})$, the maximum

likelihood of $\boldsymbol{\beta}$ satisfies

$$\sum_{i=1}^{N} R_i \frac{\partial \log f_c}{\partial \boldsymbol{\beta}} = 0 \quad \text{and} \quad \sum_{i=1}^{N} R_i \frac{\partial \log f_c}{\partial \boldsymbol{\omega}} = \xi \tag{7.1}$$

subject to the constraint

$$\log\left(\frac{n_{1l}-\xi_l}{N_{1l}-\xi_l}\right) - \log\left(\frac{n_{0l}+\xi_l}{N_{0l}+\xi_l}\right) = \omega_l \tag{7.2}$$

7.1.1 Equivalence to Scott and Wild

We want to see if the solutions of the equation (2.16) are the same as the ones stated above, for a specific value of λ yet to be found. We firstly assume that we have a saturated model for the selection probability π_1 so we can choose the parameter $\boldsymbol{\alpha}_1$ to be any one-to-one function of the cells $\{\pi_{jl} : j = 1, \ldots, J; l = 1, \ldots, L\}$. Since in our case $J = 2$, let

$$\rho_{0l} = \log \pi_{0l} \Rightarrow \pi_{0l} = e^{\rho_{0l}} \quad \text{and} \quad \rho_{1l} = \log\left(\frac{\pi_{1l}}{\pi_{0l}}\right) \Rightarrow \pi_{1l} = e^{\rho_{1l}+\rho_{0l}},$$

so that

$$\frac{\partial \pi_{1l}}{\partial \rho_{0l}} = \frac{\partial \pi_{1l}}{\partial \rho_{1l}} = \pi_{1l}, \quad \frac{\partial \pi_{0l}}{\partial \rho_{0l}} = \pi_{0l} \quad \text{and} \quad \frac{\partial \pi_{0l}}{\partial \rho_{1l}} = 0.$$

Setting $\boldsymbol{S}_\lambda = 0$, we have that

$$\boldsymbol{S} + \lambda \widetilde{\boldsymbol{S}} = \sum_{j=1}^{N_i} \begin{pmatrix} \boldsymbol{S}_0 \\ \boldsymbol{S}_1^{(0)} + \lambda \widetilde{\boldsymbol{S}}_1^{(0)} + \boldsymbol{S}_1^{(1)} + \lambda \widetilde{\boldsymbol{S}}_1^{(1)} \end{pmatrix} = \boldsymbol{0},$$

where $\boldsymbol{S}_1^{(i)}$ stands for $\boldsymbol{S}_1$ (see equation (2.8)) for $Y = i$, which results in three estimating equations, one for each parameter: $\boldsymbol{\beta}$, ρ_{1l} and ρ_{0l}.

For $\boldsymbol{\beta}$, we have that

$$\sum_{i=1}^{N_1} \boldsymbol{S}_0 = \sum_{l=1}^{N_1} R_{1l} \frac{\partial \log f_c}{\partial \boldsymbol{\beta}} = 0;$$

for ρ_{1j},

$$\sum_{i=1}^{N_1} \left[\left(\frac{R_{1lj} - \pi_{1l}}{\pi_{1l}(1-\pi_{1l})} \right) \frac{\partial \pi_{1l}}{\partial \rho_{1l}} + \lambda R_{1lj} \frac{\partial \log f_c}{\partial \rho_{1l}} \right] = 0,$$

which implies

$$\sum_{i=1}^{N_1} R_{1l} \frac{\partial \log f_c}{\partial \rho_{1l}} = -\frac{1}{\lambda} \left(\frac{n_{1l} - N_{1l}\pi_{1l}}{1-\pi_{1l}} \right);$$

and for ρ_{0j},

$$\sum_{i=1}^{N_1} \left[\left(\frac{R_{1l} - \pi_{0l}}{\pi_{0l}(1-\pi_{0l})} \right) \frac{\partial \pi_{0l}}{\partial \rho_{0l}} + \lambda R_{1l} \left(-\frac{\partial \log f_c}{\partial \rho_{0l}} \right) + \left(\frac{R_{1l} - \pi_{1l}}{\pi_{1j}(1-\pi_{1l})} \right) \frac{\partial \pi_{1l}}{\partial \rho_{0l}} + \lambda R_{1l} \frac{\partial \log f_c}{\partial \rho_{0l}} \right] = 0,$$

implying

$$\left(\frac{n_{0l} - N_{0l}\pi_{0l}}{1-\pi_{0l}} \right) = -\left(\frac{n_{1l} - N_{1l}\pi_{1l}}{1-\pi_{1l}} \right).$$

If $\lambda = -1$, we can set $\xi_l = (n_{1l} - N_{1l}\pi_{1l})/(1-\pi_{1l})$ so that

$$\pi_{1l} = \frac{n_{1l} - \xi_l}{N_{1l} - \xi_l} \quad \text{and} \quad \pi_{0l} = \frac{n_{0l} + \xi_l}{N_{0l} + \xi_l}$$

and finally,

$$\rho_{1l} = \log\left(\frac{\pi_{1l}}{\pi_{0l}} \right) = \log\left(\frac{n_{1l} - \xi_l}{N_{1l} - \xi_l} \right) - \log\left(\frac{n_{0l} + \xi_l}{N_{0l} + \xi_l} \right).$$

7.1.2 Additional sample

The CML+$\widetilde{\boldsymbol{S}}$ method also works under a slightly different sampling scheme, namely that used by Zhou et al. (2002) and Song et al. (2009). The authors use an additional sample of size n_0 randomly selected among all phase-1 data so that the phase-2 data consists of n individuals, the outcome-dependent subsample sampled based on Y, plus

the additional sample n_0. We show next that both methods, CML+$\widetilde{\boldsymbol{S}}$ and the Scott and Wild's method, are still equivalent, despite the changes made on the design.

Under this new sampling scheme, the complete or full likelihood is

$$L = \left(\prod_i \prod_j (p_{ij}\delta_j)^{n_{ij}} \prod_i \left(\sum_l p_{il}\delta_l \right)^{N_i - n_{i+}} \right)^{I_A} \left(\prod_i \prod_j (p_{ij}\delta_j)^{n_{ij}} \right)^{I_B},$$

where A is the set of units with $\pi = \pi(y;\boldsymbol{\alpha})$ with indicator function I_A and B is the set of units with $\pi = c$, where c is a constant, with indicator function I_B.

For the profile likelihood approach, we have to maximize the above likelihood with respect to δ_j with the constraint $\sum_j \delta_j = 1$, where $j = 1, \ldots, N_t$ and N_t is the total number of individuals. The loglikelihood is given by

$$\ell = \left(\sum_i \sum_j n_{ij} \log p_{ij} + \sum_j n_{+j} \log \delta_j + \sum_i (N_i - n_{i+}) \log \left(\sum_l p_{il}\delta_l \right) \right) I_A + \left(\sum_i \sum_j n_{ij} \log p_{ij} + \sum_j n_{+j} \log \delta_j \right) I_B.$$

By introducing a Lagrange multiplier η to take care of the constraint and after multiplying by δ_j and summing over j, we have that

$$\left(n_{+i} + \sum_i (N_i - n_{i+}) \right) I_A + (n_{i+}) I_B + \eta = 0 \quad \rightarrow \quad \eta = -(NI_A + nI_B).$$

Then, replacing η back into $\partial l / \partial \delta_j$, we get

$$\frac{\partial l}{\partial \delta_j} = \left(\frac{n_{+j}}{\delta_j} + \sum_i \left(\frac{(N_i - n_{i+})}{\sum_l p_{il}\delta_j} p_{ij} \right) \right) I_A + \left(\frac{n_{+j}}{\delta_j} \right) I_B - (NI_A + nI_B) = 0$$

which implies

$$\delta_j = \frac{n_{+j} I_A + n_{+j} I_B}{\left(N - \sum_i (N_i - n_{i+}) \frac{p_{ij}}{\sum_l p_{il}\delta_j} \right) I_A + nI_B}.$$

Note that δ_j reduces to n_{+j}/N if $i \in B$. We can rewrite δ_j as

$$\delta_j = \frac{n_{+j}I_A + n_{+j}I_B}{N\sum_i \mu_i p_{ij}}, \quad \text{where} \quad \mu_i = 1 - \frac{1}{N}\frac{N_i - n_i}{\sum_l p_{il}\delta_l}.$$

Now, $\mu_i = 1/N$, the probability of being selected through the simple random sampling part, if $i \in B$.

Replacing δ_j back into the definition of μ_i, we have that for all $i \in A$,

$$\mu_i = \frac{n_i - \xi_i}{N_i - \xi_i},$$

where

$$\xi = n_i - \sum_l p^*_{il} n_{+l} \quad \text{and} \quad p^*_{ij} = \frac{\mu_i p_{ij}}{\sum_i \mu_i p_{ij}}.$$

Finally, ξ_i can also be written as

$$\xi_i = \frac{\partial}{\partial \mu} \sum_i \sum_j \log p^*_{ij},$$

for all $i \in A$. If $i \in B$, p^*_{ij} is not a function of μ_i and so $\xi_i = 0$. Note that the conditions stated on theorem 7.1 are the same for all $i \in A$ and for $i \in B$, $\xi_i = 0$ and $\omega = 0$. Therefore, in order to show equivalence between the Scott and Wild (1997) and the CML+$\widetilde{\boldsymbol{S}}$ methods we have only to check the expressions from CML+$\widetilde{\boldsymbol{S}}$ when $i \in B$. And since

$$\pi_i = n_o/N_t \quad \rightarrow \quad f_c = f \quad \rightarrow \quad \boldsymbol{\xi}_i = \frac{\partial \log f_c}{\partial \boldsymbol{\omega}} = 0.$$

and

$$\boldsymbol{\omega} = \log\left(\frac{\pi_{1l}}{\pi_{0l}}\right) = \log(\mu_1/\mu_0) = 0,$$

the CML+$\widetilde{\boldsymbol{S}}$ method satisfies the conditions of theorem 7.1 and both methods are

equivalent.

7.2 Lower bound for the variance

In this section we will calculate the lower bound for the variance of any semiparametric estimator of $\boldsymbol{\beta}$, where a response $\boldsymbol{Y}$ has been fully observed while some covariate $\boldsymbol{X}$ has been partially measured. We assume that the selection probability depends only on the response, so the data is missing at random and the methods studied before can be applied. That is, we are interested in the likelihood

$$\prod_{i=1}^{N} (\pi_i(y) f(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}) g(\mathbf{x}))^{R_i} \prod_{i=1}^{N} \left((1-\pi_i(\mathbf{y})) \int f(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}) g(\mathbf{x}) d\mathbf{x} \right)^{1-R_i}.$$

Note that, by the missing at random assumption, π, the selection probability, does not depend on $\boldsymbol{\beta}$ and inference concerning $\boldsymbol{\beta}$ is independent of whether π is completely known or unknown. Hence, the lower bound for the variance must be same in both situations. A representation of this lower bound has already been obtained by Robins et al. (1995) and Zhang and Rockette (2006) in terms of an integral equation. Here we work with the Zhang and Rockette (2006) representation, solving the integral equation and calculating the efficient information for the specific sampling design described above.

First, however, it is important to fix some notation that is going to be used throughout this and the following sections. In order to make it simpler and more concise, we will consider the case where $\boldsymbol{X}$ is univariate and we will write

- f, for the conditional distribution $f(\mathbf{y}|\mathbf{x};\boldsymbol{\beta})$ and f', for $f(\mathbf{y}|\mathbf{x}';\boldsymbol{\beta})$;
- g (or g') and G, for the marginal density $g(\mathbf{x})$ (or $g(\mathbf{x}')$) and distribution $G(\mathbf{x})$ of

X and G_0, for the true distribution G;

- $\boldsymbol{S}_0$, for the score vector with respect $\boldsymbol{\beta}$, evaluated at the true distribution G and true parameters $\boldsymbol{\beta}$. That is,

$$\boldsymbol{S}_0 = R\frac{\partial}{\partial \boldsymbol{\beta}} \log(L(\boldsymbol{\beta}, G_0)) + (1-R)E\left(\frac{\partial}{\partial \boldsymbol{\beta}} \log(L(\boldsymbol{\beta}, G_0))\Big|\mathbf{y}\right)$$

- $\boldsymbol{S}_1$, for the score vector with respect to nuisance parameter g;
- $\mathcal{X}$, for the space in which $\boldsymbol{X}$ is defined;
- $\mathcal{B}$, for the parameter space.

7.2.1 Efficient score

In order to calculate the lower bound for the variance, we need first to define and derive the nuisance tangent space.

