Efficiency of Two-Phase Single and Multiple Response-Dependent Sampling Designs

by

© **Ananthika Nirmalkanna**

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics

Memorial University

December 2018

St. John's, Newfoundland and Labrador, Canada

Abstract

In some observational studies, it may be relatively more affordable to measure the response variable, while some covariates' data might be expensive to obtain. Therefore, collection of the covariate information might be restricted by the available budget of the study. In this situation, we need to consider sampling designs which yield powerful association tests for a given sample size by selecting more informative subjects. Using cost-efficient response-dependent two-phase sampling designs is a way to select more informative sampling units. In phase I, we have easily measured variables including the response variable for all individuals in the cohort or in a large random sample from the population, and in phase II, we obtain expensive variables for a subset of individuals selected according to their response variable and inexpensive covariates obtained in phase I. We consider the likelihood and pseudo-likelihood based methods for incomplete data analysis to make inference on the association between the expensive covariate and the response variable. We also consider multiple response-dependent sampling designs. The objective is to compare the efficiency of estimators and to identify efficient sampling design settings under each estimation method.

I dedicate this thesis to my father Ratnam Raveendranathan.

Lay summary

In genetic association studies, the aim is to identify the relation between genetic variants and phenotypes or traits. For example, using genetic association studies, we can identify genes involved in human disease. In such studies, it may be relatively more affordable to measure the trait values such as blood pressure, while a genetic variant data might be expensive to obtain. Therefore, collection of the genetic variant information might be restricted by the available budget of the study. In this situation, we need to find a way to select informative individuals based on their available trait values to collect genetic variant information. To obtain informative results on the relation between genetic variants and trait values, we investigate the best way for selecting individuals based on their trait values to obtain the genetic variants information. We consider different statistical methods to estimate the association between genetic variants and trait values, and we compare the performance of each estimation method.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Dr. Yildiz Yilmaz for her invaluable support, guidance and encouragement throughout my research. This thesis would not have been possible without her guidance, advices, criticisms and most importantly without her care.

Special thanks to my co-supervisor Dr. Candemir Cigsar who guide me to get into this degree. I shall eternally be grateful to him for his valuable advice and guidance throughout my study.

I would like to express my gratitude to all the professors at Memorial University of Newfoundland who taught me subjects during my program.

I am also grateful to the School of Graduate Studies and the Department of Mathematics and Statistics, my supervisor and my co-supervisor for their financial support in the form of Graduate Fellowship and Teaching Assistantship.

I would like to thank my husband Nirmalkanna for his unconditional love and support during my study at university and my parents Raveendranathan and Indra for a lifetime of encouragement.

Last but not least, it is my great pleasure to thank my family and friends for their support throughout these years.

Statement of contribution

Dr. Yildiz Yilmaz proposed the research question that was investigated throughout this thesis. The overall study was jointly designed by Dr. Yildiz Yilmaz, Dr. Candemir Cigsar, and Ananthika Nirmalkanna. The simulation study was conducted and the manuscript was drafted by Ananthika Nirmalkanna. Dr. Yildiz Yilmaz and Dr. Candemir Cigsar supervised the study and contributed to the final manuscript.

Table of contents

List of tables

List of Symbols

N	Cohort size
N_{cases}	Number of cases within cohort
$N_{control}$	Number of control within cohort
n	Sample size
n_{cases}	Number of cases selected in the sample
$n_{control}$	Number of controls selected in the sample
R_i	Phase II sampling indicator for individual i
V	The set of completely observed data
$\bar{V}$	The set of units with incomplete data
π_i	Probability of selecting the i^{th} unit in phase II sample
S_j	The j^{th} stratum $(j = 1, 2, \ldots, K)$
N_j	Number of units within the j^{th} stratum $(j = 1, 2, \ldots, K)$
n_j	Number of units sampled from the j^{th} stratum $(j = 1, 2, \ldots, K)$
C_{k-1}	Lower cut-point value of the k^{th} stratum
C_k	Upper cut-point value of the k^{th} stratum
δ_{ij}	Phase II sampling indicator of unit i being in the j^{th} stratum
D_j	The set of indices of all fully observed units in stratum S_j
N_m	Number of units observed in the m^{th} level of inexpensive covariate $(m = 0, 1, \ldots, M-1)$
n_m	Number of units sampled from the m^{th} level of inexpensive covariate $(m = 0, 1, \ldots, M-1)$
S_{mj}	The j^{th} stratum of m^{th} level of inexpensive covariate $(m = 0, 1, \ldots, M-1,\ j = 1, 2, \ldots, K)$
N_{mj}	Number of units in the j^{th} stratum of m^{th} level of inexpensive covariate $(m = 0, 1, \ldots, M-1,\ j = 1, 2, \ldots, K)$

n_{mj}	Number of units sampled from the j^{th} stratum of the m^{th} level of inexpensive covariate $(m = 0, 1, \ldots, M-1,\ j = 1, 2, \ldots, K)$
C_{mj-1}	Lower cut-point value of the j^{th} stratum of the m^{th} level of inexpensive covariate $(m = 0, 1, \ldots, M-1,\ j = 1, 2, \ldots, K)$
C_{mj}	Upper cut-point value of the j^{th} stratum of the m^{th} level of inexpensive covariate $(m = 0, 1, \ldots, M-1,\ j = 1, 2, \ldots, K)$
δ_{imj}	Phase II sampling indicator of unit i being in the j^{th} stratum of the m^{th} level of inexpensive covariate
D_{mj}	The set of indices of all fully observed units in stratum S_{mj}
R_{ij}	Phase II sampling indicator for individual i based on the j^{th} response variable
π_{ij}	Probability of selecting the i^{th} unit in phase II sample depending on the j^{th} response variable
n_j	Number of units sampled depending on the j^{th} response variable $(j = 1, 2, \ldots, J)$
S_{jk}	The k^{th} stratum of the j^{th} response variable $(j = 1, 2, \ldots, J,\ k = 1, 2, \ldots, K)$
N_{jk}	Number of units in the k^{th} stratum of the j^{th} response variable $(j = 1, 2, \ldots, J,\ k = 1, 2, \ldots, K)$
n_{jk}	Number of units sampled from the k^{th} stratum depending on the j^{th} response variable $(j = 1, 2, \ldots, J,\ k = 1, 2, \ldots, K)$
C_{jk-1}	Lower cut-point value of the k^{th} stratum based on the j^{th} response variable $(j = 1, 2, \ldots, J,\ k = 1, 2, \ldots, K)$
C_{jk}	Upper cut-point value of the k^{th} stratum based on the j^{th} response variable $(j = 1, 2, \ldots, J,\ k = 1, 2, \ldots, K)$
δ_{ijk}	Phase II sampling indicator of unit i depending on the j^{th} response from the k^{th} stratum

List of abbreviations

BSS	Basic Stratified Sampling
CPP	Collaborative Perinatal Project
FLM	Full Likelihood-based Method
IPW	Inverse Probability Weighting
MLE	Maximum Likelihood Estimation
MSE	Mean Square Error
PE	Point Estimate
SE	Standard Error
SNHL	Sensorineural Hearing Loss
SNP	Single Nucleotide Polymorphism

Chapter 1

Introduction

In various studies investigating the association between a response variable and some covariates is the primary interest. In some observational studies, it may be relatively more affordable to measure the response variable, while a covariate's data might be expensive to obtain. Therefore, collection of the covariate information might be restricted by the available budget of the study. For example, in genetic association studies, where the aim is to identify genetic markers associated with a phenotype or a trait, although trait values might be available for each individual in a cohort, it might be expensive to obtain genetic data (Yilmaz and Bull, 2011; Barnett et al., 2013). The genetic data could only be obtained for a limited number of individuals. The sample size is determined based on the available budget. Thus, we might need to select a sample with a given size from the cohort to collect the genetic data. In many genetic association studies, due to the multiple testing issue, the power of the association tests is a main concern. These studies require a much larger sample size to achieve adequate statistical power. However, when the sample size is fixed based on the available budget, we can increase the power of the association tests for a given sample size by selecting more informative subjects. Response-dependent two-phase sampling design is a way to select more informative sampling units. For example, Zhou et al. (2007) showed the improved statistical efficiency obtained by using response-dependent two-phase sampling design compared to simple random sampling. Yilmaz and Bull (2011) suggested that given a fixed sample size, trait-dependent sampling designs can lead to more powerful association tests compared to simple random sampling design. Also, quantitative trait-dependent sampling can be quite effective in reducing costs of

genetic data collection. Many other studies including Allison (1997), Page and Amos (1999), Slatkin (1999), Xiong et al. (2002), Chen et al. (2005), Wallace et al. (2006), Huang and Lin (2007), Li et al. (2011), Chen and Li (2011) suggested that sampling subjects with the extreme trait values can substantially increase statistical power compared to simple random sampling.

The objectives of this thesis are to investigate efficiency of the response-dependent two-phase sampling designs and to identify informative sampling designs which give efficient estimates of the coefficient of an expensive covariate for the given sample size. We consider different estimation methods including likelihood-based methods and pseudo-likelihood based methods. We assess the asymptotic properties of the estimation methods under different response-dependent sampling designs and compare their efficiency under each design. We also discuss response-dependent two-phase sampling designs under bivariate response variable models.

1.1 Response-Dependent Sampling

Sampling designs where sampling probabilities for subjects depend on their response values are called response-dependent sampling designs. They are retrospective designs. There are different response-dependent sampling designs based on the type of the response variable. For example, case-control design (Breslow and Cain, 1988) is a well-known response-dependent sampling design where the response variable is defined as a case or a control (i.e., a binary variable). Observational epidemiology study designs often investigate the relationship between a disease outcome and an exposure given other characteristics. The disease outcome might be known for a large number of subjects. However, the covariate data can only be collected for a small number of subjects since collecting the covariate information might be expensive.

In a case-control study, the response variable (disease status) is binary where cases have the disease and controls are free of the disease. Suppose there are N individuals in a cohort in which N_{cases} of them are cases and $N_{control}$ of them are controls. Suppose the budget allows the covariate information to be collected for n subjects. Then, one may collect n_{cases} individuals from the cases group and $n_{control}$ individuals from the control group where $n_{cases} + n_{control} = n$, and measure expensive covariate for all n sampled units. When the cases are rare, we can select all the cases and select a

random sample of controls to measure the expensive covariate.

In the studies of investigating the association between an exposure measure and a continuous response variable, a common approach is to discretize the response variable values using some cut-point values which leads to the mutually exclusive intervals called strata. Then, a basic stratified sampling (BSS) can be conducted (Lawless et al., 1999).

1.2 Response-Dependent Two-Phase Sampling Designs

In a response-dependent two-phase sampling design, in phase I, we have easily measured variables including the response variable for all individuals in a cohort or in a large random sample from the population, and in phase II, we obtain expensive covariate(s) for a subset of individuals selected according to their response variable obtained in phase I (Neyman, 1938; Zhao and Lipsitz, 1992).

The reason for applying a response-dependent two-phase sampling design is to enhance the efficiency of estimates with a limited budget. In a study of the relationship between a rare exposure and a rare disease, White (1982) proposed a stratified response-dependent two-phase design. They obtained the value of a response variable which is a disease status and a binary (inexpensive) covariate, which is an exposure status, for a large sample in phase I, but in phase II, another covariate's (expensive covariate) data were collected only for a subsample. The subsample was selected by sampling from each of the four groups constructed based on the disease and the inexpensive exposure status. Efficiency of this design was enhanced by sampling a large proportion of the subjects from the small groups and a smaller proportion of subjects from the large groups.

Variations of response-dependent two-phase sampling designs and estimation methods have been proposed. Breslow and Cain (1988) considered the preliminary sample to be separate samples of cases and controls. These samples are selected from a subpopulation of diseased and non-diseased subjects. They proposed a modified logistic regression model for the analysis of data obtained through a two-stage case-control design. They considered conditional maximum likelihood estimation

under the logistic regression model, which was developed for choice-based data by Manski and McFadden (1981) and Hsieh et al. (1985). A choice-based sampling is a stratified sampling scheme where each stratum is defined according to the response variable which is a discrete variable. In this sampling scheme, the sampling probabilities for rare response categories are high and the size of the selected sample from each response category is predetermined.

Prentice (1986) considered a different type of case-control study within a cohort study which is a case-cohort design. In standard case-cohort studies, the covariate values are obtained for all the cases and for a sub-cohort which is randomly chosen from the whole cohort. This approach is used when the event of interest is rare. When the event of interest is not rare, a generalized case-cohort design could be considered (Chen, 2001). In a generalized case-cohort design, expensive covariate values are obtained for a random sample from the case group and for a sub-cohort which is randomly chosen from the whole cohort. Here, the response variable is time-to-disease event. Prentice performed the estimation by maximizing a Cox-type likelihood under a proportional hazard regression model of time-to-disease event.

Zhao and Lipsitz (1992) studied a class of twelve possible designs within the framework of two-phase designs and considered three different statistical methods. They considered three sampling schemes to select phase I sample and four sampling schemes to select phase II samples. In phase I, disease status D and exposure status Z were available for all individuals, and in phase II, they gathered information about the confounders X on selected individuals. In phase I, they selected individuals randomly from a population and obtained the (D, Z) values for the selected individuals, selected individuals based on their D values and obtained Z values for the selected individuals or selected individuals based on their Z values and obtained D values for the selected individuals. In phase II, they selected a random sample, selected individuals based on their D values, selected individuals based on their Z values or selected individuals based on their D and Z values. The confounder information X was gathered for the individuals selected in phase II sample. The sampling designs were formed by choosing one of the three phase I designs and one of four phase II sampling designs. They analyzed the efficiency of the three estimation methods under these twelve sampling designs. The three estimation methods considered were the maximum likelihood estimation, Breslow and Cain's method, and an estimating equation method.

The studies discussed above were on the sampling designs when the response variable is binary except the case-cohort design. The response variable could also be a continuous variable. For example, Gray et al. (2005) studied the association between utero exposure to background levels of polychlorinated biphenyls with the cognitive functioning among school-age children. In this study, individuals were selected from Collaborative Perinatal Project (CPP) participants. The CPP is a prospective cohort study designed to identify determinants of neurodevelopmental deficits in children. There were 55,908 pregnancies recruited into the CPP study from 1959 to 1966. The polychlorinated biphenyls levels measured through blood serum assay where the blood samples were collected during the pregnancy and stored at -20^oC. The cost associated with blood serum assay is high, therefore, 732 women were selected at random and additional 162 chosen women whose children had either a low or high intelligence quotient score.

In whole-genome association studies to identify rare and common variants associated with complex traits, Yilmaz and Bull (2011) assumed an existing cohort of unrelated individuals, well phenotyped for a quantitative trait in phase I and discussed four different sampling designs to select phase II sample, whose sample size is the same in all four designs. Half of the cohort was selected in phase II. In design I, they selected a simple random sample. In design II, all observations in each of the 25% tail of the quantitative trait distribution were selected that is called extreme sampling where the individuals having the highest or the lowest quantitative trait values are selected. In design III, they selected all observations in each of the 20% tail and the central 10% of quantitative trait distribution. Designs II and III are systematic and do not actually involve any random selection. Design IV considered 50% sample by distance from the median of the quantitative trait distribution. That is, the selection probabilities for individuals are inversely weighted according to their distance from the median of the trait distribution in the cohort. Their simulation study results suggest that the quantitative trait-dependent sampling designs (Designs II-IV) generally produce greater than 50% relative efficiency compared to using the entire cohort. The quantitative trait-dependent sampling designs with oversampling or complete selection of the extremes of the distribution are the most efficient designs. They concluded that in comparison to using the entire cohort, extreme sampling acquires some loss of efficiency and power to detect the association, but compared to simple random sampling it is far more efficient. Also, they suggested that quantitative

trait-dependent sampling can be effective in reducing the cost to collect genetic data. Barnett et al. (2013) also concluded both analytically and numerically that sampling individuals with extreme phenotypes can increase the power for identifying causal rare variants compared to random sampling.

1.2.1 Response-Dependent Two-Phase Sampling Design Setting

Suppose Y denotes a continuous response variable, and X and Z are the expensive covariate and the inexpensive covariate, respectively. We are interested in measuring the association between Y and X while adjusting the model for Z. Now, suppose Y, X and Z are observed for all units and the observations $\{(y_i, x_i, z_i),\ i = 1, 2, \ldots, N\}$ in a cohort of size N are generated from

$$f(y|x, z; \boldsymbol{\theta})g(x|z; \boldsymbol{\alpha})h(z) \tag{1.1}$$

where $f(y|x, z; \boldsymbol{\theta})$ is the conditional density function of Y given X and Z, $\boldsymbol{\theta}$ is a vector of unknown parameters including the regression coefficients of interest, $g(x|z; \boldsymbol{\alpha})$ is the conditional density or mass function of X given $Z = z$, and $h(z)$ is the marginal density or mass function of Z. Neither $g(x|z; \boldsymbol{\alpha})$ nor $h(z)$ depends on $\boldsymbol{\theta}$. Let $G(.)$ denote the conditional distribution function corresponding to $g(.)$.

Now, suppose that Y and Z are observed for all the N units in phase I. However, the expensive covariate X can only be observed for a subsample of individuals selected in phase II for a fixed sample size n. Let R_i be an indicator function where $R_i = 1$ if unit i in phase I cohort is selected for inclusion in phase II sample, and $R_i = 0$ otherwise. Thus, the observed data consist of N units where $n = \sum_{i=1}^{N} R_i$ units in the cohort provide complete data (y_i, x_i, z_i), and $N - n$ of units provide information only on response values y_i and inexpensive covariate z_i. The set of completely observed data, V, and the set of units with incomplete data, $\bar{V}$, are denoted by

$$\begin{aligned} V &= \{i : R_i = 1, i = 1, 2, \ldots, N\} \text{ and} \\ \bar{V} &= \{i : R_i = 0, i = 1, 2, \ldots, N\}, \end{aligned} \tag{1.2}$$

respectively.

Let π_i denote the probability of selecting the i^{th} unit in phase II sample. It depends on the value of the response variable Y_i in a response-dependent sampling design (possibly in addition to Z_i). We assume that the covariate X is "missing at random" in the terminology of Little and Rubin (1987). Thus, the sampling probability for unit i is

$$\pi_i = P(R_i = 1|y_i, z_i, x_i) = P(R_i = 1|y_i, z_i) \tag{1.3}$$

for $i = 1, 2, \ldots, N$.

1.2.2 Response-Dependent BSS Design

The BSS design is a well known sampling design. In BSS design, it is assumed that the number of units in a cohort are divided into mutually exclusive groups called strata and a subsample with a given size is randomly selected from each strata without replacement. Suppose that N units are generated independently from model (1.1) and that the numbers N_j in the j^{th} stratum S_j for $j = 1, 2, \ldots, K$ are observed. Then, from each stratum S_j, n_j units are randomly selected without replacement.

In response-dependent BSS design, suppose there are N units generated from model (1.1) in phase I and the observed response values $(y_1, y_2, \ldots, y_N)$ is partitioned into K number of strata, S_k, $k = 1, 2, \ldots, K$, using fixed cut-point values C_i, $i = 1, 2, \ldots, K-1$ where $C_1 \leq C_2 \leq \ldots \leq C_{K-1}$. The first stratum includes the units with y_i values less than C_1, the k^{th} $(k = 2, \ldots, K-1)$ stratum includes the units with y_i values between C_{k-1} and C_k, and the K^{th} stratum includes the units with y_i values greater than C_{k-1}. Let N_j $(j = 1, 2, \ldots, K)$ be the number of units in each stratum where $\sum_{j=1}^{K} N_j = N$. Then, from each stratum S_j, n_j units are randomly selected for inclusion in the phase II sample, where the total phase II sample size $n = \sum_{j=1}^{K} n_j$ is fixed according to budgetary constraints.

The selection probability for the i^{th} unit, π_i in (1.3), becomes $\pi_i = \sum_{j=1}^{K} \delta_{ij} \frac{n_j}{N_j}$ for

$i = 1, 2, \ldots, N$, where

$$\delta_{ij} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ unit is in the } j^{\text{th}} \text{ stratum,} \\ 0 & \text{if the } i^{\text{th}} \text{ unit is not in the } j^{\text{th}} \text{ stratum.} \end{cases} \tag{1.4}$$

1.3 Stratified Response-Dependent Two-Phase Sampling Design

In a two-phase stratified response-dependent sampling design, in phase I, we have easily measured variables including the response variable and inexpensive covariates for all individuals in a cohort or in a large random sample from the population, and in phase II, we obtain expensive covariate(s) for a subset of individuals selected based on their response variable and inexpensive covariates obtained in phase I. The stratified response-dependent two-phase sampling design might yield more efficient estimators than sampling that depends only on the response variable (Espin-Garcia et al., 2018).

1.4 Estimation Methods

Estimation methods under response-dependent two-phase sampling designs were considered in different settings. For the response-dependent two-phase sampling designs with a binary or categorical response variable, Breslow and Cain (1988) developed a conditional likelihood estimation method; Flanders and Greenland (1991) and Zhao and Lipsitz (1992) proposed a weighted likelihood approach; Breslow and Holubkov (1997) studied a nonparametric maximum likelihood estimation method; Wacholder and Weinberg (1994) obtained a maximum likelihood estimation using an EM-algorithm.

For response-dependent two-phase sampling designs with a continuous response variable, Lawless et al. (1999) proposed semiparametric likelihood and pseudo-likelihood methods for estimating $\boldsymbol{\theta}$ in (1.1) when the phase II sampling depends on the response variable only. They considered the situations in which units generated are not fully observed and modeling the covariate distribution is not possible. They presented theoretical asymptotic results for the estimators and

handled the problems from the response-dependent sampling, measurement error, and the missing data literature under a single framework. Zhou et al. (2002) developed a likelihood based approach that is a semiparametric empirical likelihood method. Chatterjee et al. (2003) proposed a pseudo-score estimation, and Weaver and Zhou (2005) proposed an estimated likelihood approach which are pseudo-likelihood methods.

Some likelihood-based and pseudo-likelihood based estimation methods proposed for response-dependent two-phase sampling designs are described in the following.

1.4.1 Likelihood-based Methods

Suppose the complete data generated from model (1.1) is observed. The likelihood of the complete data $\{(y_i, x_i, z_i),\ i = 1, 2, \ldots, N\}$ is proportional to

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(y_i|x_i, z_i; \boldsymbol{\theta}). \tag{1.5}$$

The maximum likelihood estimate of $\boldsymbol{\theta}$ is obtained by maximizing $L(\boldsymbol{\theta})$ in (1.5).

Full Likelihood Method

Consider the response-dependent two-phase sampling described in Section 1.2. Assume the response variable Y and inexpensive covariate Z have been fully observed for all N units, in phase I. A phase II sample is selected from the phase I units with selection depending upon Y only. The expensive covariate X is observed for only selected units. If Z is not of low dimension then a parametric model for X given Z, $g(x|z, \boldsymbol{\alpha})$ is essential to consider (Lawless, 2018). The full likelihood (Robins et al., 1995) which incorporates both complete and incomplete data $\{(y_i, x_i, z_i) : i \in V\} \cup \left\{(y_i, z_i) : i \in \bar{V}\right\}$ under the missing at random assumption is proportional to

$$L_R(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \prod_{i \in V} f(y_i|x_i, z_i; \boldsymbol{\theta}) g(x_i|z_i; \boldsymbol{\alpha}) \prod_{i \in \bar{V}} \int_u f(y_i|z_i, u; \boldsymbol{\theta}) g(u|z_i; \boldsymbol{\alpha}) du, \tag{1.6}$$

where V and $\bar{V}$ are defined in (1.2). The maximum likelihood estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_R$, is obtained by maximizing $L_R(\boldsymbol{\theta}, \boldsymbol{\alpha})$ in (1.6) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. If Z is not

continuous or high-dimensional, $g(x|z)$ can be modelled nonparametrically (Lawless, 2018).

If there is no inexpensive covariate, $L_R(\boldsymbol{\theta}, g)$ in (1.6) reduces to

$$L_R(\boldsymbol{\theta}, g) = \prod_{i \in V} f(y_i|x_i; \boldsymbol{\theta}) g(x_i) \prod_{i \in \bar{V}} \int_u f(y_i|u; \boldsymbol{\theta}) g(u) du. \tag{1.7}$$

The semiparametric maximum likelihood estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_R$, is obtained by maximizing $L_R(\boldsymbol{\theta}, g)$ in (1.7) with respect to $\boldsymbol{\theta}$ and g.

When we know the values of y_i for all units in phase I, the likelihood function is $L_R(\boldsymbol{\theta}, \boldsymbol{\alpha})$. Suppose under the response-dependent BSS design explained in Section 1.2.2, only the stratum information is available for unselected units. The likelihood which incorporates complete data $\{(y_i, x_i, z_i) : i \in V\}$ and stratum information of unselected data is proportional to

$$L_F(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \prod_{j=1}^{K} \left\{ \prod_{i \in D_j} f(y_i|x_i, z_i; \boldsymbol{\theta}) g(x_i|z_i; \boldsymbol{\alpha}) \right\} Q_j(\boldsymbol{\theta}, \boldsymbol{\alpha})^{N_j - n_j}, \tag{1.8}$$

where

$$Q_j(\boldsymbol{\theta}, \boldsymbol{\alpha}) = pr\{(Y, Z, X) \in S_j\} = \int pr\{(Y, z, x) \in S_j|x, z\} g(x|z; \boldsymbol{\alpha}) h(z) dx,$$

and $D_j = \{i : \delta_{ij} = 1, R_i = 1\}$ denotes the set of indices of all fully observed units in stratum S_j and δ_{ij} is defined in (1.4). The maximum likelihood estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_F$, is obtained by maximizing the likelihood function $L_F(\boldsymbol{\theta}, \boldsymbol{\alpha})$ in (1.8) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$.

