SEMIPARAMETRIC APPROACHES FOR AUXILIARY AND INCOMPLETE COVARIATE UNDER OUTCOME DEPENDENT SAMPLING DESIGNS

Wansuk Choi

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2017

Approved by:

Haibo Zhou

Gary Koch

Fei Zou

Mark Weaver

Sylvia Becker-Dreps

ABSTRACT

Wansuk Choi: Semiparametric Approaches for Auxiliary and Incomplete Covariate under Outcome Dependent Sampling Designs
(Under the direction of Haibo Zhou)

In epidemiologic and biological studies, investigators seek to establish relationships between a response variable and expensive risk factors. In reality, to obtain exact measurement about the covariate of interest can be difficult due to the limitation of budget or missingness of covariate while outcome and auxiliary information for covariate of interest are relatively cheap to collect. In these circumstances, Outcome-Dependent sampling (ODS) and Outcome-Auxiliary dependent sampling (OADS) can be used to gain efficiency since those sampling designs incorporate additional information into parameter estimates.

In this proposal, we propose three topics : (1) an estimated likelihood under ODS including missing in a covariate; (2) an updating method under a two-stage ODS for continuous response outcome; (3) an updating method under two-stage OADS for binary outcome. (1) and (2) are developed for continuous outcome variable case and (3) is developed to handle binary outcome variable cases. The first topic uses an estimated likelihood approach which is an extension of the method in Weaver and Zhou (2005) to a single-stage ODS sample that includes missing in a covariate of interest while the method in Weaver and Zhou (2005) is developed for two-stage cohort study without an auxiliary information. The second topic considers a semiparametric empirical likelihood method (Owen,1988, 1990; Qin and Lawless, 1994; Zhou et al., 2002, Wang and Zhou, 2006) at the second stage and updates estimators from the second stage by incorporating auxiliary information from data at the first stage. In the third topic we develop a new sampling scheme using auxiliary covariate information in a two-stage prospective study. For all three topics, the consistency and asymptotic distributions of proposed estimators are

established, the finite sample performances are demonstrated through simulations studies, and real data application to the Collaborative Perinatal Project (CPP, Niswander and Gordon, 1972) are illustrated. The results from our methods show that one could gain efficiency by using auxiliary information.

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Dr. Haibo Zhou, for enlightening me about doing statistical scientific research. He taught me how to be a good statistician by always demonstrating ethical practice and critical insight in all of his research experience. (His insight, experiences, and research ethics will remind me of being a good statistician.) Dr. Sylvia Becker-Dreps has also been a positive role model as well as a passionate researcher. She has been very supportive and considerate during our work together. In addition, I would like to thank Dr. Mark Weaver for sharing his knowledge of Outcome-Dependent sampling design. I would also like to thank Dr. Gary Koch and Dr. Fei Zou for pleasantly accepting to be my committee members. I owe my thanks to Dr. Michale Hudgens too. He has shown me to be a great problem-solver in real data problems. Finally, I would like to specially thank to Dr. Jianwen Cai, Dr. Michael Kosorok, and Dr. Donglin Zeng for helping me and supporting me while I was preparing for my qualifying exams.

I could not complete my journey without my parents, Kwangdeok and Insuk, for always rooting for me in South Korea. I would like to thank my two sarcastic sisters, Jinkyoung and Jinhye as well for their constant humor through all my years of work. All my family members have given me unconditional support and love leading to my success on this long journey.

I would also like to thank all my friends in Chapel Hill. During my time at UNC, they made it possible for me to recharge and relax. They also opened my eyes to different perspectives attitudes to research and life. I would like to mention all of them here: Hyowon Ahn, Jean Ahn, Chanil Boo, Byeongyeob Choi, Heesun Choi, Hyoyoung Choi, Wonil Chung, Yunro Chung, Jonathan Hibbard, Melissa Hobgood, Kuan-Chieh Huang, Junpyo Hong, Jungin Kim, Kyung Hee Kim, Kyungsu Kim, Minki Kim, Sangwan Kim, Sunhyung Kim, Duyeol Lee, Joohwi Lee,