Definition 7.1 (Tsiatis (2006))**.** *The nuisance tangent space for a semiparametric model, denoted* Λ*, is defined as the mean-square closure of parametric submodel nuisance tangent spaces, where a parametric submodel nuisance tangent space is the set of elements*

$$\left\{\boldsymbol{M}^{q\times r} S_{\boldsymbol{\alpha}}^{r\times 1}\right\},$$

$\boldsymbol{S}_{\boldsymbol{\alpha}}$ *is the score vector for the nuisance parameter* Λ *for some parametric submodel, and* $\boldsymbol{M}^{q\times r}$ *is a conformable matrix with q-rows. Specifically, the mean-square closure of the spaces above is defined as the space* $\Lambda \subset \mathcal{H}$*, where* $\Lambda = \{h^{q\times 1} \in \mathcal{H}$ *such that* $E\left(h^T h\right) < \infty$ *and there exists a sequence* $\boldsymbol{M}_j \boldsymbol{S}_{\boldsymbol{\alpha}j}$ *such that*

$$||h - \boldsymbol{M}_j \boldsymbol{S}_{\boldsymbol{\alpha}j}||^2 \to 0 \quad as\ j \to \infty,$$

for a sequence of parametric submodels indexed by j}, where $||h||^2 = E\left(h^T h\right)$.

Consider a Hilbert space $\mathcal{H}$ generated by all measurable functions of $\boldsymbol{X}$ with mean zero and finite variance, and with inner product $\langle g_1, g_2 \rangle = E(g_1' g_2)$. The infinite-dimensional linear subspace $\Lambda \subset \mathcal{H}$ generated by spanning the nuisance score vector $\boldsymbol{S}_1$ can be constructed as follows. Define a measurable and bounded function $h : \mathcal{X} \to \mathbb{R}$ under G_0 and let

$$\frac{dG_t}{dG_0} = 1 + t\left(h - \int h dG_0(\mathbf{x})\right),$$

with $|t|$ sufficiently small so that $1 + t(h - \int h dG_0(\mathbf{x})) \geq 0$, for all $\mathbf{x}$. Moreover, we have that

$$\int dG_t = \int dG_0 + t \int \left(h - \int h dG_0\right) dG_0 = 1$$

so that G_t is also a probability distribution function on $\mathcal{X}$ and equivalent to G_0 when $t = 0$. Replacing G_t back into the loglikelihood, we have that

$$\begin{aligned}
\ell(\boldsymbol{\beta}, G_t) &= \log f + R \log dG_t + (1 - R) \log \left(\int f dG_t \right) \\
&= \log f + R \log \left[\left(1 + t \left(h - \int h dG_0 \right) \right) dG_0 \right] \\
&+ (1 - R) \log \left[\int f \left(1 + t \left(h - \int h dG_0 \right) \right) dG_0 \right],
\end{aligned}$$

and since G_0 is the true distribution of $\boldsymbol{X}$, the map $t \to \ell(\boldsymbol{\beta}, G_t)$ will be maximized at $t = 0$. Taking the derivative of $\ell(\boldsymbol{\beta}, G_t)$ with respect to t and evaluating at $t = 0$, we have that

$$\boldsymbol{S}_h = Rh + (1 - R) \frac{\int h f dG_0}{\int f dG_0} - \int h dG_0 = B_0 h - \int h dG_0,$$

where $\boldsymbol{S}_h$ is a score function and

$$B_0 h = Rh + (1 - R) \frac{\int h f dG_0}{\int f dG_0}.$$

Note that, for any function $h \in L_2(G)$ with mean zero and bounded second moment, from the L2-completeness (Tsiatis, 2006, pp. 14), the space generated by $B_0 h$ is a Hilbert space with covariance inner product $\langle h_1, h_2 \rangle = E(h_1 h_2)$. If we choose $\mathbf{M}$ to be the identity matrix, $B_0 h$ is an element of Λ. Finally, since any element of $B_0 h$ can be taken as a limit of mean zero functions of h, all elements of $B_0 h$ are either elements of a parametric submodel nuisance tangent space or a limit of such elements. That is,

$$\Lambda \supseteq \{B_0 h : h \in L_2 \text{ with } \int h dG_0 = 0\},$$

For the inverse inequality we have to show that any element of Λ can be written as $B_0 h$, with $h \in L_2$ and first moment equal to zero. Assuming a parametric model $G(\mathbf{x}; \boldsymbol{\alpha})$ for the covariates $\boldsymbol{X}$, the score with respect to $\boldsymbol{\alpha}$ is

$$R \boldsymbol{S}_{\boldsymbol{\alpha}} + (1 - R) \frac{\int \boldsymbol{S}_{\boldsymbol{\alpha}} f dG_0}{\int f dG_0},$$

where $\boldsymbol{S}_{\boldsymbol{\alpha}} = \partial \log g(\mathbf{x}; \boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$. Note that this score can be written as $B_0 h$, with $h = \boldsymbol{S}_{\boldsymbol{\alpha}}$, which has mean zero under the true distribution of G. By the finite second moment assumption, we have that

$$\Lambda \subseteq \{B_0 h : h \in L_2 \text{ with } \int h dG_0 = 0\}.$$

Now that we have derived the nuisance tangent space, we can calculate the efficient score and, consequently, the semiparametric efficiency bound.

Definition 7.2 (Tsiatis (2006))**.** *The efficient score is defined as the residual of the score vector with respect to the parameter of interest after projecting it onto the nuisance*

tangent space Λ*, i.e.,*

$$\boldsymbol{S}_e = \boldsymbol{S}_0 - \prod(\boldsymbol{S}_0|\Lambda).$$

Definition 7.2 says that the efficient score is the projection of the score vector $\boldsymbol{S}_0$ onto Λ^T, the orthogonal complement of Λ. That is, for any $h \in \Lambda$,

$$\langle \boldsymbol{S}_e, h \rangle = 0,$$

under $\mathbb{P}$. An intuitive argument is given by Nan et al. (2004). The authors say that "when G_0 is unknown, information about $\boldsymbol{\beta}$ can only come from that component of $\boldsymbol{S}_0$ that is statistically independent of variability in the data controlled by the nuisance parameter. This component is $\boldsymbol{S}_e$." Once $\boldsymbol{S}_e$ has been obtained, we can use the following theorem to get the efficiency bound.

Theorem 7.2 (Tsiatis (2006))**.** *The semiparametric efficiency bound is equal to the inverse of the variance matrix of the semiparametric efficient score, i.e.,*

$$E^{-1}\left(\boldsymbol{S}_e \boldsymbol{S}_e^T\right).$$

In order to obtain S_e we have to find its projection onto Λ. First, however, let B_0^* be the adjoint of B_0 and let $\mathcal{Z}$ be the space where $(\boldsymbol{X}, \boldsymbol{Y}, R)$ are defined, $h_1 : \mathcal{Z} \to \mathbb{R}$

and $h_2 : \mathcal{X} \to \mathbb{R}$. Then,

$$\begin{aligned}
\langle h_1, B_0 h_2 \rangle &= \int h_1 \pi h_2 f g d\mathbf{y} d\mathbf{x} + \int h_1^0 (1-\pi) \left(\int \frac{h_2 f g}{f_y} d\mathbf{x}' \right) f g d\mathbf{y} d\mathbf{x} \\
&= \int h_2 \pi h_1 f g d\mathbf{y} d\mathbf{x} + \int \left(\int \frac{h_1^0 (1-\pi) f g}{f_y} d\mathbf{x}' \right) h_2 f g d\mathbf{y} d\mathbf{x} \\
&= \int \left(\int \pi h_1 f d\mathbf{y} \right) h_2 g d\mathbf{x} + \int \left(\int \left(\int \frac{(1-\pi) h_1^0 f g}{f_y} d\mathbf{x}' \right) f d\mathbf{y} \right) h_2 g d\mathbf{x} \\
&= \langle B_0^* h_1, h_2 \rangle ,
\end{aligned}$$

where

$$\begin{aligned}
B_0^* h_1 &= \int \pi h_1 f d\mathbf{y} + \int \left(\int \frac{(1-\pi) h_1^0 f g}{f_y} d\mathbf{x} \right) f d\mathbf{y} \\
&= E\left(R h_1 | \boldsymbol{X} = \mathbf{x}\right) + E\left(E\left((1-R) h_1 | \, \boldsymbol{Y} = \mathbf{y}\right) | \boldsymbol{X} = \mathbf{x}\right).
\end{aligned}$$

Finally, since

$$B_0 1 = R + (1-R) \frac{\int f dG_0}{\int f dG_0} = 1,$$

we have that $\langle B_0^* B_0 h, 1 \rangle = \langle B_0^* h, 1 \rangle = \langle h, B_0 1 \rangle = \langle h, 1 \rangle$. That is, for $h \in L_2(G)$, $B_0^* B_0 h$ is mean preserving. In addition, assuming for now that $B_0^* B_0$ has an inverse and noting that $\langle B_0 B_0 h, 1 \rangle = \langle B_0 h, B_0^* 1 \rangle = \langle B_0 h, 1 \rangle$, its projection exists and is unique, and given by $B_0 \left(B_0^* B_0\right)^{-1} B_0^* S_0$. The efficient score is then equal to

$$\boldsymbol{S}_e = \boldsymbol{S}_0 - B_0 \left(B_0^* B_0\right)^{-1} B_0^* \boldsymbol{S}_0. \tag{7.3}$$

Zhang and Rockette conclude that $B_0^* B_0$ is invertible by showing that it is positive-definite, but this result does not hold for infinite dimensional matrix (consider, for example, the case where the eigenvalues of $B_0^* B_0$ converge to 0). We will use instead a version of the LAX-Milgram theorem (Kress, 1999, pp. 242), which requires the

operator to be strictly coercive.

Definition 7.3 (Coercive)**.** *A bounded linear operator $A : \mathcal{H} \to \mathcal{H}$ in a Hilbert space $\mathcal{H}$ is called strictly coercive if there exists a constant $c > 0$ such that*

$$\langle Ah, h \rangle \;\geq\; c||h||^2 .$$

Theorem 7.3 (LAX-Milgram)**.** *In a Hilbert space $\mathcal{H}$ a strictly coercive bounded linear operator $A : \mathcal{H} \to \mathcal{H}$ has a bounded inverse $A^{-1} : \mathcal{H} \to \mathcal{H}$.*

First, note that $B_0^* B_0$ is linear and continuous on $\boldsymbol{X}$ and so it is a bounded operator. Now, in order to show that the operator is also strictly coercive, we have that

$$\begin{aligned}
\langle B_0^* B_0 h, h \rangle &= \int \left(\int \pi h f d\mathbf{y} \right) h g d\mathbf{x} + \int \left(\int \left(\int \frac{(1-\pi) h f g}{f_y} d\mathbf{x}' \right) f d\mathbf{y} \right) h g d\mathbf{x} \\
&= \int \pi h^2 f g d\mathbf{y} d\mathbf{x} + \int \left(\int \frac{h f g}{f_y} d\mathbf{x}' \right) \left(\int \frac{h f g}{f_y} d\mathbf{x} \right) (1-\pi) f_y d\mathbf{y} \\
&\geq \int \pi h^2 f g d\mathbf{y} d\mathbf{x} \\
&\geq c||h||^2 ,
\end{aligned}$$

where $c = \min\{\pi(\mathbf{y}) : \mathbf{y} \in \mathcal{R}\} > 0$ and satisfied by assumption. Thus, $B_0^* B_0$ is strictly coercive and by the LAX-Milgram theorem it has a bounded inverse $(B_0^* B_0)^{-1}$.