If there is no inexpensive covariate, $L_F(\boldsymbol{\theta}, g)$ in (1.8) reduces to (Lawless et al., 1999)

$$L_F(\boldsymbol{\theta}, g) = \prod_{j=1}^{K} \left\{ \prod_{i \in D_j} f(y_i|x_i; \boldsymbol{\theta}) g(x_i) \right\} Q_j(\boldsymbol{\theta}, g)^{N_j - n_j}, \tag{1.9}$$

where

$$Q_j(\boldsymbol{\theta}, g) = pr\{(Y, X) \in S_j\} = \int pr\{(Y, x) \in S_j|x\} dG(x).$$

The semiparametric maximum likelihood estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_F$, is obtained by

maximizing the likelihood function $L_F(\boldsymbol{\theta}, g)$ in (1.9) with respect to $\boldsymbol{\theta}$ and g.

When there are no inexpensive covariates Z, if X is discrete with relatively few points of support then we can obtain the maximum likelihood estimator of $\boldsymbol{\theta}$ and G by maximizing the likelihoods $L_R(\boldsymbol{\theta}, g)$ or $L_F(\boldsymbol{\theta}, g)$ (Scott and Wild, 1991). This is not possible when the range of possible values for X is not small or X is continuous. In this study, we assume that X is a binary variable. Thus, we can apply the maximum likelihood estimation.

Conditional Likelihood Method

Suppose that only completely observed units information is available. Thus, the observed data is $\{(y_i, x_i, z_i) : i \in V\}$. Then, the resulting likelihood conditioning on being selected in phase II is given by (Carroll et al., 1995; Lawless et al., 1999)

$$\begin{aligned} L_{C0}(\boldsymbol{\theta}, \boldsymbol{\alpha}) &= \prod_{i:R_i=1} pr\left\{(y_i, x_i, z_i)|R_i = 1\right\} \\ &= \prod_{i:R_i=1} \left\{ \frac{f(y_i|x_i, z_i; \boldsymbol{\theta}) g(x_i|z_i; \boldsymbol{\alpha}) h(z_i) \pi_i}{\sum_{j=1}^{K} \frac{n_j}{N_j} Q_j(\boldsymbol{\theta}, \boldsymbol{\alpha})} \right\}, \end{aligned} \tag{1.10}$$

where

$$Q_j(\boldsymbol{\theta}, \boldsymbol{\alpha}) = pr\{(Y, X, Z) \in S_j\} = \int pr\{(Y, z, x) \in S_j|x, z\} g(x|z; \boldsymbol{\alpha}) h(z) dx,$$

L_{C0} in (1.10) is equivalent to L_{C1} which is a conditional likelihood that does not involve G. L_{C1} is given by (Wild, 1991; Scott and Wild, 1997, 1998; Lawless et al.,

1999)

$$
\begin{aligned}
L_{C1}(\boldsymbol{\theta}) &= \prod_{i:R_i=1} pr\left\{(y_i, x_i, z_i)|R_i = 1\right\} \\
&= \prod_{i:R_i=1} \left\{ \frac{f(y_i|x_i, z_i; \boldsymbol{\theta})\pi_i}{\sum_{l=1}^{K} \frac{n_l}{N_l} Q_l^*(x_i, z_i; \boldsymbol{\theta})} \right\},
\end{aligned} \tag{1.11}
$$

where $Q_l^*(x, z; \boldsymbol{\theta}) = pr\{(Y, z, x) \in S_l|x, z\}$.

The semiparametric profile likelihood L_{C1} for $\boldsymbol{\theta}$ is obtained from L_{C0} using the maximization process employed in Wild (1991) and Scott and Wild (1997, 1998). That is, $L_{C1}(\boldsymbol{\theta}) = L_{C0}(\boldsymbol{\theta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}))$ where $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ is obtained by maximizing L_{C0} with respect to $\boldsymbol{\alpha}$ for fixed $\boldsymbol{\theta}$ over the space of all discrete distributions whose support includes the observed values of X.

The estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{C1}$, is obtained by maximizing $L_{C1}(\boldsymbol{\theta})$ in (1.11) with respect to $\boldsymbol{\theta}$.

If there are no inexpensive covariate, $L_{C1}(\boldsymbol{\theta})$ in (1.11) reduces to

$$
\begin{aligned}
L_{C1}(\boldsymbol{\theta}) &= \prod_{i:R_i=1} pr\left\{(y_i, x_i)|R_i = 1\right\} \\
&= \prod_{i:R_i=1} \left\{ \frac{f(y_i|x_i; \boldsymbol{\theta})\pi_i}{\sum_{l=1}^{K} \frac{n_l}{N_l} Q_l^*(x_i; \boldsymbol{\theta})} \right\},
\end{aligned} \tag{1.12}
$$

where $Q_j^*(x; \boldsymbol{\theta}) = pr\{(Y, x) \in S_j|x\}$.

1.4.2 Pseudo-likelihood Methods

By replacing the empirical estimate for the function G in the likelihood function (1.7), the pseudo-likelihood can be formulated. This likelihood is also known as an estimated likelihood. The estimates of the parameters can be obtained by solving the

estimating equation for $\boldsymbol{\theta}$. Here, the estimating function is the first derivative of the log of pseudo-likelihood.

Estimated Pseudo-likelihood

Suppose there is no inexpensive covariate Z. The estimated pseudo-likelihood method is obtained by replacing G in the likelihood function (1.9) with an empirical estimate $\tilde{G}$. The estimate $\tilde{G}$ of G is the empirical cumulative distribution function of X obtained by

$$\tilde{G}(x) = \sum_{j=1}^{K} \tilde{G}_j(x) \frac{N_j}{N}, \tag{1.13}$$

where $\tilde{G}_j(x)$ is the empirical cumulative distribution function obtained by using the x_i values for units $i \in D_j$.

The log-pseudo-likelihood function is obtained by inserting $\tilde{G}(x)$ in (1.13) into the logarithm of likelihood $L_F(\boldsymbol{\theta}, G)$ in (1.9) and is given by (Lawless et al., 1999)

$$l_P(\boldsymbol{\theta}) = \sum_{i:R_i=1} \log f(y_i|x_i;\boldsymbol{\theta}) + \sum_{j=1}^{K} (N_j - n_j) \left\{ \log \sum_{l=1}^{K} \frac{N_l}{n_l} \sum_{i \in D_l} Q_j^*(x_i;\boldsymbol{\theta}) \right\}, \tag{1.14}$$

where $Q_j^*(x;\boldsymbol{\theta}) = pr\{(Y,x) \in S_j|x\}$. The estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_P$, is obtained by maximizing $l_P(\boldsymbol{\theta})$ in (1.14) with respect to $\boldsymbol{\theta}$.

Inverse Probability Weighting (IPW) Method

Suppose that only completely observed units' information is available, then the pseudo-likelihood estimating function is obtained by weighting the contributions of completely observed units inversely according to their probability of selection. This is called as the IPW method. The IPW method uses the Horvitz-Thompson approach in problems involving response-dependent sampling. The resulting estimating function

is given by (Robins et al., 1994)

$$S_w(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_i(\boldsymbol{\theta}), \tag{1.15}$$

where π_i is defined in (1.3), and

$$U_i(\boldsymbol{\theta}) = \frac{\partial \log f(y_i|x_i, z_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{1.16}$$

If there are no inexpensive covariate, $U_i(\boldsymbol{\theta})$ in (1.16) reduces to

$$U_i(\boldsymbol{\theta}) = \frac{\partial \log f(y_i|x_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

To obtain the estimate of $\boldsymbol{\theta}$, we solve the estimating equation $S_w(\boldsymbol{\theta}) = 0$. The IPW method is potentially less efficient because only the completely observed units information is used in the estimating equation and units with incomplete observation are not considered. In this estimation method, we can get the estimates without making any model assumption on the expensive covariate distribution.

Under some regularity conditions (White, 1982), the estimator $\hat{\boldsymbol{\theta}}$ obtained by solving estimating equations is a consistent estimator of the true parameter vector $\boldsymbol{\theta}$. In addition, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed with zero mean vector and covariance matrix

$$C(\boldsymbol{\theta}) = A(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}) [A(\boldsymbol{\theta})^{-1}]', \tag{1.17}$$

where

$$A(\boldsymbol{\theta}) = -\frac{1}{n} \left(\frac{\partial S_w(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right),$$

$$B(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{N} S_{wi}(\boldsymbol{\theta}) S_{wi}(\boldsymbol{\theta})'.$$

A consistent estimator of $C(\boldsymbol{\theta})$ is obtained by inserting $\hat{\boldsymbol{\theta}}$ instead of $\boldsymbol{\theta}$ as given in the following

$$C(\hat{\boldsymbol{\theta}}) = A(\hat{\boldsymbol{\theta}})^{-1} B(\hat{\boldsymbol{\theta}}) [A(\hat{\boldsymbol{\theta}})^{-1}]'. \tag{1.18}$$

1.5 Multiple Response-Dependent Two-Phase Sampling Design

Many studies discussed and developed response-dependent two-phase sampling designs based on a univariate continuous response variable. In practice, the data may consist of multiple response variables which are of interest. In some epidemiological cohort studies, the response variable values are recorded for multiple members of families, and in some clinical trials, study individuals may experience multiple events. A common feature among all these studies is that response variables might be correlated. As the field of epidemiology expands and evolves, an increasing number of studies have been conducted using the multiple response-dependent sampling designs. For example, Longnecker et al. (2004) studied the association between polychlorinated biphenyls levels in maternal pregnancy serum and audiometrically determined hearing thresholds among offspring when they were approximately 8 years old. Subjects were selected from the Collaborative Perinatal Project (CPP), where eligible children met the criteria of live-born singleton and 3-ml third trimester maternal serum specimen was available. Samples were selected using a multiple response-dependent sampling design. From the eligible children, 1200 subjects were selected at random, of whom 726 had an 8-year audiometric evaluation. To get the additional sample, they defined sensorineural hearing loss (SNHL) as a hearing threshold $\geq$ 13.3dB, based on the average across both ears at 1000, 2000, and 4000 Hz, in conjunction with no evidence of conductive hearing loss. Evidence of conductive hearing loss was defined by the air-bone difference in hearing threshold being $\geq$ 10 dB, based on the average across both ears at 1000, 2000, and 4000 Hz. An additional 200 eligible children were randomly selected from the 440 children whose 8-year audiometric evaluation showed SNHL.

Sampling designs under the multiple response-dependent sampling framework and statistical methods accounting for the multiple response-dependent sampling method is still underdeveloped. Therefore, new and efficient development of sampling methods under the multiple response-dependent sampling design is needed. In addition, efficient statistical inference procedures are needed in order to take advantage of limited data under the multiple response-dependent sampling design. In our study, we will investigate inference procedures under the multiple response-dependent sampling

designs and identify the efficient sampling designs. In this study, copula functions are used to model the dependence between response variables. An introduction to copula modeling is given in the following section.

1.6 Copula Models

Copulas are functions used to construct a joint distribution function by combining the marginal distributions (Nelsen, 2006). For simplicity, consider a pair of random variables Y_1 and Y_2, with marginal distribution functions $F_1(y_1) = P(Y_1 \leq y_1)$ and $F_2(y_2) = P(Y_2 \leq y_2)$, respectively, and a joint distribution function $F(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2)$. A result due to Sklar (1959) says there is for any $F(y_1, y_2)$ a unique copula function $C(u_1, u_2),\ 0 < u_1, u_2 < 1$, such that

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2))$$

for all $y_1, y_2 \in \mathbb{R}^2$.

The function $C(u_1, u_2)$ is termed a copula, and it is a joint cumulative distribution function on the unit square, with marginal distributions that have Uniform distribution on $(0, 1)$. A common model used in many applications is the Clayton copula model (Clayton, 1978) given by:

$$C_\phi(u_1, u_2) = (u_1^{-\phi} + u_2^{-\phi} - 1)^{-\frac{1}{\phi}},\ \phi > 0, \tag{1.19}$$

where the parameter ϕ specifies the dependence level between U_1 and U_2.

A frequently used measure of dependence is Kendall's tau. Kendall's tau is the probability of concordance minus the probability of discordance, that is

$$\tau = P((Y_{1i} - Y_{1j})(Y_{2i} - Y_{2j}) > 0) - P((Y_{1i} - Y_{1j})(Y_{2i} - Y_{2j}) < 0),$$

where two observations (Y_{1i}, Y_{2i}) and (Y_{1i}, Y_{2i}), $i \neq j$ are called concordant if $(Y_{1i} - Y_{1j})(Y_{2i} - Y_{2j}) > 0$ and discordant if $(Y_{1i} - Y_{1j})(Y_{2i} - Y_{2j}) < 0$. There is a one-to-one relation between Kendall's tau τ and ϕ which is given by

$$\tau = \frac{\phi}{\phi + 2}, \quad \phi \geq 0,$$

and the range of possible values for τ is [-1, 1].

Another measure of dependence is tail dependence. Tail dependence describes the amount of dependence in the tails of a bivariate distribution. In other words, tail dependence measures the dependence between the variables in the upper-right quadrant and in the lower-left quadrant of a bivariate distribution. The concept of tail dependence has been discussed in some data applications, for example, in financial applications related to market or credit risk (Embrechts et al., 2003; Hauksson et al., 2001).

The tail dependence describes the limiting proportion that one margin exceeds a certain threshold given that the other margin has already exceeded that threshold. The upper tail dependence parameter λ_U is the limit of the conditional probability of Y_1 which is greater than the $100p^{th}$ percentile of F_1 given that Y_2 is greater than the $100p^{th}$ percentile of F_2 as p approaches 1, that is

$$\lambda_U = \lim_{p \to 1^-} P\left[Y_1 > F_1^{-1}(p) | Y_2 > F_2^{-1}(p)\right]$$

if the limit exists.

Similarly, the lower tail dependence parameter λ_L is the limit of the conditional probability of Y_1 is less than or equal to the $100p^{th}$ percentile of F_1 given that Y_2 is less than or equal to the $100p^{th}$ percentile of F_2 as p approaches 0, that is

$$\lambda_L = \lim_{p \to 0^+} P\left[Y_1 \leq F_1^{-1}(p) | Y_2 \leq F_2^{-1}(p)\right]$$

if the limit exists.

There is upper tail dependence between two random variables if $\lambda_U \in (0, 1]$, and if $\lambda_U = 0$, they are upper tail-independent. Similarly, there is lower tail dependence between two random variables if $\lambda_L \in (0, 1]$, and if $\lambda_L = 0$, they are lower tail-independent. Under the Clayton copula, the lower tail dependence parameter is $\lambda_L = 2^{-\frac{1}{\phi}}$ and there is no upper tail dependence (Joe, 1997).

1.7 Aim and the Outline of the Study

In this study, the main objective is to investigate efficiency of response-dependent two-phase sampling designs. We identify informative sampling designs which give efficient estimates of the coefficient of an expensive covariate for a given sample size. We consider sampling based on a continuous response variable Y, in addition to the inexpensive covariate Z. Further, we investigate efficiency of response-dependent two-phase sampling designs under bivariate response variable models. We consider different estimation methods. Unbiasedness and efficiency of estimators under different estimation methods for each sampling design are assessed. We also check the robustness of estimators when the conditional distribution of the response variable given covariates is misspecified.

In this study, we assume an existing cohort of individuals in phase I, where response variable values are available for all individuals. The response-dependent BSS design in Section 1.2.2 is considered at phase II sample selection. The main objective of this study is to investigate how to specify sampling probabilities for each stratum for a fixed phase II sample size n to achieve an efficient estimate for the coefficient of an expensive covariate. To make inference on the coefficient of the expensive covariate, we consider the likelihood-based methods and the pseudo-likelihood methods described in Section 1.4.

We also investigate the estimation methods under the multiple response-dependent sampling designs and phase II sampling schemes giving the most efficient estimate under each estimation method. We investigate the gain in efficiency under multiple response-dependent sampling designs compared to single response-dependent sampling designs.

In Chapter 2, we present the response-dependent sampling design settings when there is a single response variable. Efficiency of the sampling designs is investigated through a simulation study. Also, we compare the performance of estimation methods under different sampling design settings. We assess the robustness of the estimators when the model distribution is misspecified.

In Chapter 3, efficiency of different phase II sampling designs depending on inexpensive covariate in addition to response variable is evaluated. A simulation study is performed to identify efficient sampling designs under each estimation method and

also to compare the performance of the methods. Further, robustness of estimators is discussed.

In Chapter 4, efficiency of different sampling designs is discussed when the sampling depends on multiple response variables. We compare efficiency of the sampling designs using a simulation study and assess performance of the estimation methods under different sampling design settings. Also, we compare the efficiency gain under multiple response-dependent designs to single response-dependent designs. In Chapter 5, we summarize the results obtained in this study.

Chapter 2

Efficiency of Two-Phase Response-Dependent Sampling Designs

Suppose the response variable values $\{y_1, y_2, \ldots, y_N\}$ were observed for a cohort of size N in phase I. In phase II, n units are selected to obtain their expensive covariate data. To select these units, the response variable observations are partitioned into $K = 3$ mutually exclusive intervals constructed based on the fixed constants C_1 and C_2 as follows:

$$\underbrace{y_1 \leq \ldots \leq y_{N_1}}_{\substack{\text{Low stratum} \\ (S_1)}} \leq C_1 \leq \underbrace{y_{N_1+1} \leq \ldots \leq y_{N_1+N_2}}_{\substack{\text{Middle stratum} \\ (S_2)}} \leq C_2 \leq \underbrace{y_{N_1+N_2+1} \leq \ldots \leq y_N}_{\substack{\text{High stratum} \\ (S_3)}} \tag{2.1}$$

Suppose there are N_1 units in the first stratum S_1 which includes units with response values lower than C_1, N_2 units in the second stratum S_2 which includes units with response values between C_1 and C_2, and $N_3 = N - N_1 - N_2$ units in the third stratum S_3 which includes units with response values greater than C_2. The first and third strata consist of extreme values, where in this study the sizes of the extreme strata are set small compared to central stratum to address the importance of sampling from the extreme strata. Then, the response-dependent BSS design described in Section 1.2.2 is applied to select the phase II sample. That is, from each stratum S_j, n_j units

are randomly selected for inclusion in the phase II sample, where the total phase II sample size n is fixed according to the budgetary constraint.

In this study, we investigate different sampling designs to select sampling units from each stratum for phase II as summarized in Table 2.1. In sampling design (I), we consider sampling all units in the extreme strata and no units from the middle stratum. In other words, sampling probabilities from the extreme strata are 1 and the sampling probability from the middle stratum is 0. In design (II), number of units selected from the middle stratum is less than the number of sampling units selected from the extreme strata. In design (III), more units are selected only from one extreme stratum. In design (IV), more units are selected from the middle stratum compared to the extreme strata. Design (V) is a simple random sampling design. We consider different scenarios under these sampling designs but in all of these designs, we assume the available budget to measure the expensive covariate is only for n units. Therefore, the phase II sample size $n = \sum_{j=1}^{3} n_j$ is the same over each scenario.

Table 2.1: Response-dependent BSS designs

Sampling scenario	**Stratum-specific sample sizes**			**Sampling design**
	Low stratum sample size n_1	**Middle stratum sample size** n_2	**High stratum sample size** n_3	
1	Large	0	Large	(I) Sampling from extreme strata only
2	Medium	Small	Medium	(II) Oversampling from extreme strata
3	Large	Small	Medium	
4	Medium	Small	Large	
5	Large	Medium	Small	(III) Oversampling from only one extreme stratum
6	Small	Medium	Large	
7	Medium	Large	Medium	(IV) Oversampling from middle stratum
8	Small	Large	Medium	
9	Medium	Large	Small	
10	Small	Large	Small	(V) Simple random sampling

2.1 Model Description

Suppose that the data $\{(y_i, x_i) : \ i = 1, 2, \ldots, N\}$ come from the linear model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{2.2}$$

where β_0, β_1 are the regression coefficients and ϵ_i is a random error for subject i ($i = 1, 2, \ldots, N$).

In our study, we assume X is a Bernoulli random variable with probability of having $X = 1$ being p, ϵ_i's are independently and identically normally distributed with mean 0 and variance σ^2. Therefore, Y_i has a Normal distribution with mean $\beta_0 + \beta_1 x_i$ and variance σ^2 when $X_i = x_i$ is given.

We have three unknown parameters, β_0, β_1, and σ. To estimate $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma)$, we obtain the samples by applying the proposed sampling designs presented in Table 2.1. We presented five different estimation methods in Section 1.4 to estimate the parameter vector $\boldsymbol{\theta}$. In this study, under each estimation method, we aim to check whether the estimator of β_1 is unbiased and to investigate the efficiency of the estimators under each sampling design.

Under the model in (2.2) with the specified assumptions,

$$f(y_i|x_i;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}}, \tag{2.3}$$

where $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma)$ and $g(x_i; p) = p^{x_i}(1-p)^{1-x_i}$ for $x_i = 0, 1$ are the conditional density function for Y_i given $X_i = x_i$ and the marginal distribution of X_i, respectively.

The full likelihood $L_R(\boldsymbol{\theta}, p)$ which incorporates both complete and incomplete data $\{(y_i, x_i) : i \in V\} \cup \{y_i : i \in \bar{V}\}$ under the missing at random assumption was given in (1.7). Under the current setting, the logarithm of the likelihood function becomes

$$\begin{aligned} l_R(\boldsymbol{\theta}, p) = &\sum_{i \in V} [\log f(y_i|x_i;\boldsymbol{\theta}) + \log g(x_i; p)] \\ &+ \sum_{i \in \bar{V}} \log[p f(y_i|x=1;\boldsymbol{\theta}) + (1-p) f(y_i|x=0;\boldsymbol{\theta})]. \end{aligned} \tag{2.4}$$

The maximum likelihood estimates of $\boldsymbol{\theta}$ and p are obtained by maximizing $l_R(\boldsymbol{\theta}, p)$ in (2.4) with respect to $\boldsymbol{\theta}$ and p. We denote the resulting estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_R$.

The likelihood $L_F(\boldsymbol{\theta}, g)$ which incorporates complete data $\{(y_i, x_i) : i \in V\}$ and the stratum information only for unobserved data was given in (1.9). Under the current

setting, the logarithm of the likelihood function becomes

$$l_F(\boldsymbol{\theta}, p) = \sum_{j=1}^{3} \left\{ \left(\sum_{i \in D_j} \log f(y_i|x_i;\boldsymbol{\theta}) + \log g(x_i;p) \right) \right.$$
$$\left. +(N_j - n_j) \log \left(p \int_{C_{j-1}}^{C_j} f(y|x=1;\boldsymbol{\theta})dy + (1-p) \int_{C_{j-1}}^{C_j} f(y|x=0;\boldsymbol{\theta})dy \right) \right\}, \tag{2.5}$$

where C_1, C_2 are fixed cut-point values and $C_0 = -\infty, C_3 = \infty$. The maximum likelihood estimates of $\boldsymbol{\theta}$ and p are obtained by maximizing $l_F(\boldsymbol{\theta}, p)$ in (2.5) with respect to $\boldsymbol{\theta}$ and p. We denote the resulting estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_F$.

The likelihood of complete data $\{(y_i, x_i) : i \in V\}$, $L_{C1}(\boldsymbol{\theta})$, was given in (1.12). Under the current setting, the logarithm of the likelihood function becomes

$$l_{C1}(\boldsymbol{\theta}) = \sum_{i:R_i=1} \left\{ \log f(y_i|x_i;\boldsymbol{\theta}) + \log \pi_i - \log \left(\sum_{l=1}^{3} \frac{n_l}{N_l} \int_{C_{l-1}}^{C_l} f(y|x_i;\boldsymbol{\theta})dy \right) \right\}, \tag{2.6}$$

where C_1, C_2 are fixed cut-point values, $C_0 = -\infty, C_3 = \infty$ and $\pi_i = P(R_i = 1|y_i, z_i)$. The estimate of $\boldsymbol{\theta}$ is obtained by maximizing $l_{C1}(\boldsymbol{\theta})$ in (2.6) with respect to $\boldsymbol{\theta}$. We denote the resulting estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_{C1}$.

Under the current setting, the logarithm of the pseudo-likelihood function, $l_P(\boldsymbol{\theta})$, given in (1.14) becomes

$$l_P(\boldsymbol{\theta}) = \sum_{i:R_i=1} \log f(y_i|x_i;\boldsymbol{\theta}) + \sum_{j=1}^{3} (N_j - n_j) \log \left(\sum_{l=1}^{3} \frac{N_l}{n_l} \sum_{i \in D_l} \int_{C_{j-1}}^{C_j} f(y|x_i;\boldsymbol{\theta})dy \right), \tag{2.7}$$

where C_1, C_2 are fixed cut-point values and $C_0 = -\infty, C_3 = \infty$. The estimate of $\boldsymbol{\theta}$ is obtained by maximizing $l_P(\boldsymbol{\theta})$ in (2.7) with respect to $\boldsymbol{\theta}$. We denote the resulting estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_P$.