Sooyoung Lee, Rachel Lithman, Paul Little, Fang-Shu Ou, Jackie Eunjung Relyea, Pratyaydipta Rudra, Hojin Yang.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Observational study and randomized clinical trials (RCT) are two main streams of epidemiological/medical research. RCT is used when researchers can control an experimental situation in investigating efficacy of interventions. However, such strict control may bring losses in generalizing results and raise some ethical issues. Additionally, even though investigators can assign patients to smoking or non-smoking groups, since there are many other factors that cause lung cancer, strict allocation may not reflect reality. For instance, suppose that an investigator wants to investigate the relationship between lung cancer and smoking. In this case, there is an ethical issue if an investigator assigns participants to smoking group or non-smoking group. Observational studies, on the other hand, are useful to generalize results and can properly deal with ethical issues as well. These differences make observational study an available alternative to RTC able to solve a variety of questions in epidemiological studies.

Well-designed observational studies can provide results similar to those of randomized controlled trials. Cohort studies and case-control studies are two primary examples of observational study types that aid in evaluating associations between diseases and exposures.

First, the cohort study is a widely used type of study in epidemiological studies. The cohort has been used in epidemiological study to define a group of people followed over a period of time. The modern definition is of a group of people with defined characteristics with whom researchers follow-up to determine incidence of, or mortality from, some specific disease, cause of death, or some other outcome. In examining rare exposure diseases, cohort studies are useful because subjects are selected by their exposure status. However, the large sample size and long

follow-up period of the cohort study design may require big budget. There are two types of cohort studies: prospective and retrospective. Prospective studies are conducted from the present time to the future. They offer an advantage in that researchers can design a sampling scheme. However, the follow-up for a prospective study could take time to obtain the aimed sample size. Thus, prospective studies would create cost issues. In contrast, retrospective cohort studies are relatively inexpensive. Retrospective cohort studies are conducted from the past to the present and examine events or outcomes in which researchers are interested. In retrospective cohort studies that are based on a cohort of subjects is chosen at the present time, and outcome data (i.e. disease status, event status), are reconstructed for analysis. A limitation of this study design is that the existing data may be incomplete or inaccurate. An advantage of this study design is that it is less costly than prospective cohort studies. In cohort studies, researchers develop sampling designs to obtain consistent and efficient estimators under given statistical models. The comparison and explanation of prospective and retrospective designs introduced this section are nicely described in Mann (2003) and Song and Chung (2010).

On the other hand, in the two-stage sampling design, studied in Breslow and Cain (1988), Breslow and Chatterjee (1999), Breslow and Wellner (2007), and Wang and Zhou (2010), disease and exposure status are obtained in the first stage and the covariates of interest are determined in the second stage. This design allows for more efficient estimates since information about covariates is obtained only from the second stage with its stratified subsample of first stage subjects. Prentice (1986) proposed a two stage design with the identified disease status in the first stage ; those methods are in line with the case-control design where oversampling the cases provides more information about the rare event population than an SRS design with the same sample size and cost. Breslow and Cain (1988) put forth a double sampling design in which the preliminary sample includes separate sampling from subpopulations of responders and non-responders. In the two-phase sampling design that we are interested in here, a response variable and an auxiliary variable are observed in the first phase, and, in the second phase, covariates of interests are observed. Two-phase sampling is particularly useful when the cost

for observing auxiliary variables is relatively cheap compared to the cost for obtaining covariates of interest. In summary, two-phase sampling can be described as follows : (1) Draw N primary units in first stage; (2) at second stage, draw n_i samples from the *i*th unit in first stage. Note that a subset of elements in primary units is randomly selected in second stage. White (1982) showed that, in studies of the relationship between a rare disease and a rare exposure to a risk factor, the efficiency of the two-stage design is gained by sampling a large proportion of the subjects from the small groups and a smaller proportion of those from the large groups. In particular, this approach can be useful to enhance study efficiency under a budget limitation.