Since $B^* B$ has an inverse, we can calculate the efficient score given by equation

(7.3). By noting that

$$\begin{aligned} B_0^* B_0 h &= \int h\pi f d\mathbf{y} + \int \left(\int \frac{h(1-\pi) f g}{f_y} d\mathbf{x} \right) f d\mathbf{y} \\ &= h \int \pi f d\mathbf{y} + \int \left(\int \frac{h'(1-\pi) f' g'}{f'_y} d\mathbf{x}' \right) f d\mathbf{y} \\ &= h \int \pi f d\mathbf{y} + \int \left(\int \frac{(1-\pi) f' f}{f'_y} d\mathbf{y} \right) h' g' d\mathbf{x}' \\ &= h\phi(\mathbf{x}) + \int K(\mathbf{x}, \mathbf{x}') h' g' d\mathbf{x}'. \end{aligned}$$

and that

$$\begin{aligned} B_0^* \boldsymbol{S}_0 &= E(R\boldsymbol{S}_0 | \boldsymbol{X} = \mathbf{x}) + E\left(E\left((1-R)\boldsymbol{S}_0 | \, \boldsymbol{Y} = \mathbf{y} \right) | \boldsymbol{X} = \mathbf{x} \right) \\ &= \int \pi \left(\frac{\partial}{\partial \boldsymbol{\beta}} \log L \right) f d\mathbf{y} + \int \left(\int (1-\pi) \left(\frac{\partial}{\partial \boldsymbol{\beta}} \log L \right) \frac{fg}{f_y} d\mathbf{x} \right) f d\mathbf{y} \\ &= \psi(\mathbf{x}), \end{aligned}$$

we see that the function h can be obtained by solving the integral equation $B_0^* B_0 h = B_0^* \boldsymbol{S}_0$, which can be written as

$$h(\mathbf{x})\phi(\mathbf{x}) + \int K(\mathbf{x}, \mathbf{x}') h(\mathbf{x}') g(\mathbf{x}') d\mathbf{x}' = \psi(\mathbf{x})$$

and is of the form of a Fredholm equation of the second kind. The Nystrom routine is a simple way to solve such equations and works as follows. We first approximate the integral by any quadrature method as

$$h(\mathbf{x})\phi(\mathbf{x}) + \sum_{i=1}^{N_p} K(\mathbf{x}, \mathbf{x}_i) h(\mathbf{x}_i) g(\mathbf{x}_i) \mathbf{w}_i = \psi(\mathbf{x}), \tag{7.4}$$

where $\mathbf{x}_i$s and N_p are the points and total number of points of the chosen quadrature method, repsectively, with weights $\mathbf{w}_i$s. Then, by replacing $\mathbf{x}$ by $\mathbf{x}_i$ in equation (7.4),

we have that

$$\psi(\mathbf{x}_i) = h(\mathbf{x}_i)\phi(\mathbf{x}_i) + \sum_{i=1}^{N_p} K(\mathbf{x}, \mathbf{x}_i)h(\mathbf{x}_i)g(\mathbf{x}_i)\mathbf{w}_i$$

where $\psi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_i)$ are the approximations for $\psi(\mathbf{x})$ and $\phi(\mathbf{x})$, respectively, and $h(\mathbf{x}_i)$ is the solution of the system

$$\left(I_{N_p \times N_p}\phi(\mathbf{x}_i) + \tilde{K}(\mathbf{x}_i, \mathbf{x}_i)\right) h(\mathbf{x}_i) = \psi(\mathbf{x}_i), \quad i = 1, \cdots, N_p,$$

where $I_{N_p \times N_p}$ is a $N_p \times N_p$ identity matrix and $\tilde{K}(\mathbf{x}_i, \mathbf{x}_i) = K(\mathbf{x}_i, \mathbf{x}_i)g(\mathbf{x}_i)\mathbf{w}_i$. Once $h(\mathbf{x}_i)$ is obtained, we replace it back into equation (7.4) and $h(\mathbf{x})$ can then be calculated for any value of $\mathbf{x}$. Finally, the efficient information will be given by

$$\boldsymbol{\mathcal{I}}_e = \int ||\boldsymbol{S}_0 - Bh|| \, d\mathbb{P} \tag{7.5}$$

and the lower bound for the variance by $\boldsymbol{\mathcal{I}}_e^{-1}$.

Extra covariate

We are now interested in the case where the response $\boldsymbol{Y}$ and a covariate $\boldsymbol{X}_1$ are known at phase-1 and an expensive covariate $\boldsymbol{X}_2$ is observed only at phase-2. We are going to work with the likelihood

$$L(\boldsymbol{\beta}, g) = \prod_{i=1}^{N} \left(f(\mathbf{y}_i|\mathbf{x}_i;\beta)g_2(\mathbf{x}_{2i}|\mathbf{x}_{1i})g_1(\mathbf{x}_{1i})\right)^{R_i} \prod_{i=1}^{N} \left(\int f(\mathbf{y}_i|\mathbf{x}_i\beta)g_1(\mathbf{x}_{1i})g_2(\mathbf{x}_2|\mathbf{x}_{1i})d\mathbf{x}_2\right)^{1-R_i},$$

where g_1 (or G_1) and g_2 (or G_2) are the marginal and conditional pdf (or cdf) of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2|\boldsymbol{X}_1$, and $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$. Note that, if g_2 was known, the efficient estimator of $\boldsymbol{\beta}$ would be obtained by solving $\partial \log L(\boldsymbol{\beta}, g_0)/\partial\boldsymbol{\beta} = 0$.

The nuisance tangent space can be obtained using the same ideas as in the previous

section. That is, by defining a measurable and bounded function $h : \mathcal{X} \to \mathbb{R}$ under G_0, where $\mathcal{X}$ is now the space where $\boldsymbol{X}_2|\boldsymbol{X}_1$ is defined and G_0 is the conditional distribution evaluated at the true parameters, we have that

$$\Lambda = \{B_0 h : h \in L_2 \text{ with } \int h dG_0 = 0\}$$

where

$$B_0 h = Rh + (1-R)\frac{\int h f g_2 d\mathbf{x}_2}{\int f g_2 d\mathbf{x}_2} = Rh + (1-R)E\left(h|\mathbf{y}, \mathbf{x}_1\right),$$

which has the same structure as in the previous section. Similarly, by defining $h_1 : \mathcal{X} \to \mathbb{R}$ and $h_2 : \mathcal{X}_1 \to \mathbb{R}$, where $\mathcal{X}_1$ is the space where $\boldsymbol{X}_1$ takes value, under dG_{10} (where the index 0 denotes, as before, the distribution evaluated at the true parameters),

$$\begin{aligned}
\langle h_1, B_0 h_2 \rangle &= \int h_1 \pi h_2 f g_1 g_2 d\mathbf{y} d\mathbf{x}_1 d\mathbf{x}_2 + \int h_1^0 (1-\pi) \left(\int \frac{h_2 f g_2}{f_{\mathbf{y}\mathbf{x}_1}} d\mathbf{x}_2 \right) f g_1 g_2 d\mathbf{y} d\mathbf{x}_1 d\mathbf{x}_2 \\
&= \int h_2 \pi h_1 f g_1 g_2 d\mathbf{y} d\mathbf{x}_1 d\mathbf{x}_2 + \int \left(\int \frac{h_1^0 (1-\pi) f g_2}{f_{\mathbf{y}\mathbf{x}_1}} d\mathbf{x}_2 \right) h_2 f g_1 d\mathbf{y} d\mathbf{x}_1 d\mathbf{x}_2 \\
&= \int \left(\int \pi h_1 f d\mathbf{y} \right) h_2 g_1 g_2 d\mathbf{x}_1 d\mathbf{x}_2 \\
&\quad + \int \left(\int \left(\int \frac{(1-\pi) h_1^0 f g_2}{f_{\mathbf{y}\mathbf{x}_1}} d\mathbf{x}_2 \right) f d\mathbf{y} \right) h_2 g_1 g_2 d\mathbf{x}_1 d\mathbf{x}_2 \\
&= \langle B_0^* h_1, h_2 \rangle,
\end{aligned}$$

where

$$\begin{aligned}
B_0^* h_1 &= \int \pi h_1 f d\mathbf{y} + \int \left(\frac{(1-\pi) h_1^0 f g_2}{f_{\mathbf{y}\mathbf{x}_1}} d\mathbf{x}_2 \right) f d\mathbf{y} \\
&= E\left(Rh_1|\mathbf{x}_1, \mathbf{x}_2\right) + E\Big(E\big((1-R)h_1|\mathbf{y}, \mathbf{x}_1\big)|\mathbf{x}_1, \mathbf{x}_2\Big)
\end{aligned}$$

is the adjoint operator. The same arguments can be used to show that $B_0^* B_0$ is invert-

ible, resulting in

$$h(\mathbf{x}_1,\mathbf{x}_2)\phi(\mathbf{x}_1,\mathbf{x}_2)+\int K(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_2')h(\mathbf{x}_1,\mathbf{x}_2')g_2(\mathbf{x}_2'|\mathbf{x}_1)d\mathbf{x}_2'=\psi(\mathbf{x}_1,\mathbf{x}_2),$$

which is again a Freedman equation of the second kind, where

$$\phi(\mathbf{x}_1,\mathbf{x}_2)=h\int \pi f d\mathbf{y},$$

$$\psi(\mathbf{x}_1,\mathbf{x}_2)=\int \pi\left(\frac{\partial}{\partial\boldsymbol{\beta}}\log L\right)f d\mathbf{y}+\int\left(\int(1-\pi)\left(\frac{\partial}{\partial\boldsymbol{\beta}}\log L\right)\frac{fg_2}{f_{\mathbf{y}\mathbf{x}_1}}d\mathbf{x}_2\right)f d\mathbf{y}$$

and

$$k(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_2')=\int(1-\pi)g_1\frac{f'f}{f'_{\mathbf{y}\mathbf{x}_1}}d\mathbf{y}.$$

7.2.2 Simulations

In order to check the asymptotic efficiency of the proposed method discussed in chapters 2 and 4, we performed a 2-phase study assuming the linear model

$$Y=\beta_0+\beta_1 x+\sigma\epsilon,$$

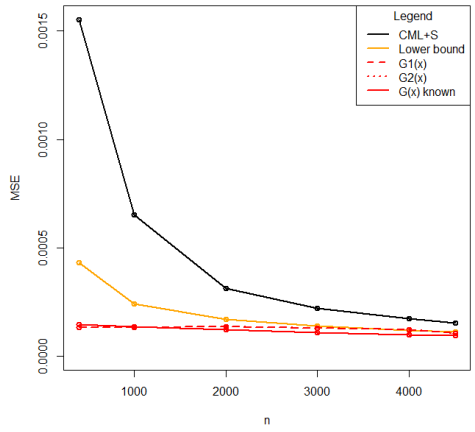
where ϵ and X follow a standard normal distribution and $\sigma=1$. We assumed that the response Y was fully observed at phase-1 and the covariate X was observed only at phase-2. The phase-2 data was sampled from each of the two strata defined by Y_d =I($Y<c_1^{(y)}$), where I is an indicator function and $c_1^{(y)}$ the $15th$ percentile of Y. We kept the intercept constant (equal to 1) and ran simulations with $\beta_1=0.5,1.0$ and 1.5, fitting the saturated model

$$\text{logit}(\pi)\sim y_d * y,$$

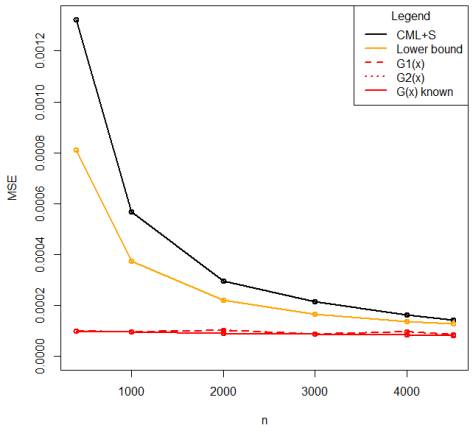
where $y_d * y$ is equivalent to $y_d + y + y_d y$, for the selection probability and varying the phase-2 sample sizes from 200 to 2250. For each sample size we ran 1000 simulations and, using equation (7.5), obtained the lower bound for the variance, i.e., the smallest MSE possible for any unbiased semiparametric estimator. This value was compared to the MSE for β_1 when using the CML+$\widetilde{\boldsymbol{S}}$ method discussed earlier in chapters 3 and 4. Our goal here is to see how far from the best possible semiparametric estimate our estimate actually is. These results are shown in figure 7.1. Figure 7.1 also shows the MSE when the true distribution of X is correctly fitted and under a fully parametric approach so that we can obtain a measure, in terms of MSE, of the amount of information lost when the distribution $G(X)$ is treated non-parametrically. The corresponding plots of log(MSE) are shown in Figure A.1.

Based on our simulations, we see that if the distribution of $G(X)$ is known (solid red line) or correctly fitted (dashed red lines) either using a smaller model denoted by $G1(X)$ (error normally distributed) or a larger model denoted by $G2(X)$ (error following a Generalized Normal distribution), these fully parametric approaches give the lowest MSE, as expected. Using the known $G(X)$ is slightly better than the others parametric approaches and there is very little difference between the smaller (2-parameters) or larger (3-parameters) fitted models. The discrepancy between these fully parametric and the other semiparametric methods increases as the correlation between Y and X increases. When the sample size is small, there is not enough information to estimate the marginal distribution of X efficiently, resulting in large MSE when compared to the fully parametric approaches. As the sample increases, the semiparametric approach is nearly as good as those fully parametric methods.