The IPW method estimating function given in (1.15) becomes

$$S_w(\boldsymbol{\theta}) = \begin{pmatrix} \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{1i}(\boldsymbol{\theta}) \\ \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{2i}(\boldsymbol{\theta}) \\ \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{3i}(\boldsymbol{\theta}) \end{pmatrix}, \tag{2.8}$$

where

$$\begin{aligned} U_{1i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_i|x_i;\boldsymbol{\theta})}{\partial \beta_0} = \frac{(y_i - \beta_0 - \beta_1 x_i)}{\sigma^2}, \\ U_{2i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_i|x_i;\boldsymbol{\theta})}{\partial \beta_1} = \frac{(y_i - \beta_0 - \beta_1 x_i)x_i}{\sigma^2}, \\ U_{3i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_i|x_i;\boldsymbol{\theta})}{\partial \sigma} = \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^3} - \frac{1}{\sigma}. \end{aligned}$$

By solving the estimating equation $S_w(\boldsymbol{\theta}) = 0$, we obtain the IPW estimators as

$$\begin{aligned} \hat{\beta}_{0,w} &= \frac{\sum_{i=1}^{N} \frac{R_i}{\pi_i} y_i x_i \sum_{i=1}^{N} \frac{R_i}{\pi_i} x_i - \sum_{i=1}^{N} \frac{R_i}{\pi_i} y_i \sum_{i=1}^{N} \frac{R_i}{\pi_i} x_i^2}{(\sum_{i=1}^{N} \frac{R_i}{\pi_i} x_i)^2 - \sum_{i=1}^{N} \frac{R_i}{\pi_i} \sum_{i=1}^{N} \frac{R_i}{\pi_i} x_i^2}, \\ \hat{\beta}_{1,w} &= \frac{\sum_{i=1}^{N} \frac{R_i}{\pi_i} y_i \sum_{i=1}^{N} \frac{R_i}{\pi_i} x_i - \sum_{i=1}^{N} \frac{R_i}{\pi_i} y_i x_i \sum_{i=1}^{N} \frac{R_i}{\pi_i}}{(\sum_{i=1}^{N} \frac{R_i}{\pi_i} x_i)^2 - \sum_{i=1}^{N} \frac{R_i}{\pi_i} \sum_{i=1}^{N} \frac{R_i}{\pi_i} x_i^2}, \end{aligned} \tag{2.9}$$

$$\hat{\sigma^2}_w = \frac{\sum_{i=1}^{N} \frac{R_i}{\pi_i} (y_i - \hat{\beta}_{0,w} - \hat{\beta}_{1,w} x_i)^2}{\sum_{i=1}^{N} \frac{R_i}{\pi_i}}.$$

To ensure correctness of the derivations for the IPW estimator $\hat{\boldsymbol{\theta}}_w = (\hat{\beta}_{0,w}, \hat{\beta}_{1,w}, \hat{\sigma}_w)$,

we also used the "survey" package in R to obtain the IPW estimate of $\boldsymbol{\theta}$ and obtained $\hat{\boldsymbol{\theta}}_{Sy} = (\hat{\beta}_{0,Sy}, \hat{\beta}_{1,Sy}, \hat{\sigma}_{Sy})$ by using the *svydesign* and *svyglm* commands.

Under some regularity conditions, the covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta})$ given in (1.17) becomes

$$C(\boldsymbol{\theta}) = A(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}) [A(\boldsymbol{\theta})^{-1}]',$$

where

$$A(\boldsymbol{\theta}) = -\frac{1}{n}\left(\frac{\partial S_w}{\partial \boldsymbol{\theta}'}\right) = \begin{pmatrix} \sum_{i=1}^{N} \frac{1}{n\sigma^2}\frac{R_i}{\pi_i} & \sum_{i=1}^{N} \frac{1}{n\sigma^2}\frac{R_i}{\pi_i} x_i & 0 \\ \sum_{i=1}^{N} \frac{1}{n\sigma^2}\frac{R_i}{\pi_i} x_i & \sum_{i=1}^{N} \frac{1}{n\sigma^2}\frac{R_i}{\pi_i} x_i^2 & 0 \\ 0 & 0 & \sum_{i=1}^{N} \frac{2}{n\sigma^2}\frac{R_i}{\pi_i} \end{pmatrix},$$

and

$$B(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{N} S_{wi}(\boldsymbol{\theta}) S_{wi}(\boldsymbol{\theta})' = \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{pmatrix},$$

with

$$B_{11} = \frac{1}{n}\sum_{i=1}^{N} \frac{R_i}{\pi_i^2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^4},$$

$$B_{22} = \frac{1}{n}\sum_{i=1}^{N} \frac{R_i}{\pi_i^2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2 x_i^2}{\sigma^4},$$

$$B_{33} = \frac{1}{n}\sum_{i=1}^{N} \frac{R_i}{\pi_i^2} \left[\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^3} - \frac{1}{\sigma}\right]^2,$$

$$B_{12} = B_{21} = \frac{1}{n}\sum_{i=1}^{N} \frac{R_i}{\pi_i^2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2 x_i}{\sigma^4},$$

$$B_{13} = B_{31} = \frac{1}{n}\sum_{i=1}^{N} \frac{R_i}{\pi_i^2} \frac{(y_i - \beta_0 - \beta_1 x_i)}{\sigma^2} \left[\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^3} - \frac{1}{\sigma}\right],$$

$$B_{23} = B_{32} = \frac{1}{n}\sum_{i=1}^{N} \frac{R_i}{\pi_i^2} \frac{(y_i - \beta_0 - \beta_1 x_i)x_i}{\sigma^2} \left[\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^3} - \frac{1}{\sigma}\right].$$

The covariance matrix of $\hat{\boldsymbol{\theta}}_w = (\hat{\beta}_{0,w}, \hat{\beta}_{1,w}, \hat{\sigma}_w)$ is $\frac{1}{n}C(\boldsymbol{\theta})$. A consistent estimator of $C(\boldsymbol{\theta})$ is obtained by plugging the estimates of $\boldsymbol{\theta}$ in the sandwich form given in (1.18) and the estimated covariance matrix of estimator is obtained as

$$\frac{1}{n}C(\hat{\boldsymbol{\theta}}_w). \tag{2.10}$$

2.2 Simulation Study

A simulation study was conducted to investigate the properties of the estimation methods under the sampling design settings described in Table 2.1. We also aim to identify the sampling design which gives the most efficient estimate of the coefficient of expensive covariate under each estimation method. We assume that there is a cohort of size $N = 50,000$ with their observed response variable Y values in the first phase. We are interested in making inference on the association between the response variable Y and an expensive covariate X. We consider a linear regression of Y on X with a normally distributed error term.

In data generation, we first generated error terms of size $N = 50,000$ from Normal distribution with mean 0 and variance $\sigma^2 = 2$, and we generated the covariate values x_i from Bernoulli distribution with $p = P(X_i = 1) = 0.4$ or 0.05. Here, we considered two different values of p to investigate the effect of changing the value of p. Note that when $p = 0.05$, X rarely takes the value of 1. Next, we generated the response variable values $y_i's$ using the model (2.2), where we choose different values of $\beta_1(= 0, 0.5, 1)$ with the intercept $\beta_0 = 10$ to investigate the effect of changing the values of β_1.

In our study, we set (C_1, C_2) in (2.1) to the (10^{th}, 90^{th}) and the (30^{th}, 70^{th}) percentiles of the response distribution which divide the cohort into three strata with approximate sizes of $N_1 \approx 5,000, N_2 \approx 40,000, N_3 \approx 5,000$ and $N_1 \approx 15,000, N_2 \approx 20,000, N_3 \approx 15,000$, respectively. We investigate the effect of changing extreme stratum sizes under these two settings. The cut-point values C_1 and C_2 under different parameter values are given in Table 2.2.

We applied the response-dependent BSS designs described in Table 2.1 to obtain

the phase II samples and only for these selected samples, we assume to have the expensive covariate X values. The stratum specific probabilities under each design are given in Table 2.3 and 2.4 for cut-point values of the (10^{th}, 90^{th}) and (30^{th}, 70^{th}) percentiles of the response distribution, respectively. In each sampling design, the phase II sample size is set to $n = 10,000$ and the stratum specific sample sizes are given in Table 2.5 under two different cut-point settings. The estimated value of β_1 is obtained using the drawn samples under each estimation method described in Section 1.4. For all ten sampling designs, we repeated the above process and calculated the point estimate of β_1, its standard error estimate and mean square error estimate.

Table 2.2: The cut-point values C_1 and C_2 under different parameter values

p	β_1	Cut-point percentiles (%)		Cut-point values	
		C_1	C_2	C_1	C_2
0.4	0	10	90	8.1846	11.8142
0.4	0.5	10	90	8.3600	12.0360
0.4	1	10	90	8.4757	12.3205
0.05	0	10	90	8.1846	11.8142
0.05	0.5	10	90	8.2047	11.8431
0.05	1	10	90	8.2146	11.8831
0.4	0	30	70	9.2583	10.7432
0.4	0.5	30	70	9.4495	10.9569
0.4	1	30	70	9.6094	11.1885

Table 2.3: Response-dependent BSS designs when $C_1 = 10^{\text{th}}$ percentile and $C_2 = 90^{\text{th}}$ percentile

Sampling scenario	**Sampling probability**			**Sampling design**
	Low stratum π_1	**Middle stratum** π_2	**High stratum** π_3	
1	100%	0%	100%	(I) Sampling from extreme strata only
2	80%	5%	80%	(II) Oversampling from extreme strata
3	100%	5%	60%	
4	60%	5%	100%	
5	100%	10%	20%	(III) Oversampling from only one extreme stratum
6	20%	10%	100%	
7	50%	12.5%	50%	(IV) Oversampling from middle stratum
8	20%	15%	60%	
9	60%	15%	20%	
10	20%	20%	20%	(V) Simple random sampling

Table 2.4: Response-dependent BSS designs when $C_1 = 30^{\text{th}}$ percentile and $C_2 = 70^{\text{th}}$ percentile

Sampling scenario	**Sampling probability**			**Sampling design**
	Low stratum π_1	**Middle stratum** π_2	**High stratum** π_3	
1	33.33%	0%	33.33%	(I) Sampling from extreme strata only
2	26.67%	10.00%	26.67%	(II) Oversampling from extreme strata
3	33.33%	10.00%	20.00%	
4	20.00%	10.00%	33.33%	
5	33.33%	20.00%	6.67%	(III) Oversampling from only one extreme stratum
6	6.67%	20.00%	33.33%	
7	16.67%	25.00%	16.67%	(IV) Oversampling from middle stratum
8	6.67%	30.00%	20.00%	
9	20.00%	30.00%	6.67%	
10	20.00%	20.00%	20.00%	(V) Simple random sampling

Table 2.5: Stratum-specific sample sizes under each response-dependent BSS design

Sampling scenario	**Stratum-specific sample sizes**			**Sampling design**
	Low stratum $n_1\left(\frac{n_1}{n} \times 100\%\right)$	**Middle stratum** $n_2\left(\frac{n_2}{n} \times 100\%\right)$	**High stratum** $n_3\left(\frac{n_3}{n} \times 100\%\right)$	
1	5000 (50%)	0 (0%)	5000 (50%)	(I) Sampling from extreme strata only
2	4000 (40%)	2000 (20%)	4000 (40%)	(II) Oversampling from extreme strata
3	5000 (50%)	2000 (20%)	3000 (30%)	
4	3000 (30%)	2000 (20%)	5000 (50%)	
5	5000 (50%)	4000 (40%)	1000 (10%)	(III) Oversampling from only one extreme stratum
6	1000 (10%)	4000 (40%)	5000 (50%)	
7	2500 (25%)	5000 (50%)	2500 (25%)	(IV) Oversampling from middle stratum
8	1000 (10%)	6000 (60%)	3000 (30%)	
9	3000 (30%)	6000 (60%)	1000 (10%)	
10	1000 (10%)	8000 (80%)	1000 (10%)	(V) Simple Random Sampling When $C_1 = 10^{\text{th}}$ and $C_2 = 90^{\text{th}}$ percentile
	3000 (30%)	4000 (40%)	3000 (30%)	When $C_1 = 30^{\text{th}}$ and $C_2 = 70^{\text{th}}$ percentile

When $C_1 = 10^{\text{th}}$ percentile and $C_2 = 90^{\text{th}}$ percentile, the strata sizes are $N_1 = 5,000$, $N_2 = 40,000$, $N_3 = 5,000$. When $C_1 = 30^{\text{th}}$ percentile and $C_2 = 70^{\text{th}}$ percentile, the strata sizes are $N_1 = 15,000$, $N_2 = 20,000$, $N_3 = 15,000$.

2.3 Simulation Results

The parameter estimates of β_1 and their standard error estimates and mean square error estimates under different parameter values and different estimation methods are presented in Tables 2.6 - 2.15. Table 2.6 shows the results under the maximum likelihood estimation using the full cohort data. As expected, we obtained consistent maximum likelihood estimates.

Tables 2.7, 2.8, and 2.9 show properties of different parameter estimators of the coefficient β_1 of the expensive covariate when $\beta_1 = 0,\ 0.5,\ 1$, respectively, $p = 0.4$ and cut-point values C_1 and C_2 set to the 10^{th} and 90^{th} percentiles. From these simulation results, we observe that the likelihood-based methods give consistent estimates and more efficient estimates than the IPW estimation method. Under the extreme sampling design (sampling scenario 1), likelihood-based methods provide the most efficient estimates as they give the lowest standard errors compared to other sampling designs. On the other hand, the pseudo-likelihood methods give biased estimates under the extreme sampling design. The sampling designs selecting more units from the central stratum give consistent and more efficient estimates under the IPW estimation method. In fact, we obtain the most efficient IPW estimate when oversampling from the middle stratum with equal sampling probabilities from extreme strata (sampling scenario 7). However, likelihood-based estimators are more efficient than the IPW method even under these sampling designs. The conditional likelihood method yields a little less efficient estimates, when the association between X and Y becomes stronger (i.e., the values of β_1 changes from 0 to 1). Estimated pseudo-likelihood estimation gives consistent estimates under all designs considered except the extreme sampling design. It gives as efficient estimates or more efficient estimates than likelihood-based methods except the extreme sampling design. It gives the most efficient estimates when oversampling from extreme strata while selecting some from the middle stratum (sampling scenarios 2, 3, and 4). When we compare the estimates under the most efficient design with the full cohort, we observe an approximately 40% loss in efficiency when the sample size reduces to $n = 10,000$ from $N = 50,000$.

Tables 2.10, 2.11, and 2.12 show the properties of different parameter estimators of the coefficient β_1 of the expensive covariate when $\beta_1 = 0,\ 0.5,\ 1$, respectively, $p = 0.05$ and cut-point values C_1 and C_2 set to the 10^{th} and 90^{th} percentiles. We observe that the estimation methods may give biased estimates under some sampling designs. Thus, we compare the MSEs to determine efficient designs. When $p = 0.4$, under the extreme sampling design (sampling scenario 1), likelihood-based methods provide the most efficient estimates. However when $X = 1$ values are rare, the relative efficiency of the extreme sampling design is low compared to many other designs. When X and Y are associated, estimated pseudo-likelihood estimation method gives the most efficient estimates under many sampling design settings except the extreme sampling

design. The IPW estimation method yields the least efficient designs when $p = 0.05$ as well. As expected, we observed that the estimation methods give less efficient estimators compared to the $p = 0.4$ setting as the standard error estimates are high when $p = 0.05$.

The results in Tables 2.13, 2.14, and 2.15 show the properties of different parameter estimators of the coefficient β_1 of the expensive covariate when $\beta_1 = 0,\ 0.5,\ 1$, respectively, $p = 0.4$ and cut-point values set to the 30^{th} and 70^{th} percentiles. In this setting, the sizes of the extreme strata are larger. Thus, for example under the extreme strata sampling, the sampling probabilities of units are not 1 anymore. Instead, we consider a simple random sampling in each extreme stratum. We observe that under the extreme sampling design, when the association between X and Y is weak, pseudo-likelihood methods give consistent estimates. When the association between X and Y is strong (i.e., $\beta_1 = 0.5$ and 1), the observed bias in pseudo-likelihood methods estimates is less compared to the setting with cut-point values $C_1 = 10^{th}$ and $C_2 = 90^{th}$ percentile. Similar to the setting with cut-point values set to the 10^{th} and 90^{th} percentile, under the extreme sampling design (sampling scenario 1), likelihood-based methods provide the most efficient estimates as they give the lowest standard errors compared to other sampling designs. The pseudo-likelihood methods give the most efficient estimates when oversampling from extreme strata while selecting some from the middle stratum. The likelihood-based methods give consistent estimates and a little more efficient estimates than the pseudo-likelihood methods. All of the estimation method estimates became less efficient under each design except the simple random sampling compared to the setting with the cut-point values set to the 10^{th} and 90^{th} percentiles.

The IPW estimates of β_1 and their standard error estimates obtained by using the estimator in equation (2.9) and the variance estimator in (2.10) and by the *survey* package in R are given in Table A.1. These two calculation ways gave very close results.

Table 2.6: Maximum likelihood estimation results under the full cohort

p	β_1	Estimates		
		$\hat{\beta}_1$	SE	MSE
0.4	0	0.008	0.013	0.0002
0.4	0.5	0.508	0.013	0.0002
0.4	1	1.008	0.013	0.0002
0.05	0	0.020	0.029	0.0012
0.05	0.5	0.520	0.029	0.0012
0.05	1	1.020	0.029	0.0012

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 2.7: Simulation results when $\beta_1 = 0$, $p = 0.4$, $C_1 = 10^{\text{th}}$ percentile and $C_2 = 90^{\text{th}}$ percentile

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	0.018	0.018	0.018	0.059	0.059
			SE	0.016	0.016	0.016	0.052	0.052
			MSE	0.0006	0.0006	0.0006	0.0062	0.0062
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	0.003	0.003	0.003	0.003	-0.024
			SE	0.018	0.018	0.018	0.018	0.036
			MSE	0.0003	0.0003	0.0003	0.0003	0.0019
3		(50%, 20%, 30%)	$\hat{\beta}_1$	0.020	0.020	0.020	0.019	0.035
			SE	0.018	0.018	0.018	0.018	0.037
			MSE	0.0007	0.0007	0.0007	0.0007	0.0026
4		(30%, 20%, 50%)	$\hat{\beta}_1$	0.013	0.013	0.013	0.019	-0.010
			SE	0.018	0.018	0.018	0.018	0.036
			MSE	0.0005	0.0005	0.0005	0.0007	0.0014
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	0.007	0.007	0.007	0.012	-0.016
			SE	0.023	0.023	0.023	0.020	0.031
			MSE	0.0005	0.0006	0.0006	0.0005	0.0012
6		(10%, 40%, 50%)	$\hat{\beta}_1$	-0.002	-0.002	-0.002	0.001	-0.037
			SE	0.023	0.023	0.023	0.020	0.030
			MSE	0.0005	0.0005	0.0005	0.0004	0.0023
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	0.014	0.014	0.014	0.014	0.014
			SE	0.021	0.021	0.021	0.021	0.026
			MSE	0.0007	0.0007	0.0007	0.0007	0.0009
8		(10%, 60%, 30%)	$\hat{\beta}_1$	-0.002	-0.002	-0.002	0.000	-0.026
			SE	0.024	0.024	0.024	0.023	0.027
			MSE	0.0006	0.0006	0.0006	0.0005	0.0014
9		(30%, 60%, 10%)	$\hat{\beta}_1$	0.018	0.018	0.018	0.017	0.033
			SE	0.024	0.024	0.024	0.023	0.028
			MSE	0.0009	0.0009	0.0009	0.0008	0.0018
10	Simple random sampling	(10%, 80%, 10%)	$\hat{\beta}_1$	-0.009	-0.009	-0.009	-0.009	-0.009
			SE	0.029	0.029	0.029	0.029	0.029
			MSE	0.0009	0.0009	0.0009	0.0009	0.0009

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 2.8: Simulation results when $\beta_1 = 0.5$, $p = 0.4$, $C_1 = 10^{\text{th}}$ percentile and $C_2 = 90^{\text{th}}$ percentile

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	0.517	0.517	0.519	1.653	1.653
			SE	0.016	0.016	0.017	0.050	0.050
			MSE	0.0006	0.0006	0.0007	1.3312	1.3322
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	0.532	0.531	0.533	0.532	0.537
			SE	0.018	0.018	0.018	0.018	0.037
			MSE	0.0013	0.0013	0.0014	0.0013	0.0028
3		(50%, 20%, 30%)	$\hat{\beta}_1$	0.508	0.507	0.507	0.505	0.494
			SE	0.018	0.018	0.018	0.018	0.037
			MSE	0.0004	0.0004	0.0004	0.0003	0.0014
4		(30%, 20%, 50%)	$\hat{\beta}_1$	0.518	0.518	0.520	0.516	0.470
			SE	0.018	0.018	0.019	0.018	0.037
			MSE	0.0007	0.0007	0.0007	0.0006	0.0023
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	0.503	0.503	0.507	0.505	0.501
			SE	0.022	0.022	0.023	0.020	0.031
			MSE	0.0005	0.0005	0.0006	0.0004	0.0010
6		(10%, 40%, 50%)	$\hat{\beta}_1$	0.510	0.510	0.512	0.503	0.514
			SE	0.022	0.022	0.023	0.019	0.030
			MSE	0.0006	0.0006	0.0007	0.0004	0.0011
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	0.518	0.518	0.521	0.517	0.521
			SE	0.021	0.021	0.022	0.021	0.026
			MSE	0.0008	0.0008	0.0009	0.0007	0.0011
8		(10%, 60%, 30%)	$\hat{\beta}_1$	0.503	0.502	0.502	0.499	0.510
			SE	0.024	0.024	0.024	0.023	0.027
			MSE	0.0006	0.0006	0.0006	0.0005	0.0008
9		(30%, 60%, 10%)	$\hat{\beta}_1$	0.504	0.504	0.507	0.506	0.494
			SE	0.024	0.024	0.025	0.023	0.028
			MSE	0.0006	0.0006	0.0007	0.0006	0.0008
10	Simple random sampling	(10%, 80%, 10%)	$\hat{\beta}_1$	0.512	0.512	0.516	0.512	0.516
			SE	0.028	0.028	0.029	0.028	0.029
			MSE	0.0009	0.0009	0.0011	0.0009	0.0011

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 2.9: Simulation results when $\beta_1 = 1$, $p = 0.4$, $C_1 = 10^{\text{th}}$ percentile and $C_2 = 90^{\text{th}}$ percentile

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	1.002	1.003	1.004	3.071	3.071
			SE	0.017	0.017	0.019	0.045	0.044
			MSE	0.0003	0.0003	0.0004	4.2921	4.2921
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	1.002	1.003	1.004	1.004	1.061
			SE	0.018	0.018	0.020	0.018	0.038
			MSE	0.0003	0.0003	0.0004	0.0003	0.0052
3		(50%, 20%, 30%)	$\hat{\beta}_1$	1.001	1.001	0.998	1.000	1.027
			SE	0.018	0.018	0.020	0.018	0.039
			MSE	0.0003	0.0003	0.0004	0.0003	0.0023
4		(30%, 20%, 50%)	$\hat{\beta}_1$	1.011	1.013	1.014	1.025	1.043
			SE	0.019	0.019	0.020	0.018	0.038
			MSE	0.0005	0.0005	0.0006	0.0010	0.0033
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	1.011	1.013	1.017	1.010	1.051
			SE	0.021	0.022	0.024	0.021	0.032
			MSE	0.0006	0.0006	0.0009	0.0005	0.0036
6		(10%, 40%, 50%)	$\hat{\beta}_1$	1.010	1.013	1.015	1.021	0.993
			SE	0.022	0.022	0.024	0.019	0.031
			MSE	0.0006	0.0007	0.0008	0.0008	0.0010
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	1.013	1.016	1.020	1.015	1.041
			SE	0.021	0.021	0.022	0.021	0.027
			MSE	0.0006	0.0007	0.0009	0.0007	0.0024
8		(10%, 60%, 30%)	$\hat{\beta}_1$	1.010	1.010	1.008	1.011	0.999
			SE	0.023	0.023	0.024	0.022	0.028
			MSE	0.0006	0.0006	0.0007	0.0006	0.0008
9		(30%, 60%, 10%)	$\hat{\beta}_1$	1.022	1.024	1.030	1.024	1.055
			SE	0.023	0.023	0.025	0.023	0.028
			MSE	0.0010	0.0011	0.0015	0.0011	0.0038
10	Simple random sampling	(10%, 80%, 10%)	$\hat{\beta}_1$	1.011	1.013	1.019	1.013	1.019
			SE	0.026	0.027	0.029	0.027	0.029
			MSE	0.0008	0.0009	0.0012	0.0009	0.0012

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 2.10: Simulation results when $\beta_1 = 0$, $p = 0.05$, $C_1 = 10^{\text{th}}$ percentile and $C_2 = 90^{\text{th}}$ percentile

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	0.052	0.052	0.051	0.167	0.168
			SE	0.036	0.036	0.036	0.118	0.118
			MSE	0.0040	0.0040	0.0040	0.0418	0.0420
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	0.051	0.051	0.051	0.053	0.027
			SE	0.041	0.041	0.040	0.041	0.087
			MSE	0.0042	0.0042	0.0042	0.0045	0.0083
3		(50%, 20%, 30%)	$\hat{\beta}_1$	0.035	0.035	0.035	0.045	-0.004
			SE	0.041	0.041	0.040	0.039	0.074
			MSE	0.0029	0.0029	0.0028	0.0036	0.0055
4		(30%, 20%, 50%)	$\hat{\beta}_1$	0.081	0.081	0.081	0.059	0.100
			SE	0.041	0.041	0.041	0.039	0.081
			MSE	0.0082	0.0082	0.0082	0.0050	0.0165
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	0.018	0.018	0.018	0.030	-0.053
			SE	0.051	0.051	0.051	0.045	0.066
			MSE	0.0029	0.0030	0.0030	0.0029	0.0071
6		(10%, 40%, 50%)	$\hat{\beta}_1$	0.038	0.038	0.038	0.037	-0.014
			SE	0.051	0.051	0.051	0.044	0.066
			MSE	0.0041	0.0041	0.0041	0.0034	0.0045
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	0.063	0.063	0.063	0.062	0.031
			SE	0.048	0.049	0.049	0.048	0.058
			MSE	0.0063	0.0063	0.0063	0.0062	0.0044
8		(10%, 60%, 30%)	$\hat{\beta}_1$	0.013	0.013	0.012	0.010	0.016
			SE	0.054	0.054	0.054	0.052	0.058
			MSE	0.0031	0.0031	0.0030	0.0028	0.0037
9		(30%, 60%, 10%)	$\hat{\beta}_1$	0.092	0.092	0.093	0.099	0.054
			SE	0.054	0.054	0.054	0.052	0.061
			MSE	0.0113	0.0114	0.0115	0.0124	0.0066
10	Simple random sampling	(10%, 80%, 10%)	$\hat{\beta}_1$	0.010	0.010	0.010	0.010	0.010
			SE	0.065	0.065	0.065	0.065	0.063
			MSE	0.0043	0.0043	0.0044	0.0043	0.0040