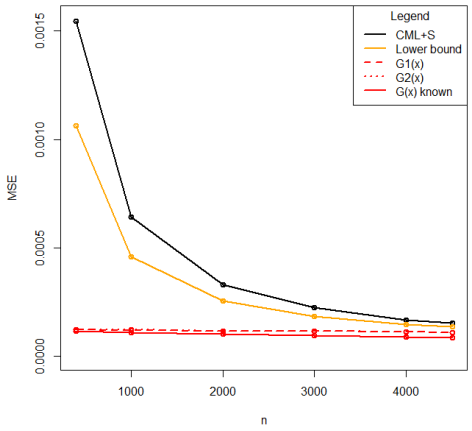
Regarding the semiparametric approaches, we see that the MSE obtained from CML+$\widetilde{\boldsymbol{S}}$ method appears to be approaching the semiparametric lower bound for the



(a) $\beta_1 = 0.5$



(b) $\beta_1 = 1.0$



(c) $\beta_1 = 1.5$

Figure 7.1: Mean squared error for $\hat{\beta}_1$ as a function of the sample size n, using the semiparametric CML+$\widetilde{\boldsymbol{S}}$ (black line) as well as the semiparametric lower bound for the variance (orange line) and the mean squared error for $\hat{\beta}_1$ when the true distribution of X is considered known (red line) or fitted by a smaller ($G1(X)$, dashed red line) or larger ($G2(X)$, dashed red line) parametric model.

variance as the sample size n, the proportion of individuals selected into phase-2, increases. For large n, the CML+$\widetilde{\boldsymbol{S}}$ method is nearly as good as the best semiparametric method to estimate β_1, for all cases considered here. For small n, however, both curves are considerably different, especially when β_1 is small. In this case, CML+$\widetilde{\boldsymbol{S}}$ gives an MSE that is almost 3 times higher than the one given by the semiparametric lower bound, but decreases to about 1.5 when $\beta_1 = 1.5$.

We also simulated a 2-phase study assuming the linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sigma\epsilon,$$

where X_1, X_2 and ϵ are independent and normally distributed with mean 0 and variance 1, and $\sigma = 1$. We assumed that (Y, X_1) were fully observed at phase-1 and that X_2 was only observed at phase-2, a sample taken from the six strata defined by (Y_d, X_{1d}), where

$$Y_d = \begin{cases} 1, & \text{if } Y < c_1^{(y)} \\ 0, & \text{otherwise} \end{cases} \qquad \text{and} \qquad X_{1d} = \begin{cases} 0, & \text{if } X_1 \leq c_1^{(x)} \\ 1, & \text{if } c_1^{(x)} < X_1 \leq c_2^{(x)} \\ 2, & \text{if } X_1 > c_2^{(x)} \end{cases}$$

where $c_1^{(y)}$ is the Y-$15th$ quantile and $c_1^{(x)}$ and $c_2^{(x)}$ are the X_1-$15th$ and X_1-$85th$ quantile, respectively. We also defined the binary coarsening X_{2d} equal to 1 if $X_2 < .5$ and 0 otherwise, which was considered known for all phase-1 individuals, but not used to select the phase-2 sample. Notice that this sampling scheme is similar to that illustrated in Fig. 4.1. As in chapter 4, we fitted two selection models for $\pi = \text{pr}(R_i = 1|\mathbf{z}, \boldsymbol{\alpha})$, where

$\boldsymbol{Z}$ must be contained in $(Y, Y_d, X_1, X_{1d}, X_{2d})$ so that the estimates are unbiased for $\boldsymbol{\beta}$:

$$\text{Model (1): } \text{logit}(\pi) \sim y_d * x_{1d} + y * x_1$$

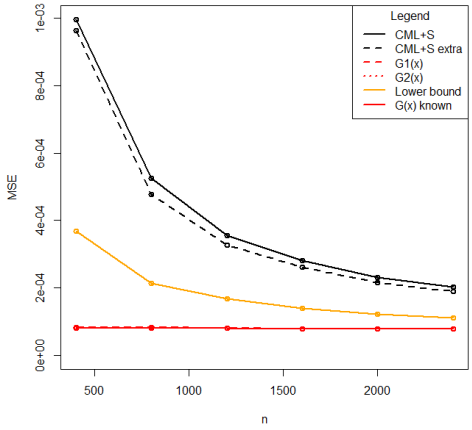
and

$$\text{Model (2): } \text{logit}(\pi) \sim y_d * x_{1d} * x_{2d} + y * x_1 * x_{2d}.$$

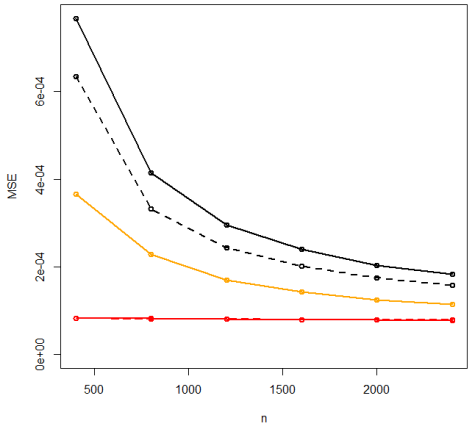
Note that model (1) is the true selection mechanism plus the interaction between the continuous variables Y and X_1. Model (2) is similar, but takes the extra information on X_{2d} into consideration. We used the CML+$\widetilde{\boldsymbol{S}}$ method to estimate the parameters of interest and we use the notation $\text{CML}^{(i)} + \widetilde{\boldsymbol{S}}$ for model (i), $i = 1$ or 2, to model the selection probabilities.

We compared the $\text{CML}^{(1)} + \widetilde{\boldsymbol{S}}$ and the $\text{CML}^{(2)} + \widetilde{\boldsymbol{S}}$ methods with respect to MSE for different $\boldsymbol{\beta}$ and phase-2 sample sizes n, against the semiparametric lower bound for the variance and fully parametric methods assuming that the distribution of $\boldsymbol{X} = (X_1, X_2)$ was known or fitted using a smaller error model (error normally distributed) and a larger error model (errors following a Generalized Normal distribution), exactly as in the previous case. Results for $\hat{\beta}_1$ and $\hat{\beta}_2$ are shown in Figures 7.2 and 7.3, respectively. The log plots are shown in Figures A.2 and A.3. Notice that only $\text{CML}^{(2)} + \widetilde{\boldsymbol{S}}$ makes use of the available information on X_{2d}. In both figures we used $\boldsymbol{\beta} = \left\{(1,1,.5)^T, (1,.5,.5)^T, (1,.5,1)^T\right\}$ to generate subfigures (a), (b) and (c), respectively. Our goal was to calculate the MSE for cases where the correlation between X_2 and Y were equal, higher or similar to the correlation between X_1 and Y.

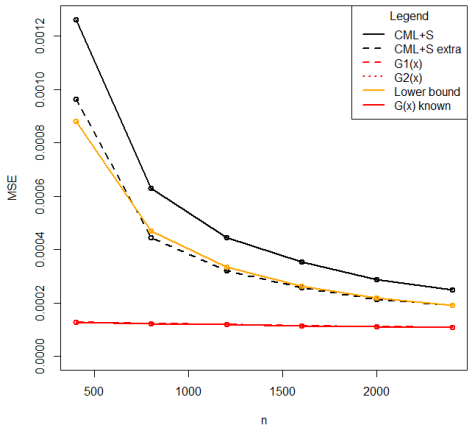
Both figures show a similar order. As in previous case (Fig. 7.1), the MSE of $\text{CML}^{(1)} + \widetilde{\boldsymbol{S}}$ becomes closer to the semiparametric lower bound for the variance as the sample size increases. The difference between these curves is slightly higher for



(a) $\boldsymbol{\beta} = (1, 1, .5)^T$

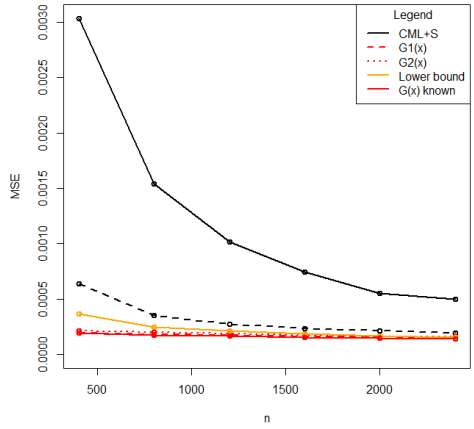


(b) $\boldsymbol{\beta} = (1, .5, .5)^T$

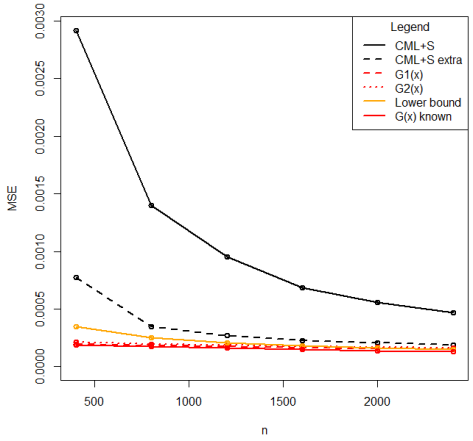


(c) $\boldsymbol{\beta} = (1, .5, 1)^T$

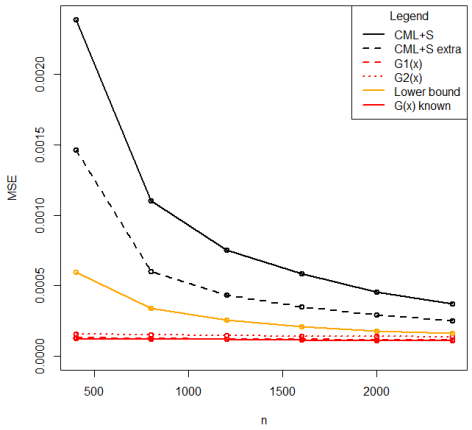
Figure 7.2: Mean squared error for $\hat{\beta}_1$ as a function of the phase-2 sample size n, using the semiparametric CML+$\widetilde{\boldsymbol{S}}$ without (solid black line) and with (dashed black line) the extra information X_{2d} added into the selection model, the semiparametric lower bound for the variance (orange line) and the mean squared error for $\hat{\beta}_1$ when the true distribution of X is considered known (red line) or fitted by a smaller ($G1(X)$, dashed red line) or larger ($G2(X)$, dashed red line) parametric model.



(a) $\boldsymbol{\beta} = (1, 1, .5)^T$



(b) $\boldsymbol{\beta} = (1, .5, .5)^T$



(c) $\boldsymbol{\beta} = (1, .5, 1)^T$

Figure 7.3: Mean squared error for $\hat{\beta}_2$ as a function of the phase-2 sample size n, using the semiparametric CML+$\widetilde{\boldsymbol{S}}$ without (solid black line) and with (dashed black line) the extra information X_{2d} added into the selection model, the semiparametric lower bound for the variance (orange line) and the mean squared error for $\hat{\beta}_2$ when the true distribution of X is considered known (red line) or fitted by a smaller ($G1(X)$, dashed red line) or larger ($G2(X)$, dashed red line) parametric model.

β_2 and decreases as β_2 increases. For example, when $\boldsymbol{\beta} = (1, .5, 1)$ and $n = 800$, the $\text{CML}^{(1)} + \widetilde{\boldsymbol{S}}$ method gives an MSE that is almost 5.5 higher than the one given by the semiparametric lower bound while estimating β_2 and about 3 times higher when $\boldsymbol{\beta} = (1, .5, 1)$ and $n = 2400$. While estimating β_1, this ratio varies from 1.8 (when $n = 800$) to 1.6 (when $n = 800$)

Adding extra information regarding X_{2d}, a simple binary coarsening, into the selection model results in much better estimates, as seen by the black dashed line in both figures and already discussed in previous chapters. The smaller the phase-2 sample size is, the greater the reduction is.

As expected, using fully parametric approaches where the distribution of $\boldsymbol{X}$ is known or fitted using a correct parametric model, leads to smaller MSE in all cases. This MSE is significantly smaller than the semiparametric approaches especially when the coefficient β_1 is large, and is still reasonably smaller for large sample size. However, because $\boldsymbol{X}$ is often of high-dimensional, modelling its joint distribution may be hard or even infeasible and small deviations from the true distribution $G(\boldsymbol{X})$ may lead to biased estimates. We conduct a small-scale investigation of this in the next section.

7.3 Parametric model for the covariates

Here we perform a small simulation study to examine the impact of a misspecified $G(\boldsymbol{X})$ on the MSE. Let $G(\boldsymbol{X})$ be the distribution function for $\boldsymbol{X}$ and $g(\mathbf{x})$ its probability density function. Let $\boldsymbol{X}$ be, as before, a partially observed variable known only for a sample of individuals selected from a larger population. Let also the response Y be fully observed, known for the entire population, and used to select the final sample. Finally, let R be an indicator of being selected for full observation or not.

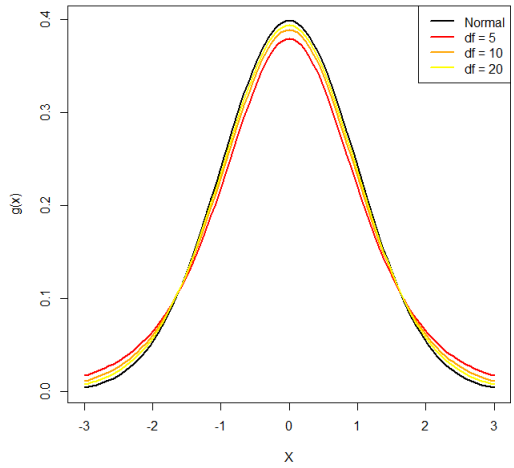


Figure 7.4: Density functions for X following a t-distribution.

Assume a parametric model for the conditional distribution of Y given X, $f(y|\mathbf{x};\beta)$, and a parametric distribution $g^*(\mathbf{x})$ for $g(\mathbf{x})$. The resulting estimating equation for β is thus given by

$$\boldsymbol{S}_0 = \sum_{i=1}^{N} \left[R_i \log f(y|\mathbf{x};\beta) + (1-R_i) \log \int f(y|\mathbf{x};\beta) g^*(\mathbf{x}) d\mathbf{x} \right] = 0, \tag{7.6}$$

where N is the total population. Note that the expected value of the first term of equation (7.6) is zero since we are assuming that model for $Y|\boldsymbol{X}$ is correctly specified, but the second term is only zero if $g^*(\boldsymbol{X}) \equiv g(\mathbf{x})$. That is, we would expect that slightly misspecified models for $\boldsymbol{X}$ lead to biased estimates and our interest here is to see, through simulations, how large the resulting bias might be.

We ran simulations generating $\boldsymbol{X}$ from a t-distribution with υ degrees of freedom. We fitted a fully parametric model with $g^*(\boldsymbol{X})$ normally distributed and the conditional distribution of $Y|\boldsymbol{X}$, the model of interest, correctly specified. Figure 7.4 shows the true (t-distribution) and the misspecified (Normal) distributions, for $\upsilon = (5, 10, 20)$.

We used a 2-phase sampling scheme with a total population of $N = 10,000$ individ-

uals with an univariate X and a linear response

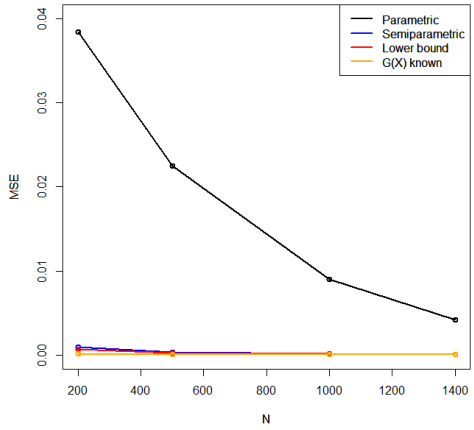
$$Y = \beta_0 + \beta_1 x + \sigma\epsilon.$$

Here $\sigma = 1$, $\boldsymbol{\beta} = (1, 1.5)^T$ and ϵ and X follow a standard normal distribution. Only the response was considered known for all individuals and the covariate X was measured only for a sample of units taken from the phase-1 population. Subjects were selected from each of the two strata defined by $Y_d = \mathrm{I}(Y < c_1^{(y)})$, where I is an indicator function and $c_1^{(y)}$ the $15th$ quantile of Y.

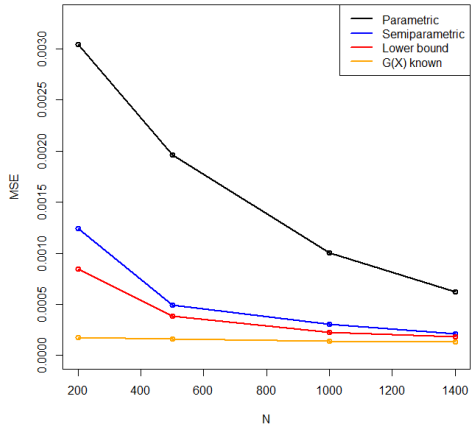
For the simulation study, we varied the phase-2 sample size from $n = 400$ to $n = 2,800$, while fitting both the selection model and the model of interest $f(y|x;\boldsymbol{\beta})$, correctly. For each sample size we calculated the MSE of β_1 and compared the following methods:

- The proposed semiparametric CML+$\widetilde{\boldsymbol{S}}$ method;
- The fully parametric approach with $g^*(x)$, the misspecified parametric distribution, used to fit the distribution of X;
- The fully parametric approach with $g(x)$, the correctly specified distribution;
- The semiparametric lower bound for the variance.

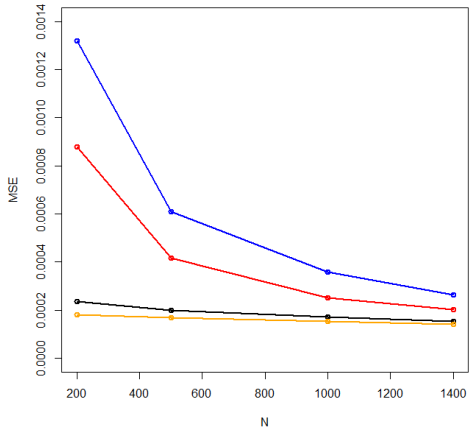
The results are shown in Fig. 7.5, when X was generated from a t-distribution. Values of the log(MSE) plots are shown in Fig. A.4 When $\upsilon = 5$, the true distribution of X is heavy-tailed and the misspecified fully parametric approach results in biased estimates and large MSE. The semiparametric method, which does not assume any distribution for X is still unbiased and gives a much smaller MSE. The same pattern is observed when $\upsilon = 10$. The misspecified model is still biased, resulting in a MSE larger than



(a) $\upsilon = 5$



(b) $\upsilon = 10$



(c) $\upsilon = 20$

Figure 7.5: Mean squared error for the parametric (black line) and semiparametric (blue line) methods, lower bounds for the variance (red line) and mean squared error when the true distribution of X is known (orange line), for X following a t-distribution with (a) 5, (b) 10 and (c) 20 degrees of freedom υ.

the one obtained by the more robust approach. Finally, only when $\upsilon = 20$ and the t-distribution is very close to the normal distribution, the fully misspecified parametric approach is more efficient than any semiparametric estimator and almost as efficient as the optimal fully parametric approach, where the distribution of X is correctly fitted.

7.4 Summary

Here we showed that the proposed CML+$\widetilde{\boldsymbol{S}}$ is fully efficient when the response is binary, for a wider range of scenarios than those considered in Scott and Wild (2011), and through simulations studied its efficiency when the response is continuous. To this end, we first derived the asymptotic lower bound for the variance by solving an integral equation and wrote an R code for computing it numerically. Comparing this lower bound against CML+$\widetilde{\boldsymbol{S}}$, we saw that their difference seems to decrease as the sample size increases when the sampling depends only on Y, but at a lower rate when the sampling depends on both Y and X_1. It is still of interest to show, mathematically, whether or not the CML+$\widetilde{\boldsymbol{S}}$ method achieves the asymptotic semiparametric lower bound for the variance and also derive its rate of convergence.

Our simulations also showed that assuming a semiparametric approach, where the distribution of the covariates $\boldsymbol{X}$ are treated non-parametrically and is thus not affected by misspecification, leads to considerable loss in efficiency if compared to fitting the correct distribution for $\boldsymbol{X}$. However, modeling $\boldsymbol{X}$ is often hard and fairly small misspecifications, as shown by our simulation study, can result in a large mean squared error.

8

Conclusions and Future Work

In this dissertation we have proposed a semiparametric approach to deal with multiphase sampling schemes that take into account all information available in the study. Cheap surrogates, for example, which are often discarded in multiphase studies, are considered in a fairly simple way with the purpose of increasing precision. The method is semiparametric in the sense that it does not require any assumption regarding how the covariates are distributed and so it is robust against such misspecifications.

The theory as well as asymptotic results were derived in chapter 2 for both discrete and continuous responses, extending Scott and Wild (2011). Estimating selection probabilities require modelling an often binary outcome given a set of variables, being thus strongly connected to the propensity score approach. However, both approaches do differ in their definition and intent, and their similarities and differences were discussed in chapter 5. While the propensity score was developed with the purpose of replicating, by balancing the background covariates, a randomized experiment while dealing with an observational study, our proposed approach was used to deal with biased or outcome-dependent sampling schemes. Therefore, unless one is willing to define

an outcome-dependent propensity score and thus ignoring its balancing property, both approaches are different.

As emphasized early, in this dissertation we worked with both discrete and continuous responses. Each one was analysed separately: Discrete responses were discussed in chapter 3 and the continuous case was treated in chapter 4. In the former, we discussed 2 and 3-phase designs with extra information that was not used in any part of the study, but available in both first and second phases of the study. We first applied it to situations where the missingness was due to design (controlled by the researcher and thus known). In such cases it is often possible to fit a saturated model for the selection probability and our simulated results showed that making use of this extra variable in a very simple way may lead to sometimes substantial gains in efficiency. The proposed method is always the best alternative for estimating the effects the variable that goes missing (i.e., only observed at the final phase).

We have also discussed, through simulations, the impact of adding an extra variable into the selection model when the response was continuous. Here we assumed that the error distribution followed a Normal, Generalized Normal, Skew-Normal or t-distribution. As for the discrete case, making use of extra variables resulted in estimates that were substantially more efficient, especially for estimating the effects the variable that goes missing. Imputation and calibration can be more efficient for estimating the effects of the complete observed variables, especially when the partially observed variable (observed only at the final phase) is weakly related to the outcome. This situation, however, is reversed as the impact of the partially observed variable on the outcome increases.

It is important to point out that, in both discrete and continuous cases, our proposed method did not make any assumption regarding this extra variable nor its distribution,

being thus robust against misspecifications. This is appealing because in many studies cheap variables are often available for part of the population, but are usually ignored. In this way we hope to encourage the collection of cheap or easy to measure variables that may contribute to the final estimates.

The proposed method is also flexible enough to be applied to situations where fitting a saturated model for the selection probability is not possible, which might the case of non-response problems. Here the missingness is unknown and fitting a saturated model might not be possible. As a result, most efficient approaches may not be applied here. Our method comes as an alternative approach and its efficiency with respect to the non-response rate was studied through simulations, extending the work of Jiang et al. (2011). The missing at random assumption, of course, must be valid so that the method can provide unbiased or nearly unbiased estimates.

This approach can in addition be applied to a variety of scenarios, which encompasses the expensive covariate or expensive response problems, the secondary analysis and many others, as discussed in chapter 6. Even though the data may be obtained via different sampling schemes corresponding to perhaps completely different problems, the resulting estimating equations all have the same structure and so the method is readily extended to all those scenarios. It was implemented in R for both discrete and continuous variables, extending previous work of Neuhaus et al. (2006) and Jiang et al. (2006), to name a few.

A drawback, however, is its lack of robustness against model misspecification. In chapter 4 we study the worst possible misspecification and compared nearly-true models against an alternative robust approach. We noticed that there are situations in which even small deviations from the true model could lead to very inefficient estimates. Dealing with levels of misspecifications that are around the level of detectability seems

to be the right level to concentrate on. Levels that could always be detected (as in Hade and Lu (2013)) are not relevant because an analyst would not be fitting the model in a situation like that.

Its efficiency is discussed in chapter 7. Here we first showed that, in the discrete case, the proposed method is equivalent to the Scott and Wild (1997) approach and thus semiparametric fully efficient. More generally (and for a 2-phase design), we derived the semiparametric lower bound for the variance by solving numerically an integral equation. R code was then written to compute this lower bound and used to study the efficiency of the proposed method in a few scenarios. From our simulations we saw that, if the missingness depends only on the response, the proposed method seems to converge to the best possible semiparametric method as the phase-2 sample size increases. If the sampling depends on both the response and on some continuous covariate, this convergence rate (assuming that there is one) seems to be much lower. More work is required here.

Future Work

In this dissertation we have considered that the extra variable was, in most cases, a simple binary coarsening. This information was simply used by adding it into the selection model, but this may not be the most efficient way. As discussed in section 2.2.1, the reduction of the asymptotic variance depends on the relationship between the score function from the selection model and the score function from the model of interest, in the sense that the greater the correlation between them are, the greater is the reduction of the asymptotic variance. Thus, it is of interest to examine different ways and possible derive the most efficient way of making use of this extra information. This is currently being investigated.

In cases where the extra variable is correlated to the outcome, as discussed in chapter 6, we should use the expected value of this selection probability given a set of variables fully observed at phase-1, which requires assuming a parametric model for this extra variable in order to get unbiased estimates. This increases risks of model misspecifications and perhaps mixed methods that do some degree of adjustments and adds some robustness into the estimating equations should be considered.