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 2.11: Simulation results when $\beta_1 = 0.5$, $p = 0.05$, $C_1 = 10^{\text{th}}$ percentile and $C_2 = 90^{\text{th}}$ percentile

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	0.568	0.569	0.568	1.647	1.648
			SE	0.041	0.041	0.041	0.111	0.094
			MSE	0.0063	0.0064	0.0063	1.3281	1.3271
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	0.534	0.535	0.535	0.519	0.562
			SE	0.042	0.042	0.042	0.039	0.083
			MSE	0.0029	0.0030	0.0030	0.0019	0.0107
3		(50%, 20%, 30%)	$\hat{\beta}_1$	0.602	0.602	0.600	0.615	0.629
			SE	0.043	0.043	0.043	0.043	0.083
			MSE	0.0123	0.0122	0.0119	0.0152	0.0236
4		(30%, 20%, 50%)	$\hat{\beta}_1$	0.516	0.517	0.517	0.527	0.540
			SE	0.044	0.044	0.044	0.039	0.084
			MSE	0.0022	0.0022	0.0022	0.0022	0.0087
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	0.592	0.597	0.600	0.586	0.616
			SE	0.052	0.053	0.053	0.052	0.077
			MSE	0.0113	0.0122	0.0128	0.0101	0.0194
6		(10%, 40%, 50%)	$\hat{\beta}_1$	0.527	0.528	0.528	0.524	0.515
			SE	0.052	0.052	0.052	0.041	0.066
			MSE	0.0034	0.0035	0.0035	0.0023	0.0045
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	0.509	0.512	0.513	0.516	0.492
			SE	0.049	0.050	0.050	0.049	0.063
			MSE	0.0025	0.0026	0.0027	0.0026	0.0040
8		(10%, 60%, 30%)	$\hat{\beta}_1$	0.539	0.538	0.537	0.543	0.509
			SE	0.052	0.052	0.052	0.048	0.059
			MSE	0.0042	0.0042	0.0041	0.0041	0.0035
9		(30%, 60%, 10%)	$\hat{\beta}_1$	0.561	0.565	0.568	0.564	0.586
			SE	0.054	0.054	0.055	0.054	0.064
			MSE	0.0066	0.0072	0.0076	0.0070	0.0115
10	Simple random sampling	(10%, 80%, 10%)	$\hat{\beta}_1$	0.553	0.558	0.561	0.558	0.560
			SE	0.063	0.063	0.064	0.063	0.066
			MSE	0.0067	0.0073	0.0078	0.0073	0.0081

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 2.12: Simulation results when $\beta_1 = 1$, $p = 0.05$, $C_1 = 10^{\text{th}}$ percentile and $C_2 = 90^{\text{th}}$ percentile

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	1.083	1.088	1.087	2.539	2.537
			SE	0.050	0.051	0.052	0.097	0.061
			MSE	0.0095	0.0103	0.0102	2.3791	2.3673
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	1.081	1.087	1.087	1.024	1.033
			SE	0.050	0.051	0.051	0.040	0.080
			MSE	0.0090	0.0102	0.0101	0.0022	0.0076
3		(50%, 20%, 30%)	$\hat{\beta}_1$	1.036	1.035	1.033	1.003	0.963
			SE	0.048	0.048	0.049	0.043	0.086
			MSE	0.0036	0.0036	0.0035	0.0019	0.0087
4		(30%, 20%, 50%)	$\hat{\beta}_1$	1.052	1.057	1.057	1.008	1.035
			SE	0.051	0.052	0.052	0.038	0.083
			MSE	0.0053	0.0059	0.0059	0.0015	0.0082
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	1.031	1.042	1.047	1.045	1.006
			SE	0.053	0.056	0.057	0.055	0.070
			MSE	0.0038	0.0048	0.0054	0.0050	0.0050
6		(10%, 40%, 50%)	$\hat{\beta}_1$	1.014	1.019	1.019	1.028	0.976
			SE	0.053	0.054	0.054	0.040	0.067
			MSE	0.0030	0.0032	0.0033	0.0023	0.0051
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	1.053	1.062	1.065	1.055	1.085
			SE	0.053	0.054	0.054	0.050	0.062
			MSE	0.0056	0.0068	0.0072	0.0055	0.0110
8		(10%, 60%, 30%)	$\hat{\beta}_1$	0.967	0.964	0.963	0.964	0.959
			SE	0.052	0.053	0.053	0.047	0.059
			MSE	0.0039	0.0041	0.0042	0.0035	0.0052
9		(30%, 60%, 10%)	$\hat{\beta}_1$	0.994	1.003	1.008	1.004	0.976
			SE	0.057	0.059	0.060	0.059	0.066
			MSE	0.0033	0.0035	0.0037	0.0035	0.0050
10	Simple random sampling	(10%, 80%, 10%)	$\hat{\beta}_1$	0.928	0.935	0.940	0.935	0.940
			SE	0.061	0.063	0.064	0.063	0.064
			MSE	0.0089	0.0082	0.0076	0.0082	0.0077

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 2.13: Simulation results when $\beta_1 = 0$, $p = 0.4$, $C_1 = 30^{\text{th}}$ percentile and $C_2 = 70^{\text{th}}$ percentile

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	0.005	0.005	0.005	0.008	0.008
			SE	0.023	0.023	0.023	0.037	0.037
			MSE	0.0005	0.0005	0.0005	0.0014	0.0014
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	0.022	0.022	0.022	0.022	0.018
			SE	0.025	0.025	0.025	0.025	0.026
			MSE	0.0011	0.0011	0.0011	0.0011	0.0010
3		(50%, 20%, 30%)	$\hat{\beta}_1$	-0.008	-0.008	-0.007	-0.003	-0.007
			SE	0.026	0.026	0.026	0.025	0.026
			MSE	0.0007	0.0007	0.0007	0.0006	0.0007
4		(30%, 20%, 50%)	$\hat{\beta}_1$	-0.001	-0.001	-0.001	-0.002	-0.005
			SE	0.026	0.026	0.026	0.025	0.026
			MSE	0.0007	0.0007	0.0007	0.0006	0.0007
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	0.063	0.063	0.063	0.062	0.087
			SE	0.032	0.032	0.032	0.029	0.039
			MSE	0.0050	0.0050	0.0050	0.0046	0.0090
6		(10%, 40%, 50%)	$\hat{\beta}_1$	-0.012	-0.012	-0.012	-0.001	0.019
			SE	0.032	0.033	0.033	0.029	0.038
			MSE	0.0012	0.0012	0.0012	0.0008	0.0018
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	-0.014	-0.014	-0.014	-0.014	-0.017
			SE	0.032	0.031	0.031	0.031	0.031
			MSE	0.0012	0.0012	0.0012	0.0012	0.0013
8		(10%, 60%, 30%)	$\hat{\beta}_1$	-0.015	-0.015	-0.015	-0.006	0.001
			SE	0.036	0.036	0.036	0.035	0.040
			MSE	0.0015	0.0015	0.0015	0.0012	0.0016
9		(30%, 60%, 10%)	$\hat{\beta}_1$	0.048	0.048	0.048	0.053	0.085
			SE	0.036	0.036	0.036	0.035	0.041
			MSE	0.0036	0.0036	0.0036	0.0041	0.0090
10	Simple random sampling	(30%, 40%, 30%)	$\hat{\beta}_1$	-0.014	-0.014	-0.014	-0.014	-0.014
			SE	0.029	0.029	0.029	0.029	0.029
			MSE	0.0010	0.0010	0.0010	0.0010	0.0010

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 2.14: Simulation results when $\beta_1 = 0.5$, $p = 0.4$, $C_1 = 30^{\text{th}}$ percentile and $C_2 = 70^{\text{th}}$ percentile

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	0.451	0.450	0.448	0.721	0.721
			SE	0.023	0.023	0.023	0.036	0.036
			MSE	0.0029	0.0030	0.0032	0.0501	0.0501
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	0.454	0.452	0.448	0.451	0.446
			SE	0.025	0.025	0.025	0.025	0.026
			MSE	0.0027	0.0030	0.0033	0.0030	0.0036
3		(50%, 20%, 30%)	$\hat{\beta}_1$	0.474	0.472	0.470	0.472	0.473
			SE	0.025	0.025	0.026	0.025	0.027
			MSE	0.0013	0.0014	0.0016	0.0014	0.0015
4		(30%, 20%, 50%)	$\hat{\beta}_1$	0.455	0.454	0.453	0.451	0.460
			SE	0.026	0.026	0.026	0.025	0.027
			MSE	0.0027	0.0027	0.0029	0.0030	0.0023
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	0.463	0.462	0.462	0.474	0.508
			SE	0.032	0.032	0.033	0.029	0.039
			MSE	0.0024	0.0025	0.0025	0.0015	0.0016
6		(10%, 40%, 50%)	$\hat{\beta}_1$	0.474	0.475	0.475	0.475	0.497
			SE	0.032	0.032	0.032	0.028	0.038
			MSE	0.0017	0.0017	0.0017	0.0014	0.0014
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	0.483	0.481	0.478	0.481	0.475
			SE	0.031	0.031	0.031	0.031	0.031
			MSE	0.0012	0.0013	0.0015	0.0013	0.0016
8		(10%, 60%, 30%)	$\hat{\beta}_1$	0.508	0.508	0.508	0.513	0.523
			SE	0.034	0.035	0.035	0.034	0.040
			MSE	0.0013	0.0013	0.0013	0.0013	0.0021
9		(30%, 60%, 10%)	$\hat{\beta}_1$	0.482	0.482	0.482	0.489	0.507
			SE	0.035	0.036	0.036	0.034	0.040
			MSE	0.0016	0.0016	0.0016	0.0013	0.0017
10	Simple random sampling	(30%, 40%, 30%)	$\hat{\beta}_1$	0.477	0.476	0.474	0.476	0.474
			SE	0.028	0.028	0.029	0.028	0.029
			MSE	0.0013	0.0014	0.0015	0.0014	0.0015

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 2.15: Simulation results when $\beta_1 = 1$, $p = 0.4$, $C_1 = 30^{\text{th}}$ percentile and $C_2 = 70^{\text{th}}$ percentile

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	1.011	1.012	1.005	1.604	1.604
			SE	0.021	0.022	0.024	0.035	0.035
			MSE	0.0006	0.0006	0.0006	0.3663	0.3663
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	1.050	1.046	1.038	1.047	1.053
			SE	0.023	0.024	0.025	0.024	0.026
			MSE	0.0031	0.0027	0.0021	0.0027	0.0035
3		(50%, 20%, 30%)	$\hat{\beta}_1$	1.013	1.013	1.006	1.012	1.011
			SE	0.024	0.024	0.026	0.024	0.027
			MSE	0.0007	0.0008	0.0007	0.0007	0.0009
4		(30%, 20%, 50%)	$\hat{\beta}_1$	0.999	1.001	0.997	1.003	0.991
			SE	0.024	0.025	0.026	0.024	0.027
			MSE	0.0006	0.0006	0.0007	0.0006	0.0008
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	0.990	0.993	0.990	0.992	0.996
			SE	0.030	0.031	0.033	0.028	0.039
			MSE	0.0010	0.0010	0.0012	0.0008	0.0015
6		(10%, 40%, 50%)	$\hat{\beta}_1$	1.032	1.035	1.034	1.032	1.026
			SE	0.028	0.029	0.032	0.027	0.037
			MSE	0.0018	0.0020	0.0021	0.0017	0.0021
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	1.030	1.027	1.019	1.027	1.017
			SE	0.028	0.029	0.031	0.029	0.031
			MSE	0.0017	0.0016	0.0013	0.0016	0.0013
8		(10%, 60%, 30%)	$\hat{\beta}_1$	1.029	1.030	1.027	1.027	1.025
			SE	0.031	0.032	0.035	0.031	0.039
			MSE	0.0018	0.0019	0.0019	0.0017	0.0022
9		(30%, 60%, 10%)	$\hat{\beta}_1$	0.982	0.984	0.983	0.984	0.982
			SE	0.033	0.034	0.036	0.032	0.040
			MSE	0.0014	0.0014	0.0016	0.0013	0.0020
10	Simple random sampling	(30%, 40%, 30%)	$\hat{\beta}_1$	0.999	0.999	0.994	0.999	0.994
			SE	0.026	0.027	0.029	0.027	0.029
			MSE	0.0007	0.0007	0.0009	0.0007	0.0009

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

2.4 Simulation Results Under Misspecification of the Distributional Assumption

We assumed that the distribution of the error term in the regression model (2.2) is Normal distribution and the true model and the assumed model were the same in Section 2.3. In practice, the conditional distribution of Y given $X = x$ may not be

known. Therefore, in this section, we investigate the sensitivity of the sampling designs and estimation methods when the distribution of the error term is misspecified. In the simulation study, we generate the error term from Student's t-distribution with degrees of freedom 8 but the other simulation settings remain unchanged.

In Table 2.16, 2.17, and 2.18, we investigate the properties of parameter estimates of the coefficient β_1 of the expensive covariate when the error distribution is misspecified. We set $p = 0.4$, cut-point values set to the 10^{th} and 90^{th} percentiles and $\beta_1 = 0,\ 0.5,\ 1$. The simulation results show that the model misspecification affects the full likelihood and the estimated pseudo-likelihood estimation methods' estimates because they give biased estimates. On the other hand, the bias in the conditional likelihood method's and the IPW estimation method's estimator is much less under many designs. In addition, when there is an association between X and Y, the conditional likelihood method and the IPW method give more efficient estimates under many designs. Thus, they seem to be more robust compared to the full likelihood and the estimated pseudo-likelihood methods.

Table 2.16: Simulation results when $\beta_1 = 0$, $p = 0.4$, $C_1 = 10^{\text{th}}$ percentile, $C_2 = 90^{\text{th}}$ percentile and the distribution of error term is misspecified

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	0.013	0.013	0.017	0.045	0.044
			SE	0.013	0.013	0.017	0.044	0.044
			MSE	0.0003	0.0003	0.0006	0.0039	0.0039
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	0.004	0.004	0.005	0.004	-0.006
			SE	0.014	0.014	0.016	0.014	0.027
			MSE	0.0002	0.0002	0.0003	0.0002	0.0008
3		(50%, 20%, 30%)	$\hat{\beta}_1$	-0.003	-0.003	-0.003	-0.005	-0.021
			SE	0.014	0.014	0.017	0.014	0.028
			MSE	0.0002	0.0002	0.0003	0.0002	0.0012
4		(30%, 20%, 50%)	$\hat{\beta}_1$	0.001	0.001	0.001	-0.004	-0.025
			SE	0.014	0.014	0.017	0.014	0.028
			MSE	0.0002	0.0002	0.0003	0.0002	0.0014
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	-0.007	-0.007	-0.007	-0.005	-0.012
			SE	0.018	0.018	0.020	0.015	0.024
			MSE	0.0004	0.0004	0.0005	0.0003	0.0007
6		(10%, 40%, 50%)	$\hat{\beta}_1$	-0.004	-0.004	-0.004	0.001	-0.028
			SE	0.018	0.017	0.020	0.015	0.024
			MSE	0.0003	0.0003	0.0004	0.0002	0.0013
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	-0.008	-0.008	-0.009	-0.008	-0.008
			SE	0.017	0.016	0.018	0.016	0.021
			MSE	0.0003	0.0003	0.0004	0.0003	0.0005
8		(10%, 60%, 30%)	$\hat{\beta}_1$	-0.021	-0.020	-0.022	-0.019	-0.027
			SE	0.019	0.018	0.020	0.018	0.022
			MSE	0.0008	0.0007	0.0009	0.0007	0.0012
9		(30%, 60%, 10%)	$\hat{\beta}_1$	-0.014	-0.014	-0.015	-0.013	-0.023
			SE	0.019	0.018	0.020	0.018	0.022
			MSE	0.0006	0.0005	0.0006	0.0005	0.0010
10	Simple random sampling	(10%, 80%, 10%)	$\hat{\beta}_1$	-0.025	-0.023	-0.025	-0.023	-0.025
			SE	0.024	0.022	0.023	0.022	0.023
			MSE	0.0012	0.0010	0.0012	0.0010	0.0012

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 2.17: Simulation results when $\beta_1 = 0.5$, $p = 0.4$, $C_1 = 10^{\text{th}}$ percentile, $C_2 = 90^{\text{th}}$ percentile and the distribution of error term is misspecified

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	0.441	0.451	0.573	1.513	1.513
			SE	0.013	0.013	0.017	0.042	0.042
			MSE	0.0037	0.0026	0.0057	1.0273	1.0275
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	0.461	0.464	0.539	0.466	0.523
			SE	0.014	0.014	0.017	0.014	0.028
			MSE	0.0018	0.0015	0.0018	0.0014	0.0013
3		(50%, 20%, 30%)	$\hat{\beta}_1$	0.444	0.445	0.517	0.445	0.484
			SE	0.014	0.014	0.017	0.014	0.029
			MSE	0.0033	0.0032	0.0006	0.0032	0.0011
4		(30%, 20%, 50%)	$\hat{\beta}_1$	0.448	0.454	0.528	0.465	0.537
			SE	0.014	0.015	0.017	0.014	0.029
			MSE	0.0029	0.0024	0.0011	0.0014	0.0022
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	0.443	0.434	0.490	0.445	0.486
			SE	0.017	0.017	0.020	0.016	0.024
			MSE	0.0036	0.0046	0.0005	0.0032	0.0008
6		(10%, 40%, 50%)	$\hat{\beta}_1$	0.425	0.428	0.483	0.445	0.464
			SE	0.017	0.018	0.020	0.015	0.024
			MSE	0.0059	0.0056	0.0007	0.0032	0.0019
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	0.448	0.441	0.483	0.441	0.501
			SE	0.017	0.016	0.018	0.016	0.021
			MSE	0.0030	0.0038	0.0006	0.0038	0.0004
8		(10%, 60%, 30%)	$\hat{\beta}_1$	0.440	0.432	0.471	0.436	0.463
			SE	0.018	0.018	0.020	0.017	0.022
			MSE	0.0039	0.0050	0.0012	0.0044	0.0018
9		(30%, 60%, 10%)	$\hat{\beta}_1$	0.456	0.442	0.484	0.443	0.504
			SE	0.019	0.018	0.021	0.018	0.022
			MSE	0.0023	0.0037	0.0007	0.0035	0.0005
10	Simple random sampling	(10%, 80%, 10%)	$\hat{\beta}_1$	0.439	0.423	0.453	0.423	0.453
			SE	0.022	0.022	0.024	0.022	0.024
			MSE	0.0042	0.0064	0.0028	0.0064	0.0028

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 2.18: Simulation results when $\beta_1 = 1$, $p = 0.4$, $C_1 = 10^{\text{th}}$ percentile, $C_2 = 90^{\text{th}}$ percentile and the distribution of error term is misspecified

#	Sampling type	Sampling percentage		Likelihood-based methods			Pseudo likelihood -based methods	
				β_{1R}	β_{1F}	β_{1C1}	β_{1P}	β_{1W}
1	Sampling from extreme strata only	(50%, 0%, 50%)	$\hat{\beta}_1$	0.854	0.869	1.050	2.768	2.768
			SE	0.014	0.014	0.019	0.038	0.037
			MSE	0.0214	0.0173	0.0028	3.1268	3.1268
2	Oversampling from extreme strata	(40%, 20%, 40%)	$\hat{\beta}_1$	0.857	0.866	0.969	0.869	0.982
			SE	0.014	0.015	0.018	0.015	0.030
			MSE	0.0207	0.0182	0.0013	0.0174	0.0012
3		(50%, 20%, 30%)	$\hat{\beta}_1$	0.854	0.859	0.962	0.859	1.039
			SE	0.015	0.015	0.018	0.015	0.029
			MSE	0.0214	0.0201	0.0018	0.0200	0.0024
4		(30%, 20%, 50%)	$\hat{\beta}_1$	0.847	0.861	0.966	0.859	1.009
			SE	0.015	0.015	0.019	0.014	0.030
			MSE	0.0235	0.0196	0.0015	0.0201	0.0010
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	$\hat{\beta}_1$	0.873	0.874	0.957	0.876	1.019
			SE	0.017	0.017	0.021	0.017	0.025
			MSE	0.0163	0.0161	0.0023	0.0157	0.0010
6		(10%, 40%, 50%)	$\hat{\beta}_1$	0.861	0.883	0.966	0.899	0.996
			SE	0.017	0.018	0.021	0.015	0.025
			MSE	0.0195	0.0141	0.0016	0.0103	0.0006
7	Oversampling from middle stratum	(25%, 50%, 25%)	$\hat{\beta}_1$	0.882	0.887	0.948	0.888	1.006
			SE	0.016	0.017	0.019	0.017	0.022
			MSE	0.0142	0.0132	0.0031	0.0127	0.0005
8		(10%, 60%, 30%)	$\hat{\beta}_1$	0.874	0.885	0.943	0.885	0.980
			SE	0.018	0.018	0.021	0.017	0.022
			MSE	0.0163	0.0137	0.0037	0.0136	0.0009
9		(30%, 60%, 10%)	$\hat{\beta}_1$	0.883	0.886	0.947	0.886	0.991
			SE	0.018	0.018	0.021	0.018	0.023
			MSE	0.0141	0.0134	0.0032	0.0134	0.0006
10	Simple random sampling	(10%, 80%, 10%)	$\hat{\beta}_1$	0.913	0.934	0.981	0.934	0.981
			SE	0.020	0.020	0.024	0.020	0.024
			MSE	0.0079	0.0047	0.0009	0.0047	0.0009

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Chapter 3

Efficiency of Two-Phase Stratified Response-Dependent Sampling Designs

In the previous chapter, we considered response-dependent BSS designs where the phase II sampling probabilities depend only on the response variable. However, if there are some inexpensive covariates, then the sampling which depends on both the response variable and inexpensive covariates might be more efficient than the sampling that depends only on the response variable (Breslow and Chatterjee, 1999; Schaid et al., 2013; Espin-Garcia et al., 2018). For example, Espin-Garcia et al. (2018) constructed some phase II sampling designs according to the values of a response variable Y which is a quantitative trait and an inexpensive covariate Z which denotes genotypes of a single nucleotide polymorphism (SNP) and is a categorical variable with three levels. In phase I, the quantitative trait values (Y) and genotype of the SNP (Z) are observed for every subject in the study. To select the phase II sample, they discretized the quantitative trait values (Y) into three strata using some fixed cut-point values under each level of Z. Thus, they obtained nine strata determined by the Y and Z values. They considered four sampling designs in phase II to select sampling units from each stratum: proportional to stratum size, extreme, balanced, and combined sampling designs. In the proportional to stratum size sampling design, the sample selected from each stratum is proportional to the stratum size. In extreme

sampling design, equal sized samples are selected only from the four strata with extreme values of Z and Y. The same number of subjects from each strata are selected in the balanced sampling. In the combined sampling design, they combined balanced selection in Z and extreme selection in Y, that is, equal number of samples are selected from the extreme strata under each level of Z. Their extreme strata sizes were set to large compared to the middle stratum. They compared efficiency of Y- and Z-dependent sampling designs to Y-dependent sampling or Z-dependent sampling designs using semiparametric maximum likelihood estimation method. They found that the combined Y- and Z-dependent sampling design exhibits better power compared to Y-dependent sampling or Z-dependent sampling designs.

In our study, we assume that Z is a categorical variable with three levels and we consider one level of Z be rare. In Chapter 2, when sampling depends only on the response variable, likelihood-based methods yield the most efficient estimates under the extreme sampling design, the pseudo-likelihood methods yield the most efficient estimates when more units are sampled from the central stratum. Therefore, in this chapter, our aim is to assess the efficiency of such designs when the phase II sampling depends also on an inexpensive covariate in addition to the response variable. The efficiency of these two designs will be compared with the simple random sampling. In our study, extreme strata sizes are set to be small compared to the middle stratum to understand the importance of sampling from the extreme strata. We consider likelihood-based and pseudo-likelihood estimation methods under stratified response-dependent two-phase sampling designs for two different settings where the inexpensive covariate Z and the expensive covariate X are independent and dependent. We compare efficiency of sampling designs under each estimation method and each setting.

3.1 Two-Phase Stratified Response-Dependent BSS Design

In two-phase stratified response-dependent BSS design, suppose there are N units generated from the model (1.1) in phase I where Z is a discrete variable with M levels. Suppose there are N_m $(m = 0, 1, \ldots, M-1)$ units observed in the m^{th} level

of Z with $\sum_{m=0}^{M-1} N_m = N$. Then, the observed response values y in each level of Z are partitioned into K strata using fixed cut-point values. Thus, the total number of strata is $M \times K$. Suppose the observed response values $(y_{m1}, y_{m2}, \ldots, y_{mN_m})$ in the m^{th} level of Z is partitioned into K strata, $S_{mk},\ k = 1, 2, \ldots, K$, using fixed cut-point values $C_{mj},\ j = 1, 2, \ldots, K-1$, where $C_{m1} \leq C_{m2} \leq \ldots \leq C_{mK-1}$. The first stratum includes the units with y_{mi} values less than C_{m1}, the k^{th} $(k = 2, \ldots, K-1)$ stratum includes the units with y_{mi} values between C_{mk-1} and C_{mk}, and the K^{th} stratum includes the units with y_{mi} values greater than C_{mk-1}. Let N_{mj} $(j = 1, 2, \ldots, K)$ be the number of units in the j^{th} stratum of m^{th} level of Z with $\sum_{j=1}^{K} N_{mj} = N_m$. Then, from each stratum S_{mj}, n_{mj} units are randomly selected for inclusion in phase II, where the number of units selected from the m^{th} level of Z is $n_m = \sum_{j=1}^{K} n_{mj}$ and the total phase II sample size $n = \sum_{m=0}^{M-1} n_m$ is fixed according to the budgetary constraint.