Another topic of work is to develop diagnostic measures for multiphase sampling schemes. Here in this thesis we were mainly concerned on developing a set of estimating equations for different sampling designs rather than discussing model adequacy. However, since as noticed in our simulated studies, the proposed method, which strongly relies on the model of interest, may provide biased estimates if the model of interest is slightly misspecified. A few papers, such as Zhu et al. (2009), Shi et al. (2009) and Zhu et al. (2012), discuss this topic, but their approaches rely on fully parametric methods. We expect to develop semiparametric approaches that also cover more general sampling schemes as those discussed in this thesis.

Another potential research topic is to check if the proposed method is in fact fully efficient when the response is continuous. Here we only studied its efficiency via simulations, comparing it against the semiparametric lower bound for the variance. It is also of interest to derive semiparametric lower bound when extra information is available.

Finally, we also plan to polish the R package implementing the methods discussed in this dissertation and make it available on CRAN in the near future.

A

Appendix: Tables and Figures

Table A.1: Results comparing efficiency of small and large model used to fit the error distribution. For the smaller model (cml+S small), we used a simple normal model and for the larger model (cml+S large) we used the Generalized Normal distribution, discussed in chapter 4, that reduces to the Normal distribution when the shape parameter θ is equals to 2.

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\boldsymbol{\beta} = (1,1,1)^T$	wgt	-0.001	0.001	0.002	0.991	0.997	0.977	1.127	1.214	1.606	0.945	0.947	0.948
	cml+S small	0.002	0.001	0.002	1.053	1.023	1.031	1.023	0.832	1.164	0.957	0.960	0.960
	cml+S large	0.002	0.000	0.003	1.062	1.000	1.032	1.013	0.888	1.194	0.955	0.949	0.961
$\boldsymbol{\beta} = (1,1,2)^T$	wgt	-0.001	0.000	0.002	1.009	0.986	0.956	1.499	1.607	1.620	0.958	0.950	0.937
	cml+S small	0.002	0.000	0.002	1.016	1.013	0.942	1.534	1.075	1.202	0.957	0.946	0.933
	cml+S large	0.002	0.000	0.002	1.044	0.987	0.916	1.475	1.150	1.299	0.957	0.940	0.930
$\boldsymbol{\beta} = (1,2,1)^T$	wgt	0.001	-0.001	0.002	1.015	0.979	0.984	1.102	1.491	1.761	0.958	0.949	0.943
	cml+S small	0.002	0.001	0.002	1.031	0.982	1.039	1.065	0.807	1.231	0.957	0.942	0.951
	cml+S large	0.004	0.000	0.003	0.942	0.994	1.048	1.295	0.800	1.230	0.941	0.947	0.956

Table A.2: Results for the same settings as those used in chapter 4, but varying the error distribution ϵ. Here we considered that the error distribution followed a t-distribution with υ (for $\upsilon = 5$ or 10) degrees of freedom and a Skew-Normal distribution with shape parameter κ (for $\kappa = 1$ or 1.5).

		Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$\epsilon \sim T(5)$	cml	0.005	0.000	0.005	0.950	1.065	0.990	2.595	1.459	2.213	0.935	0.965	0.946
	cml+S	0.003	0.000	0.007	0.937	1.061	1.003	2.457	1.419	2.030	0.942	0.962	0.938
	wgt	0.004	0.000	0.004	0.941	1.021	0.981	2.855	2.662	3.343	0.939	0.956	0.949
$\epsilon \sim T(10)$	cml	-0.018	0.000	0.003	0.815	1.039	0.969	3.300	1.304	2.043	0.872	0.959	0.935
	cml+S	-0.016	0.000	0.004	0.810	1.033	1.038	3.119	1.313	1.679	0.862	0.961	0.947
	wgt	-0.001	0.000	0.001	0.940	1.022	1.034	2.391	2.156	2.341	0.945	0.951	0.955
$\epsilon \sim S(1)$	cml	0.023	-0.014	0.005	0.600	0.743	0.914	6.139	1.846	1.471	0.792	0.852	0.924
	cml	0.008	-0.002	0.001	0.701	0.984	1.057	3.473	0.906	1.048	0.828	0.945	0.954
	wgt	0.564	0.002	0.000	1.000	1.022	1.014	319.045	1.018	1.412	0.000	0.959	0.945
$\epsilon \sim S(1.5)$	cml	0.001	0.001	-0.001	1.018	0.983	0.973	1.877	0.722	0.890	0.946	0.938	0.939
	cml	0.002	0.002	0.001	1.002	1.012	0.978	1.873	0.669	0.862	0.947	0.947	0.941
	wgt	0.665	0.001	0.000	0.965	1.035	0.971	443.332	0.860	1.302	0.000	0.952	0.939

Table A.3: Table 5.1 extended.

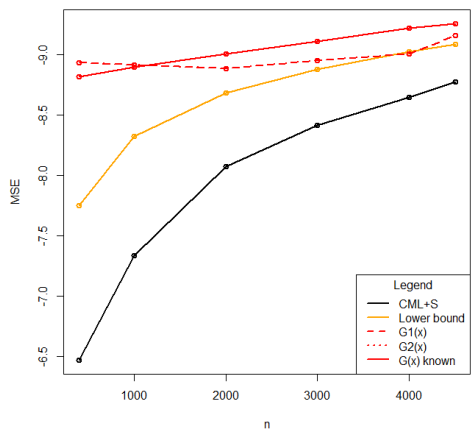
		Bias					Empirical SE				
Models	Method	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4
	Wgt	0.0034	-0.0058	0.0045	0.0023	-0.0040	0.2617	0.3524	0.2069	0.1721	0.0793
Y on (T, X_1, X_2, Z)	Adj	-0.0031	-0.0071	0.0089	0.0024	-0.0042	0.3625	0.2520	0.1852	0.1575	0.0550
	Lin	0.0038	-0.0061	0.0062	-0.0006	-0.0038	0.1682	0.2286	0.1331	0.1100	0.0533
	Wgt	0.1047	0.0023	0.3356	0.0095		0.2967	0.3752	0.2178	0.1818	
Y on (T, X_1, X_2)	Adj	-0.1305	-0.0023	0.4466	0.0914		0.4828	0.2666	0.2270	0.1959	
	Lin	0.1011	0.0794	0.3642	0.0110		0.2018	0.2578	0.1500	0.1181	
	Wgt	0.2839	-0.2565	0.6954			0.3911	0.4449	0.2830		
Y on (T, X_1)	Adj	1.7571	-0.0176	-0.0006			0.3779	0.2739	0.2508		
	Lin	0.7160	-0.9391	0.5463			0.2188	0.2731	0.1759		
	Wgt	11.4317	-0.6982				0.4678	0.7140			
Y on T	Adj	3.1182	-0.0035				0.2553	0.2801			
	Lin	2.3217	-2.6135				0.2037	0.2824			

Table A.4: Table 5.1 extended.

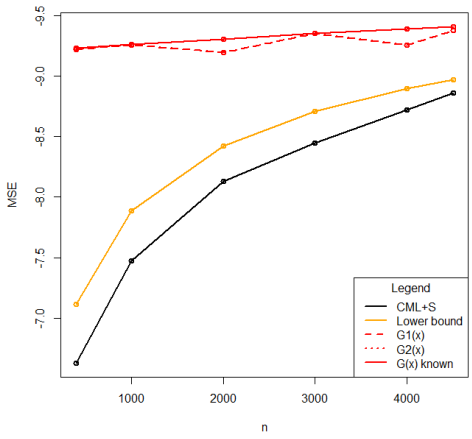
		SE					% Coverage				
Models	Method	β_0	β_T	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4
	Wgt	0.1349	0.1651	0.1251	0.0996	0.0581	0.688	0.645	0.780	0.755	0.861
Y on (T, X_1, X_2, Z)	Adj	0.3611	0.2549	0.1815	0.1570	0.0554	0.956	0.947	0.941	0.952	0.952
	Lin	0.1713	0.2315	0.1318	0.1074	0.0541	0.953	0.959	0.945	0.943	0.950
	Wgt	0.1445	0.1792	0.1158	0.1082		0.657	0.654	0.278	0.753	
Y on (T, X_1, X_2)	Adj	0.3889	0.2839	0.1735	0.1694		0.868	0.966	0.313	0.877	
	Lin	0.1883	0.2561	0.1250	0.1184		0.895	0.940	0.206	0.951	
	Wgt	0.1790	0.2222	0.1366			0.492	0.575	0.044		
Y on (T, X_1)	Adj	0.2900	0.3202	0.1791			0.002	0.974	0.819		
	Lin	0.2124	0.2757	0.1502			0.091	0.075	0.091		
	Wgt	0.2162	0.3116				0.018	0.305			
Y on T	Adj	0.1690	0.3575				0.000	0.988			
	Lin	0.1887	0.2925				0.000	0.000			

Table A.5: Model misspecification applied to *case (i)*.

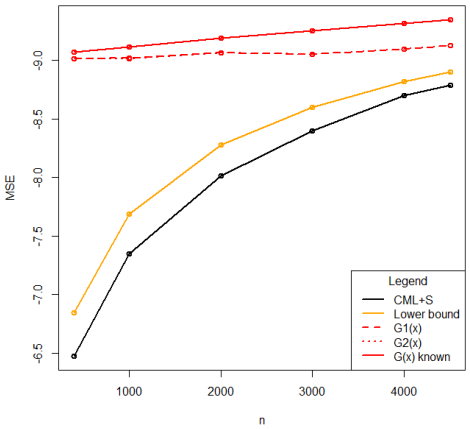
Method	Bias			Est.SE/Emp.SE			MSE $\times 10^{-3}$			Coverage %		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Fitted $v \sim y + y*x_1 + y*x_2$ (correct model) and true $\epsilon \sim T(5)$												
wgt	-0.058	0.033	0.032	1.024	0.974	0.938	59.129	50.520	52.955	0.963	0.953	0.941
cml	-0.253	-0.071	-0.08	1.029	1.000	0.961	116.598	49.376	53.197	0.881	0.928	0.916
cml*	-0.016	0.030	0.027	1.080	1.008	0.961	50.857	43.994	45.367	0.972	0.956	0.945
Fitted $v \sim y + y*x_1 + y*x_2$ (correct model) and true $\epsilon \sim T(10)$												
wgt	-0.052	0.042	0.024	0.907	0.928	0.954	74.199	56.935	51.798	0.936	0.940	0.948
cml	-0.260	-0.065	-0.079	0.921	0.936	0.919	134.432	55.585	59.360	0.860	0.910	0.888
cml*	-0.016	0.032	0.017	0.951	0.940	0.942	63.578	50.447	46.601	0.938	0.948	0.944
Fitted $v \sim y + y*x_1 + y*x_2$ (correct model) and true $\epsilon \sim T(20)$												
wgt	0.001	-0.010	0.007	0.962	0.873	0.972	59.682	59.756	46.416	0.929	0.893	0.964
cml	-0.212	-0.113	-0.096	0.976	0.875	0.974	101.284	69.331	53.506	0.893	0.857	0.914
cml*	0.0258	-0.016	-0.003	1.018	0.884	0.970	52.269	53.826	40.216	0.936	0.914	0.950



(a) $\beta_1 = 0.5$

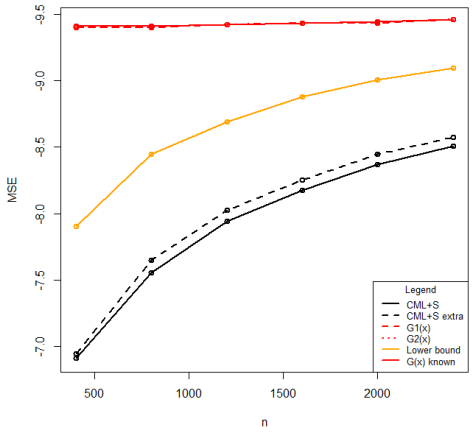


(b) $\beta_1 = 1.0$

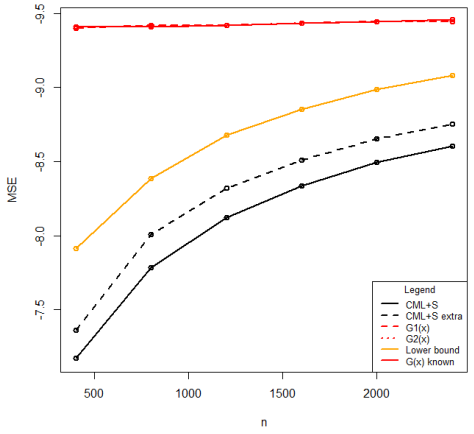


(c) $\beta_1 = 1.5$

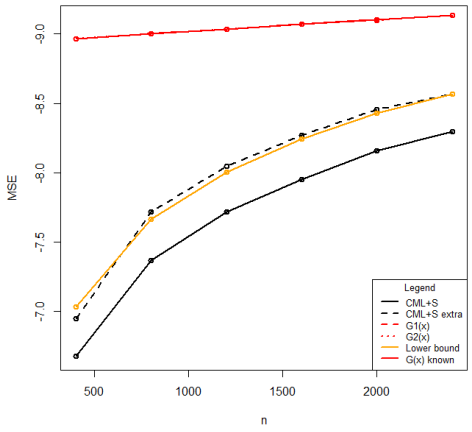
Figure A.1: Log of the mean squared error for $\hat{\beta}_1$ as a function of the phase-2 sample size n, using the semiparametric CML+$\widetilde{\boldsymbol{S}}$ (black line) as well as the semiparametric lower bound for the variance (orange line) and the mean squared error for $\hat{\beta}_1$ when the true distribution of X is considered known (red line) or fitted by a smaller ($G1(X)$, dashed red line) or larger ($G2(X)$, dashed red line) model.