The selection probability for the i^{th} unit, π_i in (1.3), becomes $\pi_i = \sum_{m=0}^{M-1} \sum_{j=1}^{K} \delta_{imj} p_{mj}$ for $i = 1, 2, \ldots, N$, where

$$\delta_{imj} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ unit has the } m^{th} \text{ level of Z} \\ & \text{and is in the } j^{\text{th}} \text{ stratum of the } m^{th} \text{ level of } Z, \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

and

$$p_{mj} = \frac{n_{mj}}{N_{mj}},\ \ m = 0, 1, \ldots, M-1,\ \ j = 1, 2, \ldots, K.$$

3.2 Phase II Sampling Design

Suppose the continuous response variable observations $\{y_1, y_2, \ldots, y_N\}$ and inexpensive covariate observations $\{z_1, z_2, \ldots, z_N\}$ are available for a cohort of size N in phase I. In our study, Z is a discrete variable with $M = 3$ levels. Suppose there are N_m individuals with $Z = m$ $(m = 0, 1, 2)$ and n_m units are selected in phase

II to obtain their expensive covariate data. To select these units, response variable observations in the m^{th} level of Z are partitioned into $K = 3$ number of strata based on the fixed constants C_{m1} and C_{m2} as follows:

$$\underbrace{y_1 \leq \ldots \leq y_{N_{m1}}}_{\substack{\text{Low stratum} \\ (S_{m1})}} \leq C_{m1} \leq \underbrace{y_{N_{m1+1}} \leq \ldots \leq y_{N_{m1}+N_{m2}}}_{\substack{\text{Middle stratum} \\ (S_{m2})}} \leq C_{m2} \leq \underbrace{y_{N_{m1}+N_{m2}+1} \leq \ldots \leq y_{N_m}}_{\substack{\text{High stratum} \\ (S_{m3})}} \tag{3.2}$$

Suppose there are N_{m1} units in the first stratum S_{m1} which includes units with $Z = m$ and lower response values than C_{m1}, N_{m2} units in the second stratum S_{m2} which includes units with $Z = m$ and response values between C_{m1} and C_{m2}, and $N_{m3} = N_m - N_{m1} - N_{m2}$ units in the third stratum S_{m3} which includes units with $Z = m$ and response values greater than C_{m2}. The first and third strata consist of extreme values for the m^{th} level of Z, where in this study the sizes of the extreme strata are set small compared to central stratum to understand the importance of sampling from the extreme stratum. Similarly, we create three strata for each level of Z. Altogether 9 strata are constructed. Then, stratified response-dependent BSS design described in Section 3.1 is applied to select the phase II sample. That is, from each stratum S_{mj}, n_{mj} units are randomly selected for inclusion in the phase II sample, where the total phase II sample size $n = \sum_{m=0}^{2} \sum_{j=1}^{3} n_{mj}$ is fixed according to the budgetary constraint.

In this study, Z can get values 0, 1 or 2. We assume that $Z = 0$ is a rare category and we investigate three sampling designs to select sampling units from each stratum within the framework of phase II design. In design (A), we consider sampling all of the units from the $Z = 0$ group and sampling all the extreme strata subjects under the other two groups $Z = 1$ and $Z = 2$. In design (B), we consider sampling all of the units from the $Z = 0$ group and very small number of subjects are selected from the extreme strata but more sampling units are selected from the middle stratum from the other two groups $Z = 1$ and $Z = 2$. Design (C) is a simple random sampling design where sampling probability for each unit is same. In all these designs the number of samples selected under each level of Z is proportional to total number of units in that level except the rare level category. Also, we assume the available budget to measure the expensive covariate is only for n individuals. Therefore, the phase II sample size $n = \sum_{m=0}^{2} n_m$ is the same over each scenario.

3.3 Model Description

Suppose that the data $\{(y_i, z_i, x_i) : \ i = 1, 2, \ldots, N\}$ comes from the linear model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i, \tag{3.3}$$

where β_0, β_1, β_2 are the regression coefficients and ϵ_i is a random error for subject i $(i = 1, 2, \ldots, N)$.

In our study, first we assume X is a Bernoulli random variable with probability of having $X = 1$ being p, Z is a discrete random variable with three levels and $Z = 0$ is a rare category. ϵ_i's are independently and identically normally distributed with mean 0 and variance σ^2. Therefore, Y_i has a normal distribution with mean $\beta_0 + \beta_1 x_i + \beta_2 z_i$ and variance σ^2 when $X_i = x_i$ and $Z_i = z_i$ are observed. We consider two settings: X and Z are independent and dependent. When X and Z are dependent, the conditional distribution of X given $Z = z$ is Bernoulli with probability of having $X = 1$ being p_z, $z = 0, 1, 2$, where $p_j \neq p_k$ for some $j \neq k$ with $j < k$.

We have four unknown parameters, β_0, β_1, β_2, and σ. To estimate $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \sigma)$, we obtain the samples by applying the proposed sampling designs presented in Section 3.2. We consider the five different estimation methods discussed in Section 1.4 to estimate the parameter vector $\boldsymbol{\theta}$. Under each estimation method, we aim to check whether the estimator of β_1 is an unbiased estimator of β_1 and to investigate the efficiency of the estimators under each sampling design setting.

Under the model in (2.2) with the specified assumptions,

$$f(y_i|x_i, z_i; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2}{2\sigma^2}}, \tag{3.4}$$

where $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \sigma)$, and

$$g(x_i|z; p_z) = \begin{cases} p_z & \text{for } x_i = 1 \\ 1 - p_z & \text{for } x_i = 0 \end{cases}$$

are the conditional density function for Y_i given $X_i = x_i$ and $Z_i = z_i$ and the conditional mass function of X_i given $Z = z$, respectively. Under the independence

assumption between X and Z, the conditional mass function of X given Z is $g(x|z;p_z) = g(x;p)$ where $p_z = p$ for $z = 0, 1, 2$.

3.4 Estimation Methods

The full likelihood $L_R(\boldsymbol{\theta}, \boldsymbol{p})$ which incorporates both complete and incomplete data $\{(y_i, x_i, z_i) : i \in V\} \cup \{(y_i, z_i) : i \in \bar{V}\}$ under the missing at random assumption was given in (1.6). Under the current setting, the logarithm of the likelihood function becomes

$$\begin{aligned} l_R(\boldsymbol{\theta}, \boldsymbol{p}) = &\sum_{i \in V} \sum_{m=0}^{2} \delta_{im} [\log f(y_i|x_i, z_i; \boldsymbol{\theta}) + \log g(x_i|z_i; p_m)] \\ &+ \sum_{i \in \bar{V}} \sum_{m=0}^{2} \delta_{im} \left(\log[p_m f(y_i|x = 1, z_i; \boldsymbol{\theta}) + (1 - p_m) f(y_i|x = 0, z_i; \boldsymbol{\theta})]\right), \end{aligned} \tag{3.5}$$

where $\boldsymbol{p} = (p_0, p_1, p_2)$ and

$$\delta_{im} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ unit has the } m^{th} \text{ level of Z,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

The maximum likelihood estimates of $\boldsymbol{\theta}$ and $\boldsymbol{p}$ are obtained by maximizing $l_R(\boldsymbol{\theta}, \boldsymbol{p})$ in (3.5) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{p}$. We denote the resulting estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_R$.

The likelihood $L_F(\boldsymbol{\theta}, \boldsymbol{p})$ which incorporates complete data $\{(y_i, x_i, z_i) : i \in V\}$ and only the stratum information for unobserved data was given in (1.8). Under the current setting, the logarithm of the likelihood function becomes

$$\begin{aligned} l_F(\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{m=0}^{2} \sum_{j=1}^{3} \Bigg\{ &\sum_{i \in D_{mj}} [\log f(y_i|x_i, z_i; \boldsymbol{\theta}) + \log g(x_i|z_i; p_m)] \\ &+ (N_{mj} - n_{mj}) \log \Bigg(p_m \int_{C_{mj-1}}^{C_{mj}} f(y|x = 1, z = m; \boldsymbol{\theta}) dy \\ &+ (1 - p_m) \int_{C_{mj-1}}^{C_{mj}} f(y|x = 0, z = m; \boldsymbol{\theta}) dy \Bigg) \Bigg\}, \end{aligned} \tag{3.7}$$

where $D_{mj} = \{i : \delta_{imj} = 1, R_i = 1\}$ denotes the set of indices of all fully observed units in stratum S_{mj}, δ_{imj} is defined in (3.1), C_{m1}, C_{m2} are fixed cut-point values and $C_{m0} = -\infty$, $C_{m3} = \infty$ when $Z = m$. The maximum likelihood estimates of $\boldsymbol{\theta}$ and $\boldsymbol{p}$ are obtained by maximizing $l_F(\boldsymbol{\theta}, \boldsymbol{p})$ in (3.7) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{p}$. We denote the resulting estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_F$.

The likelihood of complete data $\{(y_i, x_i, z_i) : i \in V\}$, $L_{C0}(\boldsymbol{\theta}, \boldsymbol{p})$, was given in (1.10). Under the current setting, the logarithm of the likelihood function becomes

$$\begin{aligned} l_{C0}(\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{i:R_i=1} \Bigg\{ & \sum_{m=0}^{2} \delta_{im} \left[\log f(y_i | x_i, z_i; \boldsymbol{\theta}) + \log g(x_i | z_i, p_m) + \log h(z_i) + \log \pi_i \right] \\ & - \log \sum_{m=0}^{2} \sum_{j=1}^{3} \frac{n_{mj}}{N_{mj}} \frac{N_m}{N} \Bigg(p_m \int_{C_{mj-1}}^{C_{mj}} f(y | x = 1, z = m; \boldsymbol{\theta}) dy \\ & + (1 - p_m) \int_{C_{mj-1}}^{C_{mj}} f(y | x = 0, z = m; \boldsymbol{\theta}) dy \Bigg) \Bigg\}, \end{aligned} \tag{3.8}$$

where δ_{im} is defined in (3.6), C_{m1}, C_{m2} are fixed cut-point values and $C_{m0} = -\infty$, $C_{m3} = \infty$ when $Z = m$. The estimate of $\boldsymbol{\theta}$ and $\boldsymbol{p}$ are obtained by maximizing $l_{C0}(\boldsymbol{\theta}, \boldsymbol{p})$ in (3.8) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{p}$. We denote the resulting estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_{C0}$.

The likelihood of complete data $\{(y_i, x_i, z_i) : i \in V\}$, $L_{C1}(\boldsymbol{\theta})$, which is a conditional likelihood was given in (1.11). Under the current setting, the logarithm of the likelihood function becomes

$$l_{C1}(\boldsymbol{\theta}) = \sum_{i:R_i=1} \left\{ \log f(y_i | x_i, z_i; \boldsymbol{\theta}) + \log \pi_i - \log \left(\sum_{m=0}^{2} \sum_{l=1}^{3} \delta_{im} \frac{n_{ml}}{N_{ml}} \int_{C_{ml-1}}^{C_{ml}} f(y | x_i, z_i; \boldsymbol{\theta}) dy \right) \right\}, \tag{3.9}$$

where δ_{im} is defined in (3.6), C_{m1}, C_{m2} are fixed cut-point values and $C_{m0} = -\infty$, $C_{m3} = \infty$ when $Z = m$. The estimate of $\boldsymbol{\theta}$, is obtained by maximizing $l_{C1}(\boldsymbol{\theta})$ in (3.9) with respect to $\boldsymbol{\theta}$. We denote the resulting estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_{C1}$.

Under the current setting, the logarithm of the pseudo-likelihood function of $l_P(\boldsymbol{\theta})$

given in (1.14) becomes

$$l_P(\boldsymbol{\theta}) = \sum_{i:R_i=1} \log f(y_i|x_i, z_i; \boldsymbol{\theta}) + \sum_{m=0}^{2}\sum_{j=1}^{3}(N_{mj} - n_{mj}) \log \left(\sum_{k=0}^{2}\sum_{l=1}^{3} \frac{N_{kl}}{n_{kl}} \sum_{i \in D_{kl}} \int_{C_{mj-1}}^{C_{mj}} \delta_{im} f(y|x_i, z_i; \boldsymbol{\theta}) dy \right), \tag{3.10}$$

where δ_{im} is defined in (3.6), C_{m1}, C_{m2} are fixed cut-point values and $C_{m0} = -\infty$, $C_{m3} = \infty$ when $Z = m$. The estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_P$, is obtained by maximizing $l_P(\boldsymbol{\theta})$ in (3.10) with respect to $\boldsymbol{\theta}$. We denote the resulting estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_P$.

The IPW method estimating function given in (1.15) becomes

$$S_w(\boldsymbol{\theta}) = \begin{pmatrix} \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{1i}(\boldsymbol{\theta}) \\ \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{2i}(\boldsymbol{\theta}) \\ \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{3i}(\boldsymbol{\theta}) \\ \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{4i}(\boldsymbol{\theta}) \end{pmatrix}, \tag{3.11}$$

where

$$\begin{aligned} U_{1i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_i|x_i, z_i; \boldsymbol{\theta})}{\partial \beta_0} = \frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)}{\sigma^2}, \\ U_{2i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_i|x_i, z_i; \boldsymbol{\theta})}{\partial \beta_1} = \frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)x_i}{\sigma^2}, \\ U_{3i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_i|x_i, z_i; \boldsymbol{\theta})}{\partial \beta_2} = \frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)z_i}{\sigma^2}, \\ U_{4i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_i|x_i, z_i; \boldsymbol{\theta})}{\partial \sigma} = \frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2}{\sigma^3} - \frac{1}{\sigma}. \end{aligned}$$

By solving the estimating equations $S_w(\boldsymbol{\theta}) = 0$, we obtain the IPW estimators $\hat{\beta}_{0,w}$, $\hat{\beta}_{1,w}$, $\hat{\beta}_{2,w}$ and $\hat{\sigma}^2_w$.

Under some regularity conditions, the covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta})$ given in

(1.17) becomes

$$C(\boldsymbol{\theta}) = A(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}) [A(\boldsymbol{\theta})^{-1}]',$$

where

$$A(\boldsymbol{\theta}) = -\frac{1}{n}\left(\frac{\partial S_w}{\partial \boldsymbol{\theta}'}\right) = \begin{pmatrix} \frac{1}{n\sigma^2}\sum_{i=1}^{N}\frac{R_i}{\pi_i} & \frac{1}{n\sigma^2}\sum_{i=1}^{N}\frac{R_i}{\pi_i}x_i & \frac{1}{n\sigma^2}\sum_{i=1}^{N}\frac{R_i}{\pi_i}z_i & 0 \\ \frac{1}{n\sigma^2}\sum_{i=1}^{N}\frac{R_i}{\pi_i}x_i & \frac{1}{n\sigma^2}\sum_{i=1}^{N}\frac{R_i}{\pi_i}x_i^2 & \frac{1}{n\sigma^2}\sum_{i=1}^{N}\frac{R_i}{\pi_i}x_i z_i & 0 \\ \frac{1}{n\sigma^2}\sum_{i=1}^{N}\frac{R_i}{\pi_i}z_i & \frac{1}{n\sigma^2}\sum_{i=1}^{N}\frac{R_i}{\pi_i}x_i z_i & \frac{1}{n\sigma^2}\sum_{i=1}^{N}\frac{R_i}{\pi_i}z_i^2 & 0 \\ 0 & 0 & 0 & \frac{2}{n\sigma^2}\sum_{i=1}^{N}\frac{R_i}{\pi_i} \end{pmatrix},$$

and

$$B(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{N} S_{wi}(\boldsymbol{\theta}) S_{wi}(\boldsymbol{\theta})' = \begin{pmatrix} B_{11} & B_{12} & B_{13} & B_{14} \\ B_{21} & B_{22} & B_{23} & B_{24} \\ B_{31} & B_{32} & B_{33} & B_{34} \\ B_{41} & B_{42} & B_{43} & B_{44} \end{pmatrix},$$

where

$$B_{11} = \frac{1}{n}\sum_{i=1}^{N}\frac{R_i}{\pi_i^2}\frac{(y_i - \beta_0 - \beta_1 x_i - \beta_z z_i)^2}{\sigma^4},$$

$$B_{22} = \frac{1}{n}\sum_{i=1}^{N}\frac{R_i}{\pi_i^2}\frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2 x_i^2}{\sigma^4},$$

$$B_{33} = \frac{1}{n}\sum_{i=1}^{N}\frac{R_i}{\pi_i^2}\frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2 z_i^2}{\sigma^4},$$

$$B_{44} = \frac{1}{n}\sum_{i=1}^{N}\frac{R_i}{\pi_i^2}\left[\frac{(y_i - \beta_0 - \beta_1 x_i - \beta_1 z_i)^2}{\sigma^3} - \frac{1}{\sigma}\right]^2,$$

$$B_{12} = B_{21} = \frac{1}{n}\sum_{i=1}^{N}\frac{R_i}{\pi_i^2}\frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2 x_i}{\sigma^4},$$

$$B_{13} = B_{31} = \frac{1}{n}\sum_{i=1}^{N} \frac{R_i}{\pi_i^2} \frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2 z_i}{\sigma^4},$$

$$B_{14} = B_{41} = \frac{1}{n}\sum_{i=1}^{N} \frac{R_i}{\pi_i^2} \frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)}{\sigma^2} \left[\frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2}{\sigma^3} - \frac{1}{\sigma}\right],$$

$$B_{23} = B_{32} = \frac{1}{n}\sum_{i=1}^{N} \frac{R_i}{\pi_i^2} \frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2 x_i z_i}{\sigma^4},$$

$$B_{24} = B_{42} = \frac{1}{n}\sum_{i=1}^{N} \frac{R_i}{\pi_i^2} \frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i) x_i}{\sigma^2} \left[\frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2}{\sigma^3} - \frac{1}{\sigma}\right],$$

$$B_{34} = B_{43} = \frac{1}{n}\sum_{i=1}^{N} \frac{R_i}{\pi_i^2} \frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i) z_i}{\sigma^2} \left[\frac{(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2}{\sigma^3} - \frac{1}{\sigma}\right].$$

The covariance matrix of $\hat{\boldsymbol{\theta}}_w = (\hat{\beta}_{0,w}, \hat{\beta}_{1,w}, \hat{\beta}_{2,w}, \hat{\sigma}_w)$ is $\frac{1}{n}C(\boldsymbol{\theta})$. A consistent estimator of $C(\boldsymbol{\theta})$ is obtained by plugging the estimates of $\boldsymbol{\theta}$ in the sandwich form given in (1.18) and the estimated covariance matrix of estimators is obtained as

$$\frac{1}{n}C(\hat{\boldsymbol{\theta}}_w).$$

If the inexpensive covariate Z and expensive covariate X are independent, then $g(x_i|z_i; p_m)$ is replaced by $g(x_i; p)$ in all the estimation methods where $p = p_0 = p_1 = p_2$.

3.5 Simulation Study

A simulation study was conducted to investigate properties of estimation methods under the sampling design settings described in Section 3.2. In Chapter 2, the sampling was only based on the response variable. Likelihood-based methods yield the most efficient estimates under the extreme sampling design, the pseudo-likelihood methods yield the most efficient estimates when sampling more units from the central stratum. In this chapter, our aim is to assess efficiency of such designs when the phase II sampling depends also on an inexpensive covariate in addition to the response variable. Efficiency of these two designs will also be compared with simple random

sampling. Among these three designs, we aim to identify the sampling design which gives the most efficient estimate of the coefficient of expensive covariate under each estimation method and to compare the performance of the estimation methods. We assume that there is a cohort of size $N = 50,000$ with their observed response variable Y values and inexpensive covariate values Z in the first phase. We are interested in making inference on the association between the response variable Y and the expensive covariate X. We consider a linear regression of Y on X and Z with a normally distributed error term.

In data generation, we first generated error terms of size $N = 50,000$ from Normal distribution with mean 0 and variance $\sigma^2 = 2$, inexpensive covariate values z_i $(i = 1, 2, \ldots, N)$ from Multinomial distribution with $P(Z_i = 0) = 0.05$, $P(Z_i = 1) = 0.45$, and $P(Z_i = 2) = 0.50$. Under the independence assumption between X and Z, we generated the expensive covariate values x_i from Bernoulli distribution with $p = P(X_i = 1) = 0.4$. Under the dependence assumption between inexpensive covariate Z and expensive covariate X, we generated the covariate values x_i $(i = 1, 2, \ldots, N)$ from the Bernoulli distributions assumed for the conditional distribution of $X|Z = z$ with $p_0 = P(X_i = 1|Z_i = 0) = 0.05$, $p_1 = P(X_i = 1|Z_i = 1) = 0.4$, and $p_2 = p(X_i = 1|Z_i = 2) = 0.2$. Next, we generated the response variable values $y_i's$ using the model (3.3), where we choose different values of $\beta_1(= 0, 0.5, 1)$, the intercept is set to $\beta_0 = 10$ and the coefficient of inexpensive covariate is set to $\beta_2 = 2$. In our study, we set the cut-point values (C_{m1}, C_{m2}) in (3.2) to the (10^{th}, 90^{th}) percentiles of the response distribution which can divide the cohort into three strata under each level of $Z = m$ $(m = 0, 1, 2)$ and the stratum sizes are given in Table 3.1. The cut-point values C_{m1} and C_{m2} for each level of $Z = m$ are given in Table 3.2.

We applied the stratified response-dependent BSS design described in Section 3.2 to obtain the phase II samples and for only the selected samples, we assume to have the expensive covariate X values. In each sampling design, the phase II sample size is set to $n = 10,000$ and the stratum specific sample sizes are given in Table 3.3. The estimated value of β_1 is obtained under each sampling design and under each estimation method described in Section 3.4.

Table 3.1: Stratum sizes under stratified response-dependent BSS design

$Z = m$	Stratum No (j)	Z and X are independent N_{mj}			Z and X are dependent N_{mj}		
		$\beta_1 = 0$	$\beta_1 = 0.5$	$\beta_1 = 1$	$\beta_1 = 0$	$\beta_1 = 0.5$	$\beta_1 = 1$
$m = 0$	$j = 1$	246	246	246	241	241	241
	$j = 2$	1966	1966	1966	1927	1927	1927
	$j = 3$	246	246	246	241	241	241
$m = 1$	$j = 1$	2259	2259	2259	2252	2252	2252
	$j = 2$	18066	18066	18066	18008	18008	18008
	$j = 3$	2259	2259	2259	2252	2252	2252
$m = 2$	$j = 1$	2496	2496	2496	2508	2508	2508
	$j = 2$	19966	19966	19966	20063	20063	20063
	$j = 3$	2496	2496	2496	2508	2508	2508

Table 3.2: The cut-point values C_{m1} and C_{m2} for different β_1 values

Assumption	$Z = m$	$\beta_1 = 0$		$\beta_1 = 0.5$		$\beta_1 = 1$	
		C_{m1}	C_{m2}	C_{m1}	C_{m2}	C_{m1}	C_{m2}
Z and X are independent	$m = 0$	8.232	11.719	8.411	11.921	8.504	12.224
	$m = 1$	10.186	13.836	10.368	14.060	10.485	14.355
	$m = 2$	12.181	15.798	12.351	16.026	12.468	16.299
Z and X are dependent	$m = 0$	8.150	11.766	8.165	11.803	8.179	11.856
	$m = 1$	10.215	13.775	10.397	14.011	10.522	14.303
	$m = 2$	12.162	15.853	12.251	15.965	12.305	16.135

Table 3.3: Stratum-specific sample sizes under each stratified response-dependent BSS design

Design	$Z = m$	Stratum No K=j	Z and X are independent n_{mj}			Z and X are dependent n_{mj}		
			$\beta_1 = 0$	$\beta_1 = 0.5$	$\beta_1 = 1$	$\beta_1 = 0$	$\beta_1 = 0.5$	$\beta_1 = 1$
Design A	$m = 0$	$j = 1$	246	246	246	241	241	241
		$j = 2$	1966	1966	1966	1927	1927	1927
		$j = 3$	246	246	246	241	241	241
	$m = 1$	$j = 1$	1697	1697	1697	1708	1708	1708
		$j = 2$	0	0	0	0	0	0
		$j = 3$	1697	1697	1697	1708	1708	1708
	$m = 2$	$j = 1$	2074	2074	2074	2088	2088	2088
		$j = 2$	0	0	0	0	0	0
		$j = 3$	2074	2074	2074	2087	2087	2087
Design B	$m = 0$	$j = 1$	246	246	246	241	241	241
		$j = 2$	1966	1966	1966	1927	1927	1927
		$j = 3$	246	246	246	241	241	241
	$m = 1$	$j = 1$	339	339	339	342	342	342
		$j = 2$	2715	2715	2715	2733	2733	2733
		$j = 3$	340	340	340	341	341	341
	$m = 2$	$j = 1$	415	415	415	418	418	418
		$j = 2$	3318	3318	3318	3340	3340	3340
		$j = 3$	415	415	415	417	417	417
Design C*	$m = 0$	$j = 1$	47	47	46	54	53	54
		$j = 2$	392	392	392	394	394	393
		$j = 3$	49	49	50	51	52	52
	$m = 1$	$j = 1$	458	463	455	476	472	473
		$j = 2$	3637	3623	3636	3540	3540	3542
		$j = 3$	433	442	437	463	467	464
	$m = 2$	$j = 1$	510	500	509	529	515	519
		$j = 2$	3966	3970	3971	3984	3997	4014
		$j = 3$	508	514	504	509	510	489

* Under simple random sampling, the sampling probability of each unit is same and n_{mj}'s were obtained based on the sampling probability.