(a) $\boldsymbol{\beta} = (1, 1, .5)^T$

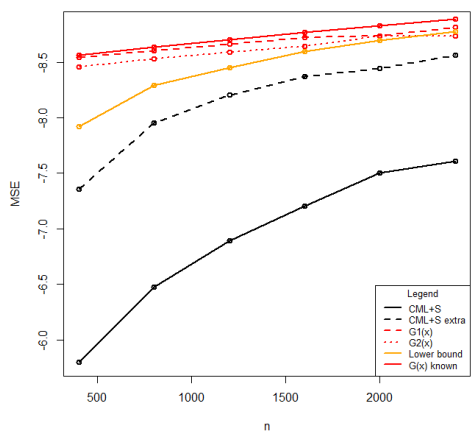


(b) $\boldsymbol{\beta} = (1, .5, .5)^T$

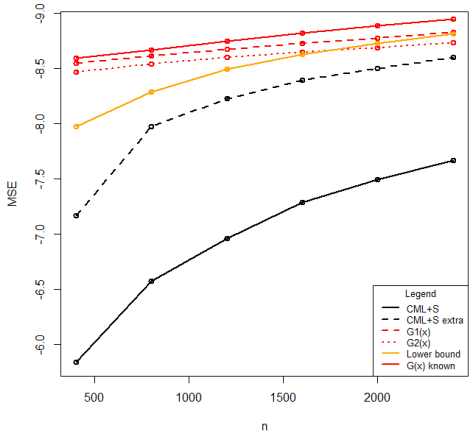


(c) $\boldsymbol{\beta} = (1, .5, 1)^T$

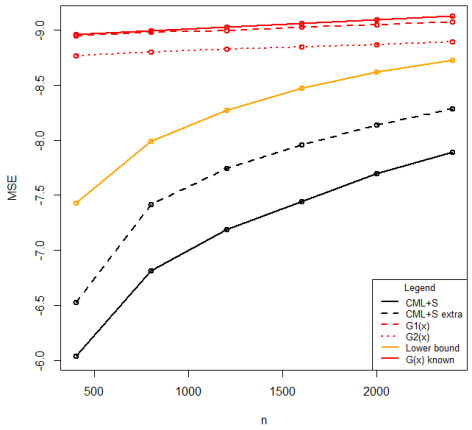
Figure A.2: Log of the mean squared error for $\hat{\beta}_1$ as a function of the phase-2 sample size n, using the semiparametric CML+$\widetilde{\boldsymbol{S}}$ without (solid black line) and with (dashed black line) the extra information X_{2d} added into the selection model, the semiparametric lower bound for the variance (orange line) and the mean squared error for $\hat{\beta}_1$ when the true distribution of X is considered known (red line) or fitted by a smaller ($G1(X)$, dashed red line) or larger ($G2(X)$, dashed red line) model.



(a) $\boldsymbol{\beta} = (1, 1, .5)^T$

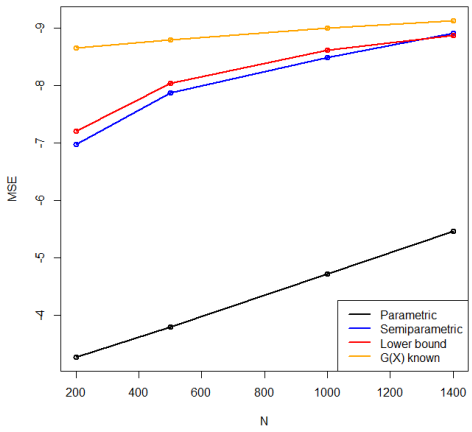


(b) $\boldsymbol{\beta} = (1, .5, .5)^T$

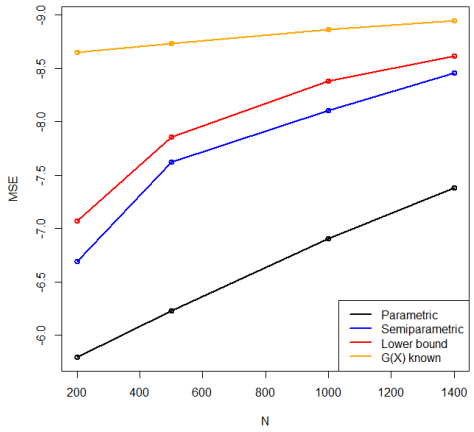


(c) $\boldsymbol{\beta} = (1, .5, 1)^T$

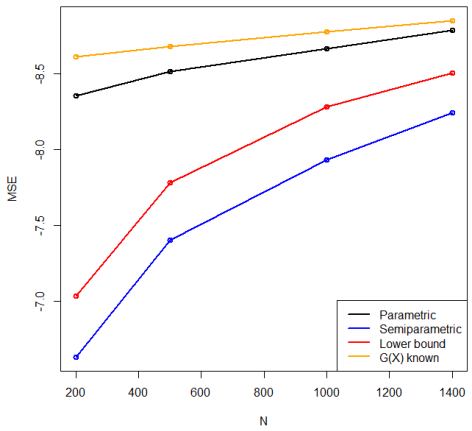
Figure A.3: Log of the mean squared error for $\hat{\beta}_2$ as a function of the phase-2 sample size n, using the semiparametric CML+$\widetilde{\boldsymbol{S}}$ without (solid black line) and with (dashed black line) the extra information X_{2d} added into the selection model, the semiparametric lower bound for the variance (orange line) and the mean squared error for $\hat{\beta}_2$ when the true distribution of X is considered known (red line) or fitted by a smaller ($G1(X)$, dashed red line) or larger ($G2(X)$, dashed red line) model.



(a) $\upsilon = 5$



(b) $\upsilon = 10$



(c) $\upsilon = 20$

Figure A.4: Log of the mean squared error for the parametric (black line) and semiparametric (blue line) methods, lower bounds for the variance (red line) and mean squared error when the true distribution of X is known (orange line), for X following a t-distribution with (a) 5, (b) 10 and (c) 20 degrees of freedom υ.

Bibliography

G. Anderson, J. Manson, R. Wallace, B. Lund, D. Halla, S. Davis, S. S. andg CY Wan, E. Stein, and R. Prentice. Implementation of the womenŠs health initiative study design. *Ann Epidemiol*, 14(4):S5–S17, 2003. 83

J. A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972. 19

O. Ashenfelter. Estimating the effect of training programs on earnings. *Review of Economics and Statistics*, 6:47–57, 1978. 120

O. Baser. Too much ado about propensity score models? comparing methods of propensity score matching. *Value Health*, 9:377–85, 2006. 125

G. Biondi-Zoccai, E. Romagnoli, P. Agostoni, D. Capodanno, D. Castagno, F. D'Ascenzo, G. Sangiorgi, and M. G. Modena. Are propensity scores really superior to standard multivariate analysis. *Contemporary Clinical Trials*, 32:731–740, 2011. 129

G. E. P. Box. A note on regions of kurtosis. *Biometrika*, 40:465–468, 1953. 93

N. Breslow and R. Holubkov. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J R Stat Soc, Ser B*, 59: 447–461, 1997. 83

N. E. Breslow and K. C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20, 1988. 7, 32

N. E. Breslow, T. Lumley, C. M. Ballantyne, L. E. Chambless, and M. Kulich. Improved horvitzŰthompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in biosciences*, 1:32–49, 2009. 26

N. E. Breslow, G. Amorim, M. B. Pettinger, and J. Rossouw. Using the whole cohort in the analysis of case-control data. *Statistics in Biosciences*, Published online in Wiley Online Library:1–18, 2013. 53

R. J. Carroll and M. P. Wand. Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society. Series B*, 53(3):573–585, 1991. 14

N. Chatterjee and Y. H. Chen. A semiparametric pseudo-score method for analysis of two-phase studies with continuous phase-i covariates. *Lifetime Data Analysis*, 13: 607–622, 2007. 13, 14, 21

N. Chatterjee, Y. H. Chen, and N. E. Breslow. A pseudo score estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 66:403–411, 2003. 12, 13

H. Y. Chen. Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the American Statistical Association*, 99(468):1176–1189, 2004. 116

G. Claeskens and N. L. Hjort. *Model selection and model averaging.* Cambridge Series in Statistical and Probabilistic Mathematics, 2008. 113

K. A. Clarke, B. Kenkel, and M. R. Rueda. Misspecification and the propensity score:

The possibility of overadjustment, 2011. URL `http://rochester.edu/college/psc/clarke`. 152

D. Clayton, D. Spiegelhalter, G. Dunn, and A. Pickles. Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society B*, 60:71–87, 1998. 157

B. J. Crowe, I. A. Lipkovich, and O. Wang. Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharmaceutical Statistics*, 9:269–279, 2010. 150

R. B. D'Agostino. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17:2265–81, 1998. 127

R. B. D'Agostino and D. B. Rubin. Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95(451): 749–759, 2000. 150

J.-C. Deville and C.-E. Sarndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):377–382, 1992. 25, 26

J. A. Domínguez-Molina, G. González-Farías, and R. M. Rodríguez-Dagnino. A practical procedure to estimate the shape parameter in the generalized gaussian distribution. report. *http://www.cimat.mx*, pages 1–37, 2001. 96

V. M. Estevao and C.-E. Sarndal. Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2):127–147, 2006. 23

R. V. Foutz. On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association*, 72:147–148, 1977. 46

D. A. Freedman and R. A. Berk. Weighting regressions by propensity scores. *Evaluation Review*, 32:392–409, 2008. 131, 132, 133, 134, 137

P. Ghosh and A. Dewanji. Analysis of spontaneous adverse drug reaction (adr) reports using supplementary information. *Statistics in Medicine*, 16:2040–2055, 2011. xi, 74, 75

E. M. Hade and B. Lu. Bias associated with using the estimated propensity score as regression covariate. *Statistics in Medicine*, Published online in Wiley Online Library, 2013. 124, 127, 135, 136, 140, 214

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometric*, 66:315–331, 1998. 124

F. R. Hampel, E. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics. The Approach Based on the Influence Function.* John Wiley & Sons, Inc., 1986. 92

S. Haneuse, T. Saegusa, and T. Lumley. osdesign: an r package for the analysis, evaluation, and design of two-phase and case-control studies. *J Stat Softw*, 43:1–29, 1997. 83

B. B. Hansen. Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99:609–618, 2004. 126

B. B. Hansen and J. Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 2:219–236, 2008. 128

F. E. Harrell. *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer Series in Statistics, 1st edition edition, 2002. 55

F. Hayashi. *Econometrics.* Princeton University Press, 2000. 128

J. Heckman, H. Ichumura, J. Smith, and P. Todd. Characterizing selection bias using experimental data. *Econometrica*, 66:1017–1098, 1999. 124

J. Hill. Discussion of research using propensity-score matching: Comments on 'a critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by peter austin. *Statistics in Medicine*, 27:2055–2061, 2008. 125

M. A. Hill and W. J. Dixon. Robustness in real life: a study of clinical laboratory data. *Biometrics*, 38:377–396, 1982. 92

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003. 124

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 86(396):945–960, 1986. 120, 129

D. Horvitz and D. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952. 22, 141

D. A. Hsieh, C. F. Manski, and McFadden. Estimation of response probabilities from augmented retrospective observations. *Journal of American Statistical Association*, 80:651–62, 1985. 34

P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964. 109

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society*, 2014. URL `http://imai.princeton.edu/research/CBPS.html`. 128

K. Imai, G. King, and E. A. Stuart. Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistics Society: Series A*, 171:481–502, 2008. 127, 128

Y. Jiang. *Semiparametric Maximum Likelihood for Multi-phase Response-Selective Sampling and Missing Data Problems.* PhD thesis, Department of Statistics, The University of Auckland, 2004. 90, 176, 177