3.6 Simulation Results

The parameter estimates of β_1 and their standard error estimates and mean square error estimates under different estimation methods are given from Table 3.4 to Table 3.10 for different β_1 values. Table 3.4 shows the results under the maximum likelihood estimation using the full cohort data. We obtained consistent maximum likelihood estimates. Under Z and X are independent settings we obtained more efficient estimates compare to the setting where Z and X are dependent.

Table 3.5, 3.6, and 3.7 show properties of different parameter estimators of the coefficient β_1 of the expensive covariate when Z and X are independent, $\beta_1 = 0$, 0.5, 1, respectively, $p = 0.4$ and cut-point values C_1 and C_2 set to the 10^{th} and 90^{th} percentiles. The likelihood-based methods give consistent and relatively efficient estimates under each sampling design considered. In addition, likelihood-based methods provide the most efficient estimates when sampling all of the units from the rare group and sampling all the extreme strata units under the other two groups of Z (Design A). On the other hand, the pseudo-likelihood methods give biased estimates under this sampling design except the case when there is no association between X and Y. Simple random sampling (Design C) give consistent and most efficient estimates under the IPW method. Under the estimated pseudo-likelihood estimation method, both the simple random sampling and the sampling in which selecting all of the units from the rare group and selecting more units from the middle stratum of the other two groups of Z (Design B) lead to efficient estimates. The conditional likelihoods yield a little less efficient estimates under Design B and C compared to the full likelihoods and the estimated pseudo-likelihood. When we compare the estimates under the most efficient design with the full cohort, we observe an approximately 40% loss in efficiency when the sample size reduces to 10,000 from 50,000.

Table 3.8 to Table 3.10 show properties of different parameter estimators of the coefficient β_1 of the expensive covariate when Z and X are dependent, $\beta_1 = 0$, 0.5, 1, respectively, $p = 0.4$ and cut-point values C_1 and C_2 set to the 10^{th} and 90^{th} percentiles. We observe similar results to those obtained when Z and X are independent.

Table 3.4: Maximum likelihood estimation results under the full cohort

Assumption	β_1	Estimates		
		$\hat{\beta}_1$	SE	MSE
Z and X	0	0.008	0.013	0.0002
are	0.5	0.508	0.013	0.0002
independent	1	1.008	0.013	0.0002
Z and X	0	-0.024	0.014	0.0008
are	0.5	0.476	0.014	0.0008
dependent	1	0.976	0.014	0.0008

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 3.5: Simulation results when Z and X are independent and $\beta_1 = 0$

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	0.009	0.009	0.009	0.009	0.024	0.032
		SE	0.018	0.018	0.017	0.017	0.047	0.049
		MSE	0.0004	0.0004	0.0004	0.0004	0.0028	0.0034
2	Design B	$\hat{\beta}_1$	-0.035	-0.035	-0.034	-0.034	-0.035	-0.016
		SE	0.029	0.029	0.029	0.029	0.029	0.032
		MSE	0.0020	0.0020	0.0020	0.0020	0.0021	0.0013
2	Design C	$\hat{\beta}_1$	-0.009	-0.009	-0.009	-0.009	-0.010	-0.009
		SE	0.029	0.029	0.029	0.029	0.029	0.029
		MSE	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 3.6: Simulation results when Z and X are independent and $\beta_1 = 0.5$

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	0.524	0.524	0.522	0.522	1.389	1.443
		SE	0.018	0.018	0.018	0.018	0.046	0.047
		MSE	0.0009	0.0009	0.0008	0.0008	0.7921	0.8907
2	Design B	$\hat{\beta}_1$	0.488	0.486	0.482	0.482	0.487	0.504
		SE	0.029	0.028	0.029	0.029	0.028	0.032
		MSE	0.0009	0.0010	0.0012	0.0012	0.0010	0.0010
2	Design C	$\hat{\beta}_1$	0.490	0.488	0.489	0.489	0.487	0.491
		SE	0.028	0.028	0.029	0.029	0.028	0.029
		MSE	0.0009	0.0009	0.0010	0.0010	0.0010	0.0009

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 3.7: Simulation results when Z and X are independent and $\beta_1 = 1$

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	1.004	1.004	1.000	0.996	2.566	2.653
		SE	0.018	0.018	0.020	0.021	0.043	0.043
		MSE	0.0003	0.0003	0.0004	0.0004	2.4534	2.7351
2	Design B	$\hat{\beta}_1$	0.992	0.987	0.976	0.976	0.987	0.995
		SE	0.027	0.027	0.029	0.029	0.027	0.032
		MSE	0.0008	0.0009	0.0014	0.0014	0.0009	0.0010
2	Design C	$\hat{\beta}_1$	0.992	0.991	0.991	0.991	0.990	0.991
		SE	0.026	0.027	0.029	0.029	0.027	0.029
		MSE	0.0008	0.0008	0.0009	0.0009	0.0008	0.0009

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 3.8: Simulation results when Z and X are dependent and $\beta_1 = 0$

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	0.001	-0.002	-0.002	-0.002	-0.007	-0.010
		SE	0.019	0.019	0.019	0.019	0.055	0.059
		MSE	0.0004	0.0004	0.0004	0.0004	0.0031	0.0036
2	Design B	$\hat{\beta}_1$	0.010	0.006	0.007	0.007	0.006	-0.001
		SE	0.031	0.031	0.033	0.033	0.031	0.035
		MSE	0.0010	0.0010	0.0012	0.0012	0.0010	0.0012
2	Design C	$\hat{\beta}_1$	-0.001	-0.007	-0.012	-0.012	-0.007	-0.012
		SE	0.029	0.030	0.031	0.031	0.030	0.032
		MSE	0.0008	0.0009	0.0011	0.0011	0.0009	0.0012

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 3.9: Simulation results when Z and X are dependent and $\beta_1 = 0.5$

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	0.503	0.502	0.502	0.504	1.414	1.428
		SE	0.019	0.020	0.020	0.020	0.053	0.056
		MSE	0.0004	0.0004	0.0004	0.0004	0.8386	0.8642
2	Design B	$\hat{\beta}_1$	0.528	0.530	0.540	0.540	0.529	0.535
		SE	0.030	0.030	0.033	0.033	0.030	0.035
		MSE	0.0017	0.0018	0.0027	0.0027	0.0018	0.0025
2	Design C	$\hat{\beta}_1$	0.484	0.482	0.479	0.479	0.482	0.488
		SE	0.028	0.029	0.031	0.031	0.029	0.032
		MSE	0.0010	0.0012	0.0014	0.0014	0.0012	0.0012

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 3.10: Simulation results when Z and X are dependent and $\beta_1 = 1$

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	0.970	0.972	0.970	0.972	2.584	2.611
		SE	0.020	0.021	0.022	0.023	0.049	0.048
		MSE	0.0013	0.0012	0.0014	0.0013	2.5114	2.5980
2	Design B	$\hat{\beta}_1$	0.963	0.964	0.965	0.965	0.964	0.949
		SE	0.029	0.030	0.033	0.033	0.029	0.035
		MSE	0.0022	0.0022	0.0024	0.0024	0.0022	0.0038
2	Design C	$\hat{\beta}_1$	0.973	0.972	0.974	0.974	0.973	0.988
		SE	0.027	0.028	0.032	0.032	0.028	0.032
		MSE	0.0014	0.0016	0.0017	0.0017	0.0015	0.0012

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

3.7 Simulation Results Under Misspecification of the Distributional Assumption

We assumed that the distribution of the error term in the regression model (3.3) is Normal distribution and the true model and the assumed model were the same in Section 3.5. In practice, the conditional distribution of Y given $X = x$ and $Z = z$ may not be known. Therefore, in this section, we investigate the sensitivity of the sampling designs and estimation methods when the distribution of the error term is misspecified. In the simulation study, we generate the error term from Student's t-distribution with degrees of freedom 8 but the other simulation settings remain the same.

In Table 3.11, 3.12, and 3.13, we investigate the properties of parameter estimates of the coefficient β_1 of the expensive covariate when Z and X are independent, $\beta_1 = 0$, 0.5, 1, respectively, and the error distribution is misspecified. The simulation results show that the model misspecification affects the performance of the full likelihood and the estimated pseudo-likelihood estimation methods. They give biased estimates except the case when there is no association between X and Y. On the other hand,

the bias in the conditional likelihood method's and the IPW estimation method's estimator is much less under many designs. In addition, when there is an association between X and Y, the conditional likelihood method and the IPW method give more efficient estimates under many designs. Thus, they seem to be more robust compared to the full likelihood and the estimated pseudo-likelihood methods.

In Table 3.14, 3.15, and 3.16, we investigate the properties of parameter estimates of the coefficient β_1 of the expensive covariate when Z and X are dependent, $\beta_1 = 0$, 0.5, 1, respectively, the error distribution is misspecified. We observe similar results to those obtained when Z and X are independent.

Table 3.11: Simulation results when Z and X are independent, $\beta_1 = 0$ and the distribution of error term is misspecified

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	0.007	0.007	0.008	0.008	0.019	0.025
		SE	0.014	0.014	0.016	0.017	0.040	0.042
		MSE	0.0002	0.0002	0.0003	0.0004	0.0020	0.0024
2	Design B	$\hat{\beta}_1$	-0.030	-0.027	-0.029	-0.029	-0.027	-0.015
		SE	0.024	0.022	0.023	0.023	0.022	0.026
		MSE	0.0014	0.0012	0.0014	0.0014	0.0012	0.0009
2	Design C	$\hat{\beta}_1$	-0.008	-0.007	-0.008	-0.008	-0.008	-0.008
		SE	0.023	0.022	0.024	0.024	0.022	0.024
		MSE	0.0006	0.0005	0.0006	0.0006	0.0005	0.0006

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 3.12: Simulation results when Z and X are independent, $\beta_1 = 0.5$ and the distribution of error term is misspecified

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	0.456	0.459	0.531	0.564	1.293	1.340
		SE	0.014	0.014	0.017	0.018	0.039	0.040
		MSE	0.0022	0.0019	0.0013	0.0044	0.6301	0.7075
2	Design B	$\hat{\beta}_1$	0.477	0.463	0.493	0.493	0.463	0.511
		SE	0.022	0.021	0.023	0.023	0.021	0.026
		MSE	0.0010	0.0018	0.0006	0.0006	0.0018	0.0008
2	Design C	$\hat{\beta}_1$	0.470	0.456	0.491	0.491	0.455	0.492
		SE	0.022	0.021	0.024	0.024	0.021	0.024
		MSE	0.0013	0.0024	0.0006	0.0006	0.0024	0.0006

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 3.13: Simulation results when Z and X are independent, $\beta_1 = 1$ and the distribution of error term is misspecified

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	0.866	0.875	0.976	1.024	2.339	2.415
		SE	0.015	0.015	0.018	0.020	0.036	0.036
		MSE	0.0181	0.0159	0.0009	0.0010	1.7944	2.0030
2	Design B	$\hat{\beta}_1$	0.933	0.957	0.998	0.998	0.957	1.020
		SE	0.020	0.020	0.023	0.023	0.020	0.026
		MSE	0.0049	0.0023	0.0005	0.0005	0.0023	0.0011
2	Design C	$\hat{\beta}_1$	0.918	0.941	0.996	0.996	0.940	0.992
		SE	0.020	0.020	0.024	0.024	0.020	0.024
		MSE	0.0072	0.0039	0.0006	0.0006	0.0040	0.0006

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 3.14: Simulation results when Z and X are dependent, $\beta_1 = 0$ and the distribution of error term is misspecified

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	0.001	-0.002	-0.002	-0.003	-0.009	-0.011
		SE	0.015	0.015	0.018	0.019	0.046	0.050
		MSE	0.0002	0.0002	0.0003	0.0004	0.0022	0.0026
2	Design B	$\hat{\beta}_1$	0.012	0.008	0.010	0.010	0.008	0.004
		SE	0.025	0.024	0.027	0.027	0.024	0.028
		MSE	0.0008	0.0006	0.0008	0.0008	0.0006	0.0008
2	Design C	$\hat{\beta}_1$	0.000	-0.004	-0.009	-0.009	-0.004	-0.008
		SE	0.023	0.022	0.026	0.026	0.022	0.026
		MSE	0.0005	0.0005	0.0007	0.0007	0.0005	0.0007

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 3.15: Simulation results when Z and X are dependent, $\beta_1 = 0.5$ and the distribution of error term is misspecified

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	0.429	0.429	0.498	0.531	1.276	1.287
		SE	0.015	0.015	0.019	0.020	0.044	0.046
		MSE	0.0053	0.0053	0.0003	0.0014	0.6039	0.6222
2	Design B	$\hat{\beta}_1$	0.497	0.476	0.519	0.519	0.477	0.511
		SE	0.024	0.023	0.027	0.027	0.023	0.029
		MSE	0.0006	0.0011	0.0011	0.0011	0.0011	0.0009
2	Design C	$\hat{\beta}_1$	0.473	0.453	0.485	0.485	0.454	0.492
		SE	0.022	0.022	0.026	0.026	0.022	0.026
		MSE	0.0012	0.0027	0.0009	0.0009	0.0026	0.0007

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Table 3.16: Simulation results when Z and X are dependent, $\beta_1 = 1$ and the distribution of error term is misspecified

#	Sampling Type		Method					
			β_{1R}	β_{1F}	β_{1C0}	β_{1C1}	β_{1P}	β_{1W}
1	Design A	$\hat{\beta}_1$	0.862	0.866	0.971	1.024	2.391	2.415
		SE	0.016	0.017	0.020	0.022	0.040	0.040
		MSE	0.0192	0.0182	0.0012	0.0010	1.9358	2.0040
2	Design B	$\hat{\beta}_1$	0.958	0.964	1.024	1.024	0.965	1.021
		SE	0.021	0.022	0.027	0.027	0.022	0.028
		MSE	0.0023	0.0018	0.0013	0.0013	0.0017	0.0012
2	Design C	$\hat{\beta}_1$	0.933	0.926	0.975	0.975	0.928	0.992
		SE	0.021	0.021	0.026	0.026	0.021	0.026
		MSE	0.0050	0.0059	0.0013	0.0013	0.0057	0.0007

SE and MSE denote the standard error estimate and mean square error estimate of the estimator of β_1, respectively.

Chapter 4

Efficiency of Multiple Response-Dependent Two-Phase Sampling Designs

In a multiple response-dependent two-phase sampling design, in phase I, we have easily measured variables including the response variables for all individuals in a cohort or in a large random sample from the population, and in phase II, we obtain expensive covariate(s) for a subset of individuals selected according to their multiple response variables obtained in phase I. An introduction to multiple response-dependent two-phase sampling design and the aim of our study under this design setting are given in Section 1.5.

Recent genetic association studies collect data on a variety of quantitative traits. For example, the NHLBI ESP (Lin et al., 2013) contains several studies, each of which was focused on a particular trait. The NHLBI ESP was designed to identify genetic variants in all protein-coding regions of the human genome that are associated with heart, lung, and blood diseases. Samples were selected based on multiple quantitative traits from seven studies and whole exome sequencing was performed on 4494 subjects which were selected based on multiple quantitative traits values. The NHLBI ESP project is an example for the need of efficient multiple response-dependent two-phase sampling designs.

4.1 Multiple Response-Dependent Two-Phase Sampling Design Setting

Suppose we have a population with units having multiple responses and an expensive covariate to measure. We denote $\boldsymbol{Y}$ as a vector of response variables and X as an expensive covariate. Let y_{ji} be the observed value of the j^{th} response for the i^{th} unit, where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, J$ $(J \geq 2)$, X_i be an expensive covariate for the i^{th} unit, $\boldsymbol{Y_i}$ is a J-dimensional response vector $(J \geq 2)$ for the i^{th} unit.

Now, suppose that $\boldsymbol{Y}$ is observed for all N units in phase I. However, the expensive covariate X can only be observed for a subsample of individuals selected in phase II for a fixed sample size n. Let R_{ij} be an indicator function where $R_{ij} = 1$ if unit i in phase I cohort is selected for inclusion in phase II sample based on the j^{th} response variable, and $R_{ij} = 0$ otherwise. Thus, the observed data consists of N units where $n = \sum_{i=1}^{N} \sum_{j=1}^{J} R_{ij}$ of the units provide complete data $(\boldsymbol{y_i}, x_i)$, and $(N - n)$ of units provide information only on response values $\boldsymbol{y_i}$. The set of completely observed data, V, and the set of units with incomplete data, $\bar{V}$, are denoted by

$$\begin{aligned} V &= \{i : R_{ij} = 1, i = 1, 2, \ldots, N, j = 1, 2, \ldots, J\} \text{ and} \\ \bar{V} &= \{i : R_{ij} = 0, i = 1, 2, \ldots, N, j = 1, 2, \ldots, J\}, \end{aligned} \tag{4.1}$$

respectively.

Let π_{ij} denote the probability of selecting the i^{th} unit in phase II sample depending on the j^{th} response variable. We assume that the covariate X is "missing at random" in the terminology of Little and Rubin (1987). Thus, the π_{ij} is

$$\pi_{ij} = P(R_{ij} = 1 | \boldsymbol{y_i}, x_i) = P(R_{ij} = 1 | \boldsymbol{y_i}) \tag{4.2}$$

for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, J$.

4.2 Multiple Response-Dependent BSS Design

Suppose the observed values of the j^{th} response variable $(y_{j1}, y_{j2}, \ldots, y_{jN})$, $j = 1, 2, \ldots, J$ is partitioned into K number of strata, S_{jk}, $k = 1, 2, \ldots, K$, using fixed

cut-point values $C_{jl},\ l = 1, 2, \ldots, K-1$, where $C_{j1} \leq C_{j2} \leq \ldots \leq C_{jK-1}$. The first stratum includes the units with y_{ji} values less than C_{j1}, the k^{th} $(k = 2, \ldots, K-1)$ stratum includes the units with y_{ji} values between C_{jk-1} and C_{jk}, and the K^{th} stratum includes the units with y_{ji} values greater than C_{jk-1}. Let N_{jk} $(k = 1, 2, \ldots, K)$ be the number of units in each stratum where $\sum_{k=1}^{K} N_{jk} = N$. Then, from each stratum S_{jk}, n_{jk} units are randomly selected for inclusion in the phase II sample, where the number of units selected depending on the j^{th} response variable is $n_j = \sum_{k=1}^{K} n_{jk}$ and the total phase II sample size $n = \sum_{j=1}^{J} n_j$ is fixed according to the budgetary constraint.

The selection probability for the i^{th} unit depending on the j^{th} response variable in (4.2) becomes $\pi_{ij} = \sum_{k=1}^{K} \delta_{ijk} p_{jk}$ for $i = 1, 2, \ldots, N$, where

$$\delta_{ijk} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ unit is selected in phase II based on the } j^{\text{th}} \text{ response from} \\ & \text{the } k^{\text{th}} \text{ stratum,} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$p_{jk} = \frac{n_{jk}}{N_{jk}}, \quad j = 1, 2, \ldots, J, \quad k = 1, 2, \ldots, K.$$

4.3 Phase II Sampling Design

Suppose there is a bivariate response vector for each unit and the observed values of j^{th} response variable $\{y_{j1}, y_{j2}, \ldots, y_{jN}\}$ $(j = 1, 2)$ are available for a cohort of size N in phase I. Suppose n_j units are selected in phase II depending on the j^{th} response variable to obtain their expensive covariate data. To select these units, response variable observations in the j^{th} response variable are partitioned into $K = 3$ number

of strata based on the fixed constants C_{j1} and C_{j2} as follows:

$$\underbrace{y_{j1} \leq \ldots \leq y_{jN_{j1}}}_{\substack{\text{Low stratum} \\ (S_{j1})}} \leq C_{j1} \leq \underbrace{y_{jN_{j1}+1} \leq \ldots \leq y_{jN_{j1}+N_{j2}}}_{\substack{\text{Middle stratum} \\ (S_{j2})}} \leq C_{j2} \leq \underbrace{y_{jN_{j1}+N_{j2}+1} \leq \ldots \leq y_{jN}}_{\substack{\text{High stratum} \\ (S_{j3})}} \tag{4.3}$$

Suppose there are N_{j1} units in the first stratum S_{j1} which includes units with j^{th} response values lower than C_{j1}, N_{j2} units in the second stratum S_{j2} which includes units with j^{th} response values between C_{j1} and C_{j2}, and $N_{j3} = N - N_{j1} - N_{j2}$ units in the third stratum S_{j3} which includes units with j^{th} response values greater than C_{j2}. The first and the third strata consist of extreme values and sizes of the extreme strata are set small compared to the central stratum to understand the importance of sampling from the extreme strata. We create three strata based on each response variable. Then, multiple response-dependent BSS design described in Section 4.2 is applied to select the phase II sample. That is, from each stratum S_{jk}, n_{jk} $(j = 1, 2,\ k = 1, 2, 3)$ units are randomly selected for inclusion in the phase II sample, where the total phase II sample size $n = \sum_{j=1}^{2} \sum_{k=1}^{3} n_{jk}$ is fixed according to the budgetary constraint. The chance of unit being selected in the phase II sample now depends on two response variables, therefore the overall inclusion indicator R_i for the i^{th} $(i = 1, 2, \ldots, N)$ unit is:

$$R_i = R_{i1}(1 - R_{i2}) + R_{i2}(1 - R_{i1}) + R_{i1}R_{i2}, \tag{4.4}$$

and the selection probability for the i^{th} unit becomes

$$\pi_i = \pi_{i1}(1 - \pi_{i2}) + \pi_{i2}(1 - \pi_{i1}) + \pi_{i1}\pi_{i2}. \tag{4.5}$$

In the sampling design based on a univariate response variable in Chapter 2, we obtained the most efficient full likelihood estimate under the extreme sampling design (Design I), the IPW method yields the most efficient estimate under the design setting with sampling mostly from the central stratum (Design IV). Therefore, under the bivariate response variable setting, we consider these two sampling designs for one response variable and for the other response variable, we investigate the five sampling designs presented in Table 2.1. We assume the available budget to measure the expensive covariate is only for n individuals and equal number of units ($n_1 =$

$n_2 = \frac{n}{2}$) are selected based on each response variable. In this study, we investigate two sampling schemes (A and B) to select sampling units from each stratum within the framework of phase II design. In scheme (A) sampling designs, we consider sampling half of the units from extreme strata based on Y_1 with $n_{11} = n_{13} = \frac{n}{4}$ and $n_{12} = 0$ and investigate sampling designs I-V to select samples based on Y_2. In scheme (B) sampling designs, we consider sampling 12.5% units from each of extreme strata and 25% units from the central stratum based on Y_1 ($n_{11} = n_{13} = \frac{n}{8}$ and $n_{12} = \frac{n}{4}$) and investigate sampling designs I-V to select samples based on Y_2. We investigate these sampling designs when the dependence between the response variables Y_1 and Y_2 is moderate (with Kendall's $\tau = 0.4$) and strong ($\tau = 0.8$).

Under current sampling design setting, the probability π_i of selecting the i^{th} unit in phase II sample defined in (4.5) becomes

$$\pi_i = \sum_{j=1}^{2}\sum_{k=1}^{3}\frac{\delta_{ijk}n_{jk}}{N_{jk}} - \prod_{j=1}^{2}\sum_{k=1}^{3}\frac{\delta_{ijk}n_{jk}}{N_{jk}}, \tag{4.6}$$

which is predetermined by the sampling plan.

Also, we applied sampling designs based on single response variable as described in Chapter 2 for Y_1 and Y_2 separately.

4.4 Model Description

Suppose there are two response variables Y_1 and Y_2 with conditional distribution functions $F_1(y_1|x)$ and $F_2(y_2|x)$, respectively, and conditional joint distribution function $F(y_1, y_2|x)$ which is modeled by the Clayton copula function C_ϕ as below

$$\begin{aligned} F(y_{1i}, y_{2i}|x_i) &= C_\phi(F_1(y_{1i}|x_i), F_2(y_{2i}|x_i)) \\ &= (F_1(y_{1i}|x_i)^{-\phi} + F_2(y_{2i}|x_i)^{-\phi} - 1)^{-\frac{1}{\phi}},\ \phi > 0, \end{aligned} \tag{4.7}$$

where the parameter ϕ specifies the dependence level between Y_1 and Y_2 after adjusted for $X = x$ and the marginal models of Y_1 and Y_2 given $X = x$ are

$$Y_{ji} = \beta_{0j} + \beta_{1j}x_i + \epsilon_{ji} \quad (j = 1, 2;\ i = 1, 2, \ldots, N), \tag{4.8}$$

where β_{0j}, β_{1j} $(j = 1, 2)$ are the regression coefficients and ϵ_{ji} is a random error for subject i $(i = 1, 2, \ldots, N)$.

In our study, we assume X is a Bernoulli random variable with probability of having $X = 1$ being p, ϵ_{ji}'s are independently and identically normally distributed with mean 0 and variance σ_j^2. Therefore, Y_{ji} has a Normal distribution with mean $\beta_{0j} + \beta_{1j}x_i$ and variance σ_j^2 when $X_i = x_i$ is observed.

We have seven unknown parameters, β_{01}, β_{11}, β_{02}, β_{12}, σ_1, σ_2 and ϕ. To estimate $\boldsymbol{\theta} = (\beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \sigma_1, \sigma_2, \phi)$, we obtain the samples by applying the proposed sampling designs presented in Section 4.3. We consider two estimation methods to estimate the parameter vector $\boldsymbol{\theta}$. Under each estimation method, we aim to check whether the estimators of β_{1j}'s are unbiased and to investigate the efficiency of the estimators under each sampling design setting when the dependence between the response variables Y_1 and Y_2 is moderate ($\tau = 0.4$) and strong ($\tau = 0.8$).

Under the model in (4.7) and (4.8) with the specified assumptions,

$$\begin{aligned} f(y_{1i}, y_{2i}|x_i;\theta) &= \frac{\partial C_\phi(F_1(y_{1i}|x_i;\boldsymbol{\theta_1}), F_2(y_{2i}|x_i;\boldsymbol{\theta_2}))}{\partial y_{1i}\partial y_{2i}} \\ &= \frac{\partial C_\phi(F_1(y_{1i}|x_i;\boldsymbol{\theta_1}), F_2(y_{2i}|x_i;\boldsymbol{\theta_2}))}{\partial F_1(y_{1i}|x_i;\boldsymbol{\theta_1})\partial F_2(y_{2i}|x_i;\boldsymbol{\theta_2})} \frac{\partial F_1(y_{1i}|x_i;\boldsymbol{\theta_1})}{\partial y_{1i}} \frac{\partial F_2(y_{2i}|x_i;\boldsymbol{\theta_2})}{\partial y_{1i}} \\ &= (1+\phi) f_1(y_{1i}|x_i;\boldsymbol{\theta_1}) f_2(y_{2i}|x_i;\boldsymbol{\theta_2}) \left[F_1(y_{1i}|x_i;\boldsymbol{\theta_1}) F_2(y_{2i}|x_i;\boldsymbol{\theta_2})\right]^{-(1+\phi)} \\ &\quad \left[F_1(y_{1i}|x_i;\boldsymbol{\theta_1})^{-\phi} + F_2(y_{2i}|x_i;\boldsymbol{\theta_2})^{-\phi} - 1\right]^{-(2+\frac{1}{\phi})}, \end{aligned}$$

where $\boldsymbol{\theta} = (\beta_{01}, \beta_{02}, \beta_{11}, \beta_{12}, \sigma_1, \sigma_2, \phi)$ and $\boldsymbol{\theta_j} = (\beta_{0j}, \beta_{1j}, \sigma_j)$ for $j = 1, 2$,

$$f_j(y_{ji}|x_i;\boldsymbol{\theta_j}) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(y_{ji}-\beta_{0j}-\beta_{1j}x_i)^2}{2\sigma_j^2}}, \tag{4.9}$$

and

$$g(x_i;p) = \begin{cases} p & \text{for } x_i = 1 \\ 1-p & \text{for } x_i = 0 \end{cases}$$

are the conditional joint density function of Y_1 and Y_2 given $X_i = x_i$, the conditional density functions of Y_{ji} given $X_i = x_i$ and the marginal distribution of X_i, respectively.

4.5 Estimation Methods

The logarithm of the full likelihood $L_R(\boldsymbol{\theta}, p)$ which incorporates both complete and incomplete data $\{(\boldsymbol{y_i}, x_i) : i \in V\} \cup \{\boldsymbol{y_i} : i \in \bar{V}\}$ under the missing at random assumption becomes

$$\begin{aligned} l_R(\boldsymbol{\theta}, p) = &\sum_{i \in V} [\log f(y_{1i}, y_{2i}|x_i; \boldsymbol{\theta}) + \log g(x_i; p)] \\ &+ \sum_{i \in \bar{V}} \log[pf(y_{1i}, y_{2i}|x = 1; \boldsymbol{\theta}) + (1-p)f(y_{1i}, y_{2i}|x = 0; \boldsymbol{\theta})]. \end{aligned} \tag{4.10}$$

The maximum likelihood estimates of $\boldsymbol{\theta}$ and p are obtained by maximizing $l_R(\boldsymbol{\theta}, p)$ in (4.10) with respect to $\boldsymbol{\theta}$ and p. We denote the resulting estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_R$.

The IPW method estimating function becomes

$$\begin{aligned} S_w(\boldsymbol{\theta}) = \Bigg(&\sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{1i}(\boldsymbol{\theta}),\ \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{2i}(\boldsymbol{\theta}),\ \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{3i}(\boldsymbol{\theta}),\ \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{4i}(\boldsymbol{\theta}), \\ &\sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{5i}(\boldsymbol{\theta}),\ \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{6i}(\boldsymbol{\theta}),\ \sum_{i=1}^{N} \frac{R_i}{\pi_i} U_{7i}(\boldsymbol{\theta}) \Bigg)^T \end{aligned} \tag{4.11}$$

where

$$\begin{aligned} U_{1i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_{1i}, y_{2i}|x_i; \boldsymbol{\theta})}{\partial \beta_{01}}, \\ U_{2i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_{1i}, y_{2i}|x_i; \boldsymbol{\theta})}{\partial \beta_{02}}, \\ U_{3i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_{1i}, y_{2i}|x_i; \boldsymbol{\theta})}{\partial \beta_{11}}, \\ U_{4i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_{1i}, y_{2i}|x_i; \boldsymbol{\theta})}{\partial \beta_{12}}, \\ U_{5i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_{1i}, y_{2i}|x_i; \boldsymbol{\theta})}{\partial \sigma_1}, \\ U_{6i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_{1i}, y_{2i}|x_i; \boldsymbol{\theta})}{\partial \sigma_2}, \\ U_{7i}(\boldsymbol{\theta}) &= \frac{\partial \log f(y_{1i}, y_{2i}|x_i; \boldsymbol{\theta})}{\partial \phi}. \end{aligned}$$

By solving the estimating equations $S_w(\boldsymbol{\theta}) = 0$, we obtain the IPW estimators $\hat{\boldsymbol{\theta}}_w = (\hat{\beta}_{01,w}, \hat{\beta}_{02,w}, \hat{\beta}_{11,w}, \hat{\beta}_{12,w}, \hat{\sigma}_{1,w}, \hat{\sigma}_{2,w}, \hat{\phi}_w)$.

Under some regularity conditions, the covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta})$ given in (1.17) becomes

$$C(\boldsymbol{\theta}) = A(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}) [A(\boldsymbol{\theta})^{-1}]' \tag{4.12}$$

where

$$A(\boldsymbol{\theta}) = -\frac{1}{n}\left(\frac{\partial S_w(\theta)}{\partial \boldsymbol{\theta}'}\right) = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{17} \\ A_{21} & A_{22} & \dots & A_{27} \\ \vdots & \vdots & \vdots & \vdots \\ A_{71} & A_{72} & \dots & A_{77} \end{pmatrix},$$

and

$$B(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{N} S_{wi}(\boldsymbol{\theta}) S_{wi}(\boldsymbol{\theta})' = \begin{pmatrix} B_{11} & B_{12} & \dots & B_{17} \\ B_{21} & B_{22} & \dots & B_{27} \\ \vdots & \vdots & \vdots & \vdots \\ B_{71} & B_{72} & \dots & B_{77} \end{pmatrix},$$

The covariance matrix of $\hat{\boldsymbol{\theta}}_w$ is $\frac{1}{n}C(\boldsymbol{\theta})$. A consistent estimator of $C(\boldsymbol{\theta})$ is obtained by plugging the estimates of $\boldsymbol{\theta}$ in (4.12) and the estimated covariance matrix of the estimators is obtained as

$$\frac{1}{n}C(\hat{\boldsymbol{\theta}}_w).$$

4.6 Simulation Study

A simulation study was conducted to investigate the properties of the estimation methods under the sampling design settings described in Section 4.3. We aim to identify the sampling design which gives the most efficient estimates of coefficients of expensive covariate under each estimation method and to compare the performance of the estimation methods. We assume that there is a cohort of size $N = 50,000$ with their observed response variable Y_j $(j = 1, 2)$ values in the first phase. We are interested in making inference on the association between the response variables Y_j's $(j = 1, 2)$ and an expensive covariate X. We consider linear regression models of Y_j's

on X with normally distributed error terms.

In data generation, we first generated uniform random variable U_1 and U_2 of size $N = 50,000$ from (1.19) and generated normally distributed error terms, ϵ_{1i} and ϵ_{2i} by using the inverse cumulative distribution function technique with two different dependence levels represented by Kendall's tau values of $\tau = 0.4$ or $\tau = 0.8$. We considered two different values of τ to investigate the effect of changing the dependence level between response variables. The marginal distributions of ϵ_{1i} and ϵ_{2i} are Normal distribution with mean 0 and variances ${\sigma_1}^2 = 2$, ${\sigma_2}^2 = 4$, respectively and we generated the covariate values x_i from the Bernoulli distribution with $p = P(X_i = 1) = 0.4$. Next, we generated the response variable values $y'_{ji}s$ using the model (4.8), where we choose values of $\beta_{11} = 1$ and $\beta_{12} = 2$ with the intercepts $\beta_{01} = 10$ and $\beta_{02} = 5$. The response variable Y_1 has the same distribution with Y in Section 2.2.

In our study, we set the cut-point values (C_{j1}, C_{j2}) in (4.3) to the (10^{th}, 90^{th}) percentiles of the response distribution which can divide the cohort into three strata with approximate sizes of $N_{j1} \approx 5,000, N_{j2} \approx 40,000, N_{j3} \approx 5,000$, for each response variable. The cut-point values C_{j1} and C_{j2} under each response variable are given in Table 4.1.

First, we applied the sampling designs based only on the single response variable discussed in Chapter 2 for Y_1 and Y_2 separately under two different dependence levels between Y_1 and Y_2. Then, we applied the multiple response-dependent BSS design described in Section 4.2 to obtain the phase II samples and for only these selected samples, we assume to have the expensive covariate X values. In each sampling design, the phase II sample size is set to $n = 10,000$. In scheme (A) sampling designs, we consider sampling 2500 units from each extreme strata and in scheme (B) sampling designs, we consider sampling 1250 units from each of extreme strata and central 2500 of samples are selected based on Y_1. The stratum specific sample sizes based on Y_2 are given in Table 4.2. Estimated values of β_{11} and β_{12} are obtained using the two estimation methods described in Section 4.5 under each sampling design.

Table 4.1: The cut-point values C_{m1} and C_{m2}

τ	Y_1		Y_2	
	C_{11}	C_{12}	C_{21}	C_{22}
0.4	8.4757	12.3205	2.9513	8.6548
0.8	8.4757	12.3205	2.9500	8.6415

Table 4.2: Stratum-specific sample sizes based on response variable Y_2

Sampling scenario	**Stratum-specific sample sizes**			**Sampling design**
	Low stratum n_{21} $(\frac{n_{21}}{n} \times 100\%)$	**Middle stratum** n_{22} $(\frac{n_{22}}{n} \times 100\%)$	**High stratum** n_{23} $(\frac{n_{23}}{n} \times 100\%)$	
1	2500 (25%)	0 (0%)	2500 (25%)	(I) Sampling from extreme strata only
2	2000 (20%)	1000 (10%)	2000 (20%)	(II) Oversampling from extreme strata
3	2500 (25%)	1000 (10%)	1500 (15%)	
4	1500 (15%)	1000 (10%)	2500 (25%)	
5	2500 (25%)	2000 (20%)	500 (5%)	(III) Oversampling from only one extreme stratum
6	500 (5%)	2000 (20%)	2500 (25%)	
7	1250 (12.5%)	2500 (25%)	1250 (12.5%)	(IV) Oversampling from middle stratum
8	500 (5%)	3000 (30%)	1500 (15%)	
9	1500 (15%)	3000 (30%)	500 (5%)	
10	500 (5%)	4000 (40%)	500 (5%)	(V) Simple random sampling

4.7 Simulation Results

The point estimates of the coefficients β_{1j} $(j = 1, 2)$ and dependence parameter ϕ and their standard error estimates and mean square error estimates obtained through the two estimation methods are given from Table 4.3 to Table 4.11 under different design settings. Table 4.3 shows the results under the maximum likelihood estimation using the full cohort data. As expected, we obtained consistent maximum likelihood estimates.

Table 4.4 and 4.5 show properties of parameter estimators of the β_{11}, β_{12}, and ϕ when there is moderate and strong dependence between Y_1 and Y_2 ($\tau = 0.4, 0.8$, respectively) and sampling depends only on Y_1. As expected, we observe similar results to those obtained in Chapter 2 under single response-dependent designs for β_{11}. Also, we obtained similar results for β_{12} since there is a moderate or strong dependence between Y_1 and Y_2. Full likelihood-based method provides the most efficient estimate when sampling all of the units from the extreme strata or oversampling from the

extreme strata. We observe that the full likelihood-based method gives consistent estimates and more efficient estimates for ϕ under each sampling design considered. The sampling designs selecting more units from the central stratum give consistent and more efficient estimates for β_{11}, β_{12}, and ϕ under the IPW estimation method. When there is a strong dependence ($\tau = 0.8$), we observe an efficiency gain for both of the estimates under each sampling design considered. However, the IPW estimates of ϕ became less efficient compared to the $\tau = 0.4$ setting.

Table 4.6 and 4.7 show properties of parameter estimators of β_{11}, β_{12}, and ϕ when there is a moderate and strong dependence between Y_1 and Y_2, respectively, and sampling depends only on Y_2. We observe similar results to those obtained sampling depends only on Y_1 for all the parameters.

Table 4.8 and 4.9 show properties of parameter estimators when sampling extreme Y_1 strata units (Scheme A) and different sampling designs were considered based on Y_2. We observe that the full likelihood method gives consistent estimates and more efficient estimates than the IPW estimation method for all three parameters. In addition, under the extreme sampling design based on Y_1, IPW method yields biased estimates for each sampling design considered. When sampling units selected only from extreme strata based on both Y_1 and Y_2, the full likelihood method provides the most efficient estimates for β_{11} and β_{12}. When there is strong dependence between the two response variables ($\tau = 0.8$), we observe an efficiency gain for the full likelihood-based method estimates under each sampling design considered.

Table 4.10 and 4.11 show properties of parameter estimators when oversampling from the middle stratum with equal sampling probabilities from extreme strata based on Y_1 (Scheme B) and different sampling designs were considered based on Y_2. We observe that the full likelihood method gives consistent estimates and more efficient estimates than the IPW estimation method. On the other hand, the IPW method yields biased estimates under extreme sampling design or oversampling from at least one of the extreme stratum based on Y_2. We observe that the IPW estimation method may give biased estimates under some sampling designs. Thus, we compare the MSEs to determine efficient designs. In oversampling from middle stratum based on Y_2 or simple random sampling based on Y_2, the IPW method provides the most efficient estimates for each parameter. When there is a strong association ($\tau = 0.8$), we observe efficiency gain under each sampling design considered for the full likelihood-based and

IPW estimation methods. However, the IPW estimates of ϕ became less efficient compared to the $\tau = 0.4$ setting.

Table 4.3: Maximum likelihood estimation results under the full cohort

Parameter value		Estimates under the dependence level	
		$\tau = 0.4$ ($\phi = 1.33$)	$\tau = 0.8$ ($\phi = 8$)
$\beta_{11} = 1$	$\hat{\beta}_{11}$	1.007	1.011
	SE	0.012	0.010
	MSE	0.0002	0.0002
$\beta_{12} = 2$	$\hat{\beta}_{12}$	2.006	2.015
	SE	0.016	0.014
	MSE	0.0003	0.0004
ϕ	$\hat{\phi}$	1.318	8.014
	SE	0.009	0.007
	MSE	0.0003	0.0003

SE and MSE denote the standard error estimate and mean square error estimate of the corresponding estimator.

Table 4.4: Simulation results when $\tau = 0.4$ and sampling based on Y_1 only

#	Sampling type	Sampling percentage (S_{11},S_{12},S_{13})		$\beta_{11} = 1$		$\beta_{12} = 2$		$\phi = 1.33$	
				FLM	IPW	FLM	IPW	FLM	IPW
1	Sampling from extreme strata only	(50%, 0%, 50%)	PE	1.006	2.657	1.995	3.366	1.322	2.397
			SE	0.015	0.038	0.025	0.043	0.010	0.040
			MSE	0.0002	2.7466	0.0007	1.8666	0.0002	1.1333
2	Oversampling from extreme strata	(40%, 20%, 40%)	PE	1.009	1.056	2.013	2.064	1.323	1.248
			SE	0.016	0.035	0.025	0.057	0.010	0.038
			MSE	0.0003	0.0043	0.0008	0.0074	0.0002	0.0087
3		(50%, 20%, 30%)	PE	1.004	1.020	1.988	1.985	1.322	1.318
			SE	0.015	0.036	0.026	0.061	0.010	0.041
			MSE	0.0002	0.0017	0.0008	0.0039	0.0002	0.0019
4		(30%, 20%, 50%)	PE	1.008	1.020	2.014	2.029	1.320	1.334
			SE	0.016	0.035	0.025	0.059	0.010	0.041
			MSE	0.0003	0.0016	0.0008	0.0043	0.0003	0.0017
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	PE	1.008	1.030	1.986	1.995	1.320	1.321
			SE	0.017	0.029	0.026	0.046	0.010	0.029
			MSE	0.0003	0.0017	0.0009	0.0021	0.0003	0.0010
6		(10%, 40%, 50%)	PE	1.010	1.009	2.022	2.013	1.317	1.307
			SE	0.019	0.028	0.026	0.043	0.011	0.034
			MSE	0.0004	0.0009	0.0011	0.0020	0.0004	0.0018
7	Oversampling from middle stratum	(25%, 50%, 25%)	PE	1.007	1.016	2.003	2.010	1.321	1.357
			SE	0.017	0.025	0.026	0.039	0.010	0.028
			MSE	0.0003	0.0009	0.0007	0.0017	0.0003	0.0013
8		(10%, 60%, 30%)	PE	1.012	1.010	2.017	1.999	1.317	1.319
			SE	0.019	0.025	0.026	0.037	0.011	0.030
			MSE	0.0005	0.0007	0.0010	0.0014	0.0004	0.0011
9		(30%, 60%, 10%)	PE	1.012	1.036	2.021	2.053	1.320	1.351
			SE	0.018	0.026	0.026	0.039	0.011	0.027
			MSE	0.0004	0.0019	0.0011	0.0043	0.0003	0.0010
10	Simple random sampling	(10%, 80%, 10%)	PE	1.014	1.019	2.014	1.998	1.317	1.316
			SE	0.020	0.026	0.027	0.037	0.011	0.028
			MSE	0.0006	0.0011	0.0009	0.0013	0.0004	0.0011

PE denotes the point estimate of the corresponding parameter. FLM denotes the full likelihood method. SE and MSE denote the standard error estimate and mean square error estimate of the corresponding estimator.

Table 4.5: Simulation results when $\tau = 0.8$ and sampling based on Y_1 only

#	Sampling type	Sampling percentage (S_{11},S_{12},S_{13})		$\beta_{11} = 1$		$\beta_{12} = 2$		$\phi = 8$	
				FLM	IPW	FLM	IPW	FLM	IPW
1	Sampling from extreme strata only	(50%, 0%, 50%)	PE	1.014	2.080	2.021	3.346	8.064	12.175
			SE	0.012	0.030	0.017	0.040	0.007	0.192
			MSE	0.0004	1.1663	0.0007	1.8133	0.0042	17.4640
2	Oversampling from extreme strata	(40%, 20%, 40%)	PE	1.022	1.094	2.032	2.134	8.054	7.728
			SE	0.012	0.032	0.017	0.046	0.007	0.168
			MSE	0.0006	0.0099	0.0013	0.0202	0.0030	0.1021
3		(50%, 20%, 30%)	PE	1.014	0.998	2.021	1.988	8.050	7.922
			SE	0.013	0.034	0.018	0.049	0.007	0.179
			MSE	0.0004	0.0012	0.0008	0.0025	0.0025	0.0382
4		(30%, 20%, 50%)	PE	1.012	1.011	2.021	2.022	8.085	8.133
			SE	0.012	0.034	0.017	0.048	0.007	0.179
			MSE	0.0003	0.0013	0.0008	0.0028	0.0073	0.0497
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	PE	1.016	1.022	2.023	2.025	8.044	7.984
			SE	0.013	0.026	0.018	0.038	0.007	0.128
			MSE	0.0004	0.0012	0.0009	0.0020	0.0020	0.0166
6		(10%, 40%, 50%)	PE	1.012	1.000	2.021	1.998	8.078	8.056
			SE	0.013	0.025	0.018	0.036	0.007	0.141
			MSE	0.0003	0.0006	0.0007	0.0013	0.0061	0.0230
7	Oversampling from middle stratum	(25%, 50%, 25%)	PE	1.011	1.015	2.021	2.024	8.091	8.121
			SE	0.013	0.022	0.018	0.032	0.007	0.120
			MSE	0.0003	0.0007	0.0007	0.0016	0.0083	0.0289
8		(10%, 60%, 30%)	PE	1.016	1.008	2.025	2.007	8.068	8.045
			SE	0.013	0.022	0.018	0.030	0.007	0.122
			MSE	0.0004	0.0005	0.0010	0.0010	0.0046	0.0170
9		(30%, 60%, 10%)	PE	1.020	1.036	2.031	2.056	8.082	8.098
			SE	0.013	0.023	0.018	0.032	0.007	0.113
			MSE	0.0006	0.0018	0.0013	0.0041	0.0069	0.0225
10	Simple random sampling	(10%, 80%, 10%)	PE	1.022	1.020	2.033	2.023	8.058	8.054
			SE	0.013	0.022	0.019	0.031	0.007	0.115
			MSE	0.0007	0.0009	0.0014	0.0015	0.0034	0.0160

PE denotes the point estimate of the corresponding parameter. FLM denotes the full likelihood method. SE and MSE denote the standard error estimate and mean square error estimate of the corresponding estimator.

Table 4.6: Simulation results when $\tau = 0.4$ and sampling based on Y_2 only

#	Sampling type	Sampling percentage (S_{21},S_{22},S_{23})		$\beta_{11} = 1$		$\beta_{12} = 2$		$\phi = 1.33$	
				FLM	IPW	FLM	IPW	FLM	IPW
1	Sampling from extreme strata only	(50%, 0%, 50%)	PE	1.001	2.305	2.017	5.103	1.323	2.248
			SE	0.021	0.030	0.022	0.051	0.011	0.042
			MSE	0.0004	1.7035	0.0008	9.6289	0.0002	0.8387
2	Oversampling from extreme strata	(40%, 20%, 40%)	PE	0.999	0.963	2.027	2.042	1.321	1.325
			SE	0.021	0.043	0.023	0.053	0.011	0.041
			MSE	0.0004	0.0032	0.0012	0.0046	0.0003	0.0018
3		(50%, 20%, 30%)	PE	1.004	1.039	2.005	1.999	1.321	1.286
			SE	0.020	0.043	0.022	0.054	0.011	0.040
			MSE	0.0004	0.0033	0.0005	0.0029	0.0003	0.0038
4		(30%, 20%, 50%)	PE	1.003	0.976	2.016	1.939	1.321	1.343
			SE	0.021	0.043	0.023	0.052	0.011	0.041
			MSE	0.0004	0.0024	0.0008	0.0064	0.0003	0.0018
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	PE	1.010	1.004	1.997	1.955	1.320	1.328
			SE	0.020	0.032	0.024	0.042	0.011	0.030
			MSE	0.0005	0.0011	0.0006	0.0039	0.0003	0.0010
6		(10%, 40%, 50%)	PE	1.007	0.990	2.032	2.031	1.320	1.305
			SE	0.020	0.031	0.025	0.040	0.011	0.033
			MSE	0.0005	0.0010	0.0016	0.0026	0.0003	0.0019
7	Oversampling from middle stratum	(25%, 50%, 25%)	PE	1.003	0.982	2.013	1.992	1.318	1.307
			SE	0.020	0.028	0.024	0.036	0.011	0.028
			MSE	0.0004	0.0011	0.0008	0.0014	0.0004	0.0015
8		(10%, 60%, 30%)	PE	1.012	1.006	2.017	2.019	1.317	1.308
			SE	0.020	0.026	0.025	0.036	0.011	0.029
			MSE	0.0005	0.0007	0.0009	0.0016	0.0004	0.0015
9		(30%, 60%, 10%)	PE	1.014	1.002	2.001	1.971	1.319	1.334
			SE	0.020	0.028	0.025	0.037	0.011	0.026
			MSE	0.0006	0.0008	0.0006	0.0022	0.0003	0.0007
10	Simple random sampling	(10%, 80%, 10%)	PE	1.023	1.011	2.019	2.031	1.313	1.293
			SE	0.020	0.026	0.026	0.037	0.011	0.028
			MSE	0.0009	0.0008	0.0011	0.0023	0.0005	0.0024

PE denotes the point estimate of the corresponding parameter. FLM denotes the full likelihood method. SE and MSE denote the standard error estimate and mean square error estimate of the corresponding estimator.