Y. Jiang, A. J. Scott, and C. J. Wild. Secondary analysis of case-control data. *Statistics in Medicine*, 25:1323–1339, 2006. 29, 153, 157, 158, 176, 177, 178, 213

Y. Jiang, A. J. Scott, and C. J. Wild. Adjusting for non-response in two-phase case control studies. *International Statistical Review*, 79(2):145–159, 2011. 75, 78, 213

J. D. Y. Kang and J. L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 4(22):523–539, 2007. 127

R. Kress. *Linear Integral Equations (Applied Mathematical Sciences*, volume 82. Springer, 1999. 194

M. Kulich and D. Y. Lin. Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*, 99:832–844, 2004. 26

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 B*, 61(2):413–438, 1999. 8, 12

L. LeCam. A survey of sampling from contaminated distributions. *University of California Publications in Statistics*, 3:37–98, 1960. 113

A. Lee and Y. Hirose. Semi-parametric efficiency bounds for regression models under response-selective sampling: the profile likelihood approach. *Annals of the Institute of Statistical Mathematics*, 62:1023–1052, 2010. 18, 184

A. J. Lee, A. J. Scott, and A. L. McMurchy. Re-using data from case-control studies. *Statistics in Medicine*, 16:1377–1389, 1997. 157, 158, 165

R. J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993. 150

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. New York: John Wiley, 2nd edition, 2002. 2, 3, 6

T.-S. Lu. *Statistical inferences for outcome dependent sampling design with multivariate outcomes*. PhD thesis, University of North Carolina at Chapel Hill, 2009. 46

T. Lumley. *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Sons Inc., Hoboken, New Jersey, 2010. 24, 25

T. Lumley. Robustness of semiparametric efficiency in nearly-correct models for two-phase sample. *unpublished*, pages 1–14, 2013. 28, 88, 113, 114

T. Lumley, P. A. Shaw, and J. I. Dai. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review (unpublished)*, xx:1–30, 2011. 26, 34

C. F. Manski and D. McFadden. *Structural Analysis of Discrete Data with Econometric Applications*. Chapter 1, 1981. 32

D. F. McCaffrey, G. Ridgeway, and A. R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Journal of the American Statistical Association*, 9:403–425, 2004. 128

D. K. McLeish and C. A. Struthers. Estimation of regression parameters in missing data problems. *The Canadian Jornal of Statistics*, 34(2):233–259, 2006. 13, 15

B. Nan, M. J. Emond, and J. A. Wellner. Information bounds for cox regression models with missing data. *The Annals of Statistics*, 32(2):723Ď1ď7753, 2004. 193

J. M. Neuhaus, A. J. Scott, and C. J. Wild. Family-specific approaches to the analysis of caseŰcontrol family data. *Biometrics*, 62:488–494, 2006. 29, 153, 157, 158, 213

J. M. Neuhaus, A. J. Scott, C. J. Wild, Y. Jiang, and C. E. McCulloch. Likelihood-based analysis of longitudinal data from outcome-dependent sampling designs. Unpublished, 2013. 157

A. O'Hagan and T. Leonard. Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 61:201–203, 1976. 96

A. B. Owen. Empirical likelihood ration confidence intervals for a single functional. *Biometrika*, 75:237–249, 1988. 16

A. B. Owen. Empirical likelihood for confidence regions. *Annals of Statistics*, 75:90–120, 1990. 16

A. B. Owen. *Empirical Likelihood*. Boca Raton: Chapman and Hall, 2001. 128

Y.-S. Park, J.-S. Ahn, H.-B. Kwon, and S.-P. Lee. Current status of dental caries diagnosis using cone beam computed tomography. *Imaging Science in Dentistry*, 41 (2):43–51, 2011. 120

C. W. Pattanayak, D. B. Rubin, and E. R. Zell. Propensity score methods for creating covariate balance in observational studies. *Revista Española de Cardiologia*, 64(10): 897–903, 2011. 129, 130

R. L. Prentice and R. Pike. Logistic disease incidence models with case-control studies. *Biometrika*, 66:403–411, 1979. 19

J. Qin. Empirical likelihood in biased sample problems. *Annals of Statistics*, 21(3): 1182–1196, 1993. 15

J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *Annals of Statistics*, 22(1):300–325, 1994. 15

Y. Qu and I. Lipkovich. Propensity score estimation with missing values using a multiple imputation missingness pattern (mimp) approach. *Statistics in Medicine*, 28:1402–1414, 2009. 151

R Development Core Team. R: A language and environment for statistical computing, 2011. URL `http://www.R-project.org/`. ISBN 3-900051-07-0. 55

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):846–867, 1994. 11, 21, 26, 28, 34, 113, 115, 126

J. M. Robins, F. Hsieh, and W. Newey. Efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society. Series B*, 57(2):409–424, 1995. 124, 189

R. Rohrer. Der index der koperfulle als mass des ernahrungszustandes. *Munich Med Wochenschr*, 68:580–583, 1921. 176

P. R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987. 124

P. R. Rosenbaum. A characterization of optimal designs for observational studies.

Journal of the Royal Statistical Society. Series B (Methodological), 53(3):pp. 597–610, 1991. 126

P. R. Rosenbaum. Propensity score. *In Encyclopedia of Biostatistics*, 5:3551–3555, 1998. 126

P. R. Rosenbaum. Ean exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B*, 67: 515–530, 2005. 128

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983. 122, 123, 124, 125, 127

P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984. 126, 128

P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39:33–38, 1985. 123, 128, 150

J. Rossouw, G. Anderson, R. Prentice, A. LaCroix, C. Kooperberg, M. Stefanick, R. Jackson, S. Beresford, B. Howard, K. Johnson, M. Kotchen, and J. Ockene. Risks and benefits of estrogen plus progestin in healthy postmenopausal womenŮprincipal results from the womenŠs health initiative randomized controlled trial. *J Am Med Assoc*, 288:321–333, 2002. 28, 53, 81

J. Rossouw, M. Cushman, P. Greenland, D. Lloyd-Jones, P. Bray, C. Kooperberg, M. Pettinger, J. Robinson, S. Hendrix, and J. Hsia. Inflammatory, lipid, thrombotic,

and genetic markers of coronary heart disease risk in the womenŠs health initiative trials of hormone therapy. *Arch Intern Med*, 168:2245–2253, 2008. 81, 82

D. Rubin and N. Thomas. Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52:249–264, 1996. 124

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics)*. John Wiley & Sons, Inc., 99 edition, 1987. 5

D. B. Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127:757–763, 1997. 123, 128

D. B. Rubin. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, 2: 169–188, 2001. 130, 140

D. B. Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36, 2007. 129, 141

C.-E. Sarndal. The calibration approach in survey sampling theory and practice. *Survey Methodology*, 33(2):99–119, 2007. 24, 25

A. J. Scott and C. J. Wild. Fitting logistic model under case-control or choice based sampling. *Journal of the Royal Statistical Society B*, 48(2):170–182, 1986. 19

A. J. Scott and C. J. Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):57–71, 1997. 18, 19, 29, 86, 183, 184, 188, 214

A. J. Scott and C. J. Wild. Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, 96:3–27, 2001. 8, 18, 19

A. J. Scott and C. J. Wild. Calculating efficient semiparametric estimators for a broad class of missing-data problems. *Festschrift for Tarmo Pukkila on his 60th Birthday Eds. E. P. Liski, J. Isotalo, J. NiemelĎ1d7 S. Puntanen, and G. P. H. Styan Dept. of Mathematics, Statistics and Philosophy, Univ. of Tampere*, pages 301–314, 2006. 18, 19, 20, 21, 90, 116, 157

A. J. Scott and C. J. Wild. Fitting regression models with response-biased samples. *Canadian Journal of Statistics*, 39:519–536, 2011. 27, 28, 31, 36, 37, 39, 40, 51, 53, 54, 65, 85, 156, 210, 211

C. Shen. Application of multiple imputation to data from two-phase sampling: estimation of the incidence rate of cognitive impairment. *Journal of Data Science*, 5: 503–518, 2007. 6

X. Shi, H. Zhu, and J. G. Ibrahim. Local influence for generalized linear models with missing covariates. *Biometrics*, 65(4):1164–74, 2009. 215

Y. Shieh. *Imputation methods on general linear mixed models of longitudinal studies.* Washington, DC.: Federal Committee on Statistical Methodology, 2003. 6

J. Song, T. R. Belin, M. B. Lee, X. Gao, and M. J. Rotheram-Borus. Handling baseline differences and missing items in a longitudinal study of hiv risk among runaway youths. *Health Services and Outcomes Research Methodology*, 2001:317–329, 1999. 150, 151

R. Song, H. Zhou, and M. R. Kosorok. A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika*, 96 (1):221–228, 2009. 88, 89, 90, 186

E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(9):1–21, 2010. 125

E. A. Stuart, M. Azur, C. Frangakis, and P. Leaf. Multiple imputation with large data sets: A case study of the childrenŠs mental health initiative. *American Journal of Epidemiology*, 155(9):1133–1139, 2009. 6

T. Stumer, S. Schneeweiss, K. J. Rothman, J. Avorn, and R. J. Glynn. Performance of propensity score calibration - a simulation study. *American Journal of Epidemiology*, 165(10):1110–1118, 2007. 151

M. T. Subbotin. On the law of frequency errors. *Mathematicheskii Sbornik*, 31:296–301, 1923. 93

Z. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97:661–682, 2010. 128

J. M. D. Thompson and E. A. Michell. Sex specific birthweight percentiles by gestational age for new zealand. *NZMJ*, 107:1–3, 1994. 176

A. Tsiatis. *Semiparametric Theory and Missing Data (Springer Series in Statistics)*. Springer, 2006. 115, 190, 192, 193

J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, 47:448–485, 1952. 109

I. Waernbaum. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Statistics in Medicine*, 31(15):1572–1581, 2011. 125

X. Wang and H. Zhou. A semiparametric empirical likelihood method for biased sampling schemes with auxiliary covariates. *Biometrics*, 62:1149–1160, 2006. 16, 20, 86, 88

X. Wang, Y. Wu, and H. Zhou. Outcome- and auxiliary-dependent subsampling and its statistical inference. *Journal of Biopharmaceutical Statistics*, 19(6):1132–1150, 2009. 17, 155

M. A. Weaver. *Semiparametric Methods for Continuous Outcome Regression Models with Covariate Data from an Outcome-Dependent Subsample*. PhD thesis, University of North Carolina at Chapel Hill, 2001. 46

M. A. Weaver and H. Zhou. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association*, 100(470):459–469, 2005. 12

J. White. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115:119–128, 1982. 6

C. J. Wild. Fitting logistic regression models in stratified case-control studies. *Biometrics*, 47:497–510, 1991. 32

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. 14, 15, 116

Z. Zhang and H. E. Rockette. Semiparametric maximum likelihood for missing covariates in parametric regression. *Annals of the Institute of Statistical Mathematics*, 58 (4):687–706, 2006. 189, 194

Z. Zhang and H. E. Rockette. An em algorithm for regression analysis with incomplete covariate information. *Journal of Statistical Computation and Simulation*, 34(2): 163–173, 2007. 14

Y. Zhao, J. F. Lawlles, and D. L. McLeish. Likelihood methods for regression models

with expensive variables missing by design. *Biometrical Journal*, 51(1):123–136, 2009. 13, 14, 15, 157

Y. Zhao, J. F.Lawless, and D. L. McLeish. Design and relative efficiency in two-phase studies. *Journal of Statistical Planning and Inference*, 142(11):2953–2964, 2012. 7, 27

H. Zhou, M. A. Weaver, J. Qin, M. P. Longnecker, and M. C. Wang. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 58(2):413–421, 2002. 16, 88, 89, 90, 186

H. Zhou, J. Chen, T. H. Rissanen, S. A. Korrick, H. Hu, J. T. Salonen, and M. P. Longnecker. Outcome-dependent sampling an efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology*, 18(4):461–468, 2007. 16

H. Zhou, R. Song, Y. Wu, and J. Qin. Statistical inference for a two-stage outcome-dependent sampling design with a continuous outcome. *Biometrics*, 67:194–202, 2011. 90, 155, 158

X. H. Zhou, E. J. Eckert, and W. M. Tierney. Multiple imputation in public health research. *Statistics in Medicine*, 20(9-10):1541–9, 2001. 6

H. Zhu, J. G. Ibrahim, and X. Shi. Diagnostic measures for generalized linear models with missing covariates. *Scand Stat Theory Appl*, 36(4):686–712, 2009. 215

H. Zhu, J. G. Ibrahim, and H. Cho. Perturbation and scaled cook's distance. *The Annals of Statistics*, 40(2):785–811, 2012. 215