Table 4.7: Simulation results when $\tau = 0.8$ and sampling based on Y_2 only

#	Sampling type	Sampling percentage (S_{21},S_{22},S_{23})		$\beta_{11} = 1$		$\beta_{12} = 2$		$\phi = 8$	
				FLM	IPW	FLM	IPW	FLM	IPW
1	Sampling from extreme strata only	(50%, 0%, 50%)	PE	1.011	2.265	2.018	3.986	8.082	11.254
			SE	0.012	0.028	0.017	0.041	0.008	0.182
			MSE	0.0003	1.5997	0.0006	3.9467	0.0068	10.6206
2	Oversampling from extreme strata	(40%, 20%, 40%)	PE	1.013	1.006	2.021	2.021	8.069	7.772
			SE	0.013	0.034	0.018	0.048	0.007	0.176
			MSE	0.0003	0.0012	0.0007	0.0027	0.0048	0.0831
3		(50%, 20%, 30%)	PE	1.012	1.010	2.019	1.993	8.076	8.149
			SE	0.013	0.034	0.018	0.048	0.007	0.185
			MSE	0.0003	0.0013	0.0007	0.0024	0.0059	0.0563
4		(30%, 20%, 50%)	PE	1.009	1.007	2.017	2.032	8.099	7.901
			SE	0.012	0.034	0.017	0.047	0.007	0.175
			MSE	0.0002	0.0012	0.0006	0.0033	0.0099	0.0404
5	Oversampling from only one extreme stratum	(50%, 40%, 10%)	PE	1.019	1.037	2.029	2.049	8.066	8.073
			SE	0.013	0.026	0.018	0.037	0.007	0.132
			MSE	0.0005	0.0021	0.0012	0.0037	0.0044	0.0228
6		(10%, 40%, 50%)	PE	1.009	1.014	2.015	2.013	8.084	7.974
			SE	0.013	0.025	0.018	0.036	0.007	0.143
			MSE	0.0002	0.0008	0.0005	0.0014	0.0071	0.0211
7	Oversampling from middle stratum	(25%, 50%, 25%)	PE	1.007	0.962	2.013	1.947	8.074	8.154
			SE	0.013	0.023	0.018	0.033	0.007	0.121
			MSE	0.0002	0.0019	0.0005	0.0038	0.0055	0.0385
8		(10%, 60%, 30%)	PE	1.011	1.006	2.016	2.001	8.064	7.978
			SE	0.013	0.022	0.018	0.031	0.007	0.124
			MSE	0.0003	0.0005	0.0006	0.0010	0.0041	0.0159
9		(30%, 60%, 10%)	PE	1.013	1.003	2.022	2.011	8.074	7.918
			SE	0.013	0.023	0.019	0.033	0.007	0.111
			MSE	0.0003	0.0005	0.0008	0.0012	0.0056	0.0190
10	Simple random sampling	(10%, 80%, 10%)	PE	1.006	0.987	2.011	1.986	8.076	8.051
			SE	0.013	0.022	0.019	0.031	0.007	0.117
			MSE	0.0002	0.0007	0.0005	0.0012	0.0058	0.0162

PE denotes the point estimate of the corresponding parameter. FLM denotes the full likelihood method. SE and MSE denote the standard error estimate and mean square error estimate of the corresponding estimator.

Table 4.8: Simulation results under sampling scheme A when $\tau = 0.4$

#	Sampling type (based on Y_2)	Sampling percentage (S_{21},S_{22},S_{23})		$\beta_{11} = 1$		$\beta_{12} = 2$		$\phi = 1.33$	
				FLM	IPW	FLM	IPW	FLM	IPW
1	Sampling from extreme strata only	(25%, 0%, 25%)	PE	1.015	1.872	2.021	3.491	1.321	1.491
			SE	0.016	0.033	0.023	0.045	0.010	0.030
			MSE	0.0005	0.7612	0.0010	2.2256	0.0003	0.0258
2	Oversampling from extreme strata	(20%, 10%, 20%)	PE	1.014	0.897	2.036	1.863	1.319	1.126
			SE	0.017	0.038	0.024	0.055	0.011	0.042
			MSE	0.0005	0.0121	0.0019	0.0217	0.0003	0.0449
3		(25%, 10%, 15%)	PE	1.012	0.840	1.991	1.688	1.318	1.200
			SE	0.017	0.038	0.024	0.056	0.010	0.043
			MSE	0.0004	0.0271	0.0007	0.1002	0.0003	0.0196
4		(15%, 10%, 25%)	PE	1.013	0.845	2.002	1.749	1.319	1.221
			SE	0.017	0.039	0.024	0.057	0.010	0.045
			MSE	0.0004	0.0257	0.0006	0.0661	0.0003	0.0145
5	Oversampling from only one extreme stratum	(25%, 20%, 5%)	PE	0.996	0.862	2.011	1.819	1.326	1.185
			SE	0.017	0.030	0.025	0.044	0.010	0.031
			MSE	0.0003	0.0200	0.0007	0.0348	0.0002	0.0230
6		(5%, 20%, 25%)	PE	1.014	0.893	1.994	1.792	1.317	1.180
			SE	0.017	0.029	0.025	0.045	0.010	0.033
			MSE	0.0005	0.0123	0.0007	0.0452	0.0004	0.0246
7	Oversampling from middle stratum	(12.5%, 25%, 12.5%)	PE	1.019	0.920	2.007	1.874	1.315	1.174
			SE	0.017	0.027	0.025	0.041	0.010	0.029
			MSE	0.0006	0.0071	0.0007	0.0176	0.0004	0.0263
8		(5%, 30%, 15%)	PE	1.006	0.903	1.993	1.843	1.323	1.240
			SE	0.017	0.026	0.025	0.040	0.010	0.030
			MSE	0.0003	0.0100	0.0007	0.0262	0.0002	0.0097
9		(15%, 30%, 5%)	PE	0.992	0.893	2.022	1.900	1.327	1.225
			SE	0.017	0.027	0.025	0.041	0.010	0.029
			MSE	0.0003	0.0122	0.0011	0.0118	0.0001	0.0125
10	Simple random sampling	(5%, 40%, 5%)	PE	1.012	0.937	1.987	1.958	1.316	1.272
			SE	0.017	0.025	0.026	0.041	0.010	0.028
			MSE	0.0004	0.0046	0.0008	0.0035	0.0004	0.0046

PE denotes the point estimate of the corresponding parameter. FLM denotes the full likelihood method. SE and MSE denote the standard error estimate and mean square error estimate of the corresponding estimator.

Table 4.9: Simulation results under sampling scheme A when $\tau = 0.8$

#	Sampling type (based on Y_2)	Sampling percentage (S_{21},S_{22},S_{23})		$\beta_{11} = 1$		$\beta_{12} = 2$		$\phi = 8$	
				FLM	IPW	FLM	IPW	FLM	IPW
1	Sampling from extreme strata only	(25%, 0%, 25%)	PE	1.011	1.693	2.016	3.009	8.069	8.563
			SE	0.012	0.025	0.017	0.037	0.007	0.154
			MSE	0.0003	0.4806	0.0006	1.0189	0.0049	0.3407
2	Oversampling from extreme strata	(20%, 10%, 20%)	PE	1.012	0.857	2.020	1.797	8.077	7.473
			SE	0.013	0.033	0.017	0.047	0.007	0.178
			MSE	0.0003	0.0214	0.0007	0.0436	0.0060	0.3093
3		(25%, 10%, 15%)	PE	1.014	0.881	2.022	1.815	8.069	7.232
			SE	0.013	0.034	0.018	0.047	0.007	0.174
			MSE	0.0003	0.0154	0.0008	0.0363	0.0049	0.6202
4		(15%, 10%, 25%)	PE	1.014	0.932	2.021	1.910	8.052	7.107
			SE	0.013	0.035	0.017	0.050	0.007	0.178
			MSE	0.0004	0.0059	0.0008	0.0106	0.0027	0.8295
5	Oversampling from only one extreme stratum	(25%, 20%, 5%)	PE	1.009	0.867	2.014	1.828	8.060	7.179
			SE	0.013	0.027	0.018	0.038	0.007	0.136
			MSE	0.0002	0.0184	0.0005	0.0310	0.0037	0.6926
6		(5%, 20%, 25%)	PE	1.003	0.878	2.006	1.836	8.081	7.616
			SE	0.013	0.028	0.018	0.039	0.007	0.147
			MSE	0.0002	0.0157	0.0004	0.0285	0.0066	0.1693
7	Oversampling from middle stratum	(12.5%, 25%, 12.5%)	PE	1.014	0.920	2.023	1.888	8.092	7.560
			SE	0.013	0.025	0.018	0.036	0.007	0.131
			MSE	0.0003	0.0070	0.0008	0.0137	0.0085	0.2104
8		(5%, 30%, 15%)	PE	1.008	0.902	2.014	1.863	8.079	7.778
			SE	0.013	0.024	0.018	0.034	0.007	0.127
			MSE	0.0002	0.0101	0.0005	0.0199	0.0063	0.0655
9		(15%, 30%, 5%)	PE	1.015	0.940	2.023	1.918	8.068	7.544
			SE	0.013	0.024	0.018	0.034	0.007	0.124
			MSE	0.0004	0.0042	0.0009	0.0080	0.0046	0.2236
10	Simple random sampling	(5%, 40%, 5%)	PE	1.008	0.940	2.013	1.922	8.064	7.802
			SE	0.013	0.023	0.018	0.032	0.007	0.117
			MSE	0.0002	0.0041	0.0005	0.0070	0.0042	0.0531

PE denotes the point estimate of the corresponding parameter. FLM denotes the full likelihood method. SE and MSE denote the standard error estimate and mean square error estimate of the corresponding estimator.

Table 4.10: Simulation results under sampling scheme B when $\tau = 0.4$

#	Sampling type (based on Y_2)	Sampling percentage (S_{21},S_{22},S_{23})		$\beta_{11} = 1$		$\beta_{12} = 2$		$\phi = 1.33$	
				FLM	IPW	FLM	IPW	FLM	IPW
1	Sampling from extreme strata only	(25%, 0%, 25%)	PE	1.015	0.918	2.035	1.831	1.313	1.165
			SE	0.019	0.028	0.024	0.039	0.011	0.029
			MSE	0.0006	0.0075	0.0018	0.0301	0.0005	0.0293
2	Oversampling from extreme strata	(20%, 10%, 20%)	PE	1.028	0.957	2.007	1.844	1.312	1.235
			SE	0.018	0.026	0.024	0.037	0.011	0.027
			MSE	0.0011	0.0026	0.0006	0.0256	0.0006	0.0105
3		(25%, 10%, 15%)	PE	0.993	0.896	1.977	1.761	1.328	1.256
			SE	0.019	0.026	0.024	0.037	0.011	0.028
			MSE	0.0004	0.0115	0.0011	0.0585	0.0001	0.0067
4		(15%, 10%, 25%)	PE	1.014	0.905	2.034	1.863	1.315	1.258
			SE	0.019	0.026	0.024	0.037	0.011	0.028
			MSE	0.0005	0.0097	0.0017	0.0201	0.0004	0.0065
5	Oversampling from only one extreme stratum	(25%, 20%, 5%)	PE	0.996	0.918	2.010	1.896	1.326	1.227
			SE	0.019	0.025	0.025	0.036	0.011	0.026
			MSE	0.0004	0.0073	0.0007	0.0120	0.0002	0.0120
6		(5%, 20%, 25%)	PE	1.016	0.957	1.997	1.926	1.315	1.300
			SE	0.018	0.024	0.025	0.036	0.011	0.027
			MSE	0.0006	0.0025	0.0006	0.0067	0.0004	0.0018
7	Oversampling from middle stratum	(12.5%, 25%, 12.5%)	PE	1.006	0.953	1.991	1.900	1.320	1.259
			SE	0.019	0.024	0.025	0.035	0.011	0.025
			MSE	0.0004	0.0028	0.0007	0.0111	0.0003	0.0061
8		(5%, 30%, 15%)	PE	1.021	0.980	2.018	1.959	1.310	1.296
			SE	0.019	0.024	0.026	0.035	0.011	0.027
			MSE	0.0008	0.0010	0.0010	0.0029	0.0007	0.0021
9		(15%, 30%, 5%)	PE	1.013	0.967	2.041	1.958	1.316	1.251
			SE	0.019	0.024	0.025	0.035	0.011	0.025
			MSE	0.0005	0.0017	0.0023	0.0030	0.0004	0.0073
10	Simple random sampling	(5%, 40%, 5%)	PE	0.998	0.965	2.002	1.974	1.326	1.341
			SE	0.019	0.024	0.026	0.036	0.011	0.027
			MSE	0.0004	0.0018	0.0007	0.0020	0.0002	0.0008

PE denotes the point estimate of the corresponding parameter. FLM denotes the full likelihood method. SE and MSE denote the standard error estimate and mean square error estimate of the corresponding estimator.

Table 4.11: Simulation results under sampling scheme B when $\tau = 0.8$

#	Sampling type (based on Y_2)	Sampling percentage (S_{21},S_{22},S_{23})		$\beta_{11} = 1$		$\beta_{12} = 2$		$\phi = 8$	
				FLM	IPW	FLM	IPW	FLM	IPW
1	Sampling from extreme strata only	(25%, 0%, 25%)	PE	1.017	0.909	2.026	1.871	8.085	7.567
			SE	0.013	0.025	0.018	0.036	0.007	0.131
			MSE	0.0004	0.0089	0.0010	0.0180	0.0072	0.2042
2	Oversampling from extreme strata	(20%, 10%, 20%)	PE	1.017	0.939	2.026	1.906	8.063	7.557
			SE	0.013	0.023	0.018	0.032	0.007	0.119
			MSE	0.0005	0.0042	0.0010	0.0099	0.0040	0.2100
3		(25%, 10%, 15%)	PE	1.013	0.957	2.021	1.938	8.061	7.556
			SE	0.013	0.023	0.018	0.032	0.007	0.117
			MSE	0.0003	0.0023	0.0008	0.0049	0.0037	0.2106
4		(15%, 10%, 25%)	PE	1.018	0.938	2.028	1.910	8.050	7.644
			SE	0.013	0.023	0.018	0.032	0.007	0.120
			MSE	0.0005	0.0044	0.0011	0.0090	0.0026	0.1410
5	Oversampling from only one extreme stratum	(25%, 20%, 5%)	PE	1.013	0.935	2.023	1.921	8.085	7.669
			SE	0.013	0.022	0.018	0.031	0.007	0.111
			MSE	0.0003	0.0047	0.0009	0.0072	0.0073	0.1220
6		(5%, 20%, 25%)	PE	1.005	0.931	2.013	1.904	8.100	8.219
			SE	0.013	0.021	0.018	0.030	0.007	0.118
			MSE	0.0002	0.0052	0.0005	0.0101	0.0100	0.0620
7	Oversampling from middle stratum	(12.5%, 25%, 12.5%)	PE	1.009	0.976	2.013	1.978	8.049	7.599
			SE	0.013	0.022	0.018	0.030	0.007	0.109
			MSE	0.0002	0.0010	0.0005	0.0014	0.0025	0.1730
8		(5%, 30%, 15%)	PE	1.007	0.965	2.014	1.948	8.067	7.771
			SE	0.013	0.021	0.018	0.030	0.007	0.111
			MSE	0.0002	0.0017	0.0005	0.0036	0.0046	0.0646
9		(15%, 30%, 5%)	PE	1.007	0.957	2.014	1.944	8.099	7.765
			SE	0.013	0.021	0.018	0.030	0.007	0.110
			MSE	0.0002	0.0023	0.0005	0.0041	0.0099	0.0673
10	Simple random sampling	(5%, 40%, 5%)	PE	1.012	0.989	2.019	1.987	8.093	8.153
			SE	0.013	0.021	0.018	0.029	0.007	0.111
			MSE	0.0003	0.0005	0.0007	0.0010	0.0087	0.0357

PE denotes the point estimate of the corresponding parameter. FLM denotes the full likelihood method. SE and MSE denote the standard error estimate and mean square error estimate of the corresponding estimator.

Chapter 5

Conclusion

We considered response-dependent two-phase sampling designs. We assessed performance of some well-known likelihood and pseudo-likelihood based methods. In addition, efficiency of response-dependent two-phase sampling designs under likelihood-based and pseudo-likelihood methods were investigated.

In response-dependent two-phase sampling designs when there is no inexpensive covariate, we considered five sampling designs in phase II: sampling all units in the extreme strata and no units from the middle stratum (extreme sampling), sampling less number of units from the middle stratum and sampling more units from the extreme strata, sampling more units from only one extreme stratum, sampling more units from the middle stratum compared to the extreme strata and simple random sampling design. When the expensive covariate X is not excessively equal to 0, we concluded that likelihood-based estimation methods performed well under each sampling setting that we considered and they yielded the most efficient estimates under the extreme sampling design. Pseudo-likelihood estimation methods give biased estimates under the extreme sampling design. On the other hand, the sampling designs selecting more units from the central stratum give consistent and more efficient estimates under the IPW estimation method. However, likelihood-based yielded more efficient estimators than the IPW method even under these sampling designs. The estimated pseudo-likelihood method performed well under each design except the extreme sampling, and oversampling from extreme strata while selecting some from the middle stratum gives the most efficient estimate. Among all the estimation

methods, the IPW estimation method performed the worst. When the expensive covariate X is excessively equal to 0, under the extreme sampling design, the relative efficiency of the extreme sampling design is low compared to many other designs. When X and Y are associated, estimated pseudo-likelihood estimation method gives the most efficient estimates under many sampling design settings except the extreme sampling design. The IPW estimation method yields the least efficient designs. As expected, we observed that the estimation methods give less efficient estimators compared to the case where the expensive covariate X is not excessively equal to 0.

In Chapter 3, we considered sampling designs which depend on inexpensive covariate in addition to the response variable and assumed that inexpensive covariate Z is a categorical variable with three levels and one level of Z be rare. We investigated two sampling designs in phase II: sampling all of the units from the rare group and sampling all the extreme strata units under the other two groups, and sampling all of the units from the rare group and more sampling units from the middle stratum of the other two groups. We compared these designs with the simple random sampling. We conclude that the likelihood-based methods give consistent and relatively efficient estimates under each sampling design considered. In addition, likelihood-based methods provide the most efficient estimates when sampling all of the units from the rare group and sampling all the extreme strata units under the other two groups of Z. However, under the simple random sampling and sampling all of the units from the rare group of Z and sampling more units from the middle stratum of the other two groups of Z, the conditional likelihoods yield a little less efficient estimates compared to the full likelihood methods and the estimated pseudo-likelihood method. The pseudo-likelihood methods performed well under the simple random sampling design.

In Chapter 4, we extended the sampling designs that depend on multiple response variable setting. Under bivariate response variable setting, we considered the extreme sampling design and the design in which we sample more from the central stratum for one response variable. Under each of these designs, we investigated five sampling designs that we considered in Chapter 2 for the other response variable. We conclude that the full likelihood method performed well compared to the IPW method. Also, the full likelihood method gives consistent and the most efficient estimates for the coefficient of the expensive covariate under the extreme sampling for each response

variable. However, the IPW estimation method performed the worst under many of the sampling designs considered.

In summary, the extreme sampling setting is desirable under likelihood-based estimation methods but when pseudo-likelihood methods are applied this design yields biased estimates. However, when X is excessively equal to 0, the relative efficiency of the extreme sampling design is low compared to many other designs under likelihood-based estimation methods. The central sampling setting performed better under pseudo-likelihood estimation methods. However, likelihood-based estimation methods yielded more efficient coefficient estimates compared to pseudo-likelihood estimation methods even under such sampling designs.

Bibliography

[1] Allison, D. B. (1997). Transmission-disequilibrium tests for quantitative traits. *American Journal of Human Genetics*, 60(3):676–690.

[2] Barnett, I. J., Lee, S., and Lin, X. (2013). Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genetic Epidemiology*, 37(2):142–151.

[3] Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20.

[4] Breslow, N. E. and Chatterjee, N. (1999). Design and analysis of twophase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C Applied Statistics*, 48(4):457–468.

[5] Breslow, N. E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):447–461.

[6] Carroll, R. J., Wang, S., and Wang, C. Y. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association*, 90(429):157–169.

[7] Chatterjee, N., Chen, Y.-H., and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168.

[8] Chen, H. Y. and Li, M. (2011). Improving power and robustness for detecting genetic association with extremevalue sampling design. *Genetic Epidemiology*, 35(8):823–830.

[9] Chen, K. (2001). Generalized case–cohort sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):791–809.

[10] Chen, Z., Zheng, G., Ghosh, K., and Li, Z. (2005). Linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *The American Journal of Human Genetics*, 77(4):661–669.

[11] Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.

[12] Embrechts, P., Lindskog, F., and McNeil, A. (2003). *Modelling Dependence with Copulas and Applications to Risk Management*. In: Rachev S (ed) Handbook of Heavy Tailed Distributions in Finance. Elsevier, New York, pp. 329-384.

[13] Espin-Garcia, O., Craiu, R. V., and Bull, S. B. (2018). Two-phase designs for joint quantitative-trait-dependent and genotype-dependent sampling in post-gwas regional sequencing. *Genetic Epidemiology*, 42(1):104–116.

[14] Flanders, W. D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10(5):739–747.

[15] Gray, K. A., Klebanoff, M. A., Brock, J. W., Zhou, H., Darden, R., Needham, L., and Longnecker, M. P. (2005). In utero exposure to background levels of polychlorinated biphenyls and cognitive functioning among school-age children. *American Journal of Epidemiology*, 162(1):17–26.

[16] Hauksson, H. A., Dacorogna, M., Domenig, T., Mller, U., and Samorodnitsky, G. (2001). Multivariate extremes, aggregation and risk estimation. *Quantitative Finance*, 1(1):79 – 95.

[17] Hsieh, P. A., Neuman, S. P., Stiles, G. K., and Simpson, E. S. (1985). Field determination of the threedimensional hydraulic conductivity tensor of anisotropic media: 2. methodology and application to fractured rocks. *Water Resources Research*, 21(11):1667–1676.

[18] Huang, B. E. and Lin, D. Y. (2007). Efficient association mapping of quantitative trait loci with selective genotyping. *The American Journal of Human Genetics*, 80(3):567–576.

[19] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall.

[20] Lawless, J. (2018). Two-phase outcome-dependent studies for failure times and testing for effects of expensive covariates. *Lifetime Data Analysis*, 24(1):28–44.

[21] Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(2):413–438.

[22] Li, D., Lewinger, J. P., Gauderman, W. J., Murcray, C. E., and Conti, D. (2011). Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic Epidemiology*, 35(8):790–799.

[23] Lin, D.-Y., Zeng, D., and Tang, Z.-Z. (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences*, 110(30):12247–12252.

[24] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

[25] Longnecker, M. P., Hoffman, H. J., Klebanoff, M. A., Brock, J. W., Zhou, H., Needham, L., Adera, T., Guo, X., and Gray, K. A. (2004). In utero exposure to polychlorinated biphenyls and sensorineural hearing loss in 8-year-old children. *Neurotoxicology and Teratology*, 26(5):629 – 637.

[26] Manski, C. and McFadden, D. (1981). *Structural Analysis of Discrete Data with Econometric Applications*. The MIT Press.

[27] Nelsen, R. (2006). *An Introduction to Copulas, 2nd edition.* Springer, New York.

[28] Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116.

[29] Page, G. P. and Amos, C. I. (1999). Comparison of linkage-disequilibrium methods for localization of genes influencing quantitative traits in humans. *The American Journal of Human Genetics*, 64(4):1194–1205.

[30] Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11.

[31] Robins, J. M., Hsieh, F., and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):409–424.

[32] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

[33] Schaid, D. J., Jenkins, G. D., Ingle, J. N., and Weinshilboum, R. M. (2013). Two-phase designs to follow-up genome-wide association signals with DNA resequencing studies. *Genetic Epidemiology*, 37(3):229–238.

[34] Scott, A. J. and Wild, C. J. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics*, 47(2):497–510.

[35] Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):57–71.

[36] Scott, A. J. and Wild, C. J. (1998). Computing maximum likelihood estimates for case-control studies. *Unpublished*.

[37] Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231.

[38] Slatkin, M. (1999). Disequilibrium mapping of a quantitative-trait locus in an expanding population. *The American Journal of Human Genetics*, 64(6):1764–1772.

[39] Wacholder, S. and Weinberg, C. R. (1994). Flexible maximum likelihood methods for assessing joint effects in case- control studies with complex sampling. *Biometrics*, 50(2):350–357.

[40] Wallace, C., Chapman, J. M., and Clayton, D. G. (2006). Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *The American Journal of Human Genetics*, 78(3):498–504.

[41] Weaver, M. A. and Zhou, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association*, 100(470):459–469.

[42] White, H. (1982a). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

[43] White, J. E. (1982b). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128.

[44] Wild, C. J. (1991). Fitting prospective regression models to case-control data. *Biometrika*, 78(4):705–717.

[45] Xiong, M., Fan, R., and Jin, L. (2002). Linkage disequilibrium mapping of quantitative trait loci under truncation selection. *Human Heredity*, 53(3):158–172.

[46] Yilmaz, Y. E. and Bull, S. B. (2011). Are quantitative trait-dependent sampling designs cost-effective for analysis of rare and common variants? *BMC Proceedings*, 5(9):S111.

[47] Zhao, L. P. and Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine*, 11(6):769–782.

[48] Zhou, H., Chen, J., Rissanen, T. H., Korrick, S. A., Hu, H., Salonen, J. T., and Longnecker, M. P. (2007). An efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology*, 18(4):461–468.

[49] Zhou, H., Weaver, M. A., Qin, J., Longnecker, M. P., and Wang, M. C. (2002). A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 58(2):413–421.

Appendix A

Simulation results under the IPW estimation method

Table A.1: Simulation results under the IPW estimation method

π_1	π_2	π_3	Estimates		Standard Error	
			$\hat{\beta}_{1,Eq}$	$\hat{\beta}_{1,Sy}$	SE_{Eq}	SE_{Sy}
100%	0%	100%	3.071	3.071	0.044	0.044
80%	5%	80%	1.061	1.061	0.038	0.037
100%	5%	60%	1.027	1.027	0.039	0.037
60%	5%	100%	1.043	1.043	0.038	0.037
100%	10%	20%	1.051	1.051	0.032	0.030
20%	10%	100%	0.993	0.993	0.031	0.029
50%	12.5%	50%	1.041	1.041	0.027	0.026
20%	15%	60%	0.999	0.999	0.028	0.026
60%	15%	20%	1.055	1.055	0.028	0.027
20%	20%	20%	1.019	1.019	0.029	0.027

$\hat{\beta}_{1,Eq}$ and $\hat{\beta}_{1,Sy}$ are obtained by solving estimating equation and survey package in R respectively where true value of β_1 is 1.