The case-control study is widely used in epidemiological study to assess factors related to disease incidence. In case-control studies, subjects are identified by outcome status and followed to certain exposure levels. Once outcome status is identified, and subjects are categorized as cases, controls are selected. Data about exposure to a risk factor or several risk factors are then collected retrospectively. Case-control studies are widely used to investigate rare outcomes because subjects are selected from the outset by their outcome status. Since disease status is already known, researchers can start the study faster than a cohort study. Moreover, most case-control studies already have data prepared, so investigators may see other risk factors together with exposure status. Comparing with prospective cohort studies, case-control studies are relatively inexpensive and quick to implement, require fewer subjects, and are flexible in terms of allowing for multiple exposures or risk factors to be assessed. Prentice and Pyke (1979) extended results on the equivalence of odds ratio estimators when both prospective and retrospective logistic models are applied to case-control data, which were investigated by Breslow and Powers (1978). Scott and Wild (1986) compared weighted likelihood approaches with unweighted likelihood approaches in case-control or choice based sampling designs. Zhao and Lipsitz (1992) introduced introduced twelve different two-stage case-control designs. Wang and Zhou (2010) developed an Outcome Auxiliary Dependent Sampling under two-stage design for the logistic regression model.

With biased sampling schemes, estimation methods used for SRS schemes create bias issues

since those methods don't take into account characters of biased sampling schemes. For example, under a linear model, the well-known least square estimate approach may bring biased estimates. The reason is as follows : Suppose that one draws samples under a given condition such that $Y \leq c_1$. Drawing a sample from $Y \leq c_1$ can be expressed as

$$P(Y, X|Y \leq c_1) = \frac{P(Y \mid X : \theta)G_X(X)}{\pi_1(\theta, G_X)},$$

where G_X is the marginal distribution function of X and $\pi_1 = \int_X P(Y \leq c_1|X;\theta)dG_X(x)$. If we treat the biased sample above as a simple random sample, we cannot account for $\pi_1(\theta, G_X)$ and the result is that biased estimates can be obtained. The inverse-weighted probability type method, which was developed by Horvitz and Thompson (1952), gives unbiased estimates and is easy to implement, but it is not the most efficient estimator. When the outcome variable is continuous, the outcome can be categorized and one can model the categorized response with covariates of interest. However, this procedure would cause a loss of information and it could cause misclassification of subjects. Suissa (1991) pointed out that categorizing a continuous outcome could losses of information and bias issues if the categorized continuous response is used under a categorical response variable model, such as logistic regression.

In this dissertation, we propose statistical inference methods that are used under an Outcome Dependent Sampling (ODS) design. Since ODS is a biased sampling scheme, its ODS character is reflected through proposed methods. Zhou et al. (2002) proposed an outcome dependent sampling design that is reviewed in the next section. Weaver and Zhou (2005) developed the two-stage outcome dependent sampling, which incorporates non-validation samples to enhance efficiency of estimates. Along with those two designs, outcome auxiliary dependent sampling (OADS) was developed by Zhou, Wu, Liu and Cai (2011) to include more information about covariates from an auxiliary variable. With logic similar to that used in continuous response cases, Wang and Zhou (2006) and Wang and Zhou (2010) developed a two-component and two-stage OADS for a discrete response variable. In addition to ODS and OADS designs, there

have been many considerations to obtain more efficient estimators under two-stage sampling designs. For example, Chen and Chen (2000) developed the unified approach in two-stage SRS design to enhance efficiency of estimates by using auxiliary information. Chatterjee et al. (2003) proposed a pseudo-likelihood method that includes auxiliary information into their likelihood function. We will review ODS and OADS according to different sampling designs and statistical methodology in later sections.

In this dissertation, we will focus on three topics, the first one is Outcome Dependent Sampling which has missing in covariates. The second topic is an updating method that uses estimates from an Outcome Dependent Sampling under two-stage design. While the first and second topics handle continuous outcome variable, the third topic is Outcome-auxiliary dependent sampling (OADS) with a binary response under two-stage sampling design. The first topic extends the ODS design in Zhou et al. (2002) to an ODS that includes missing observations in covariate. Since missing data can be an issue in rare event studies, it is more closer to reality than the one considered in Zhou et al. (2002). The second topic considers two-phase ODS design having continuous auxiliary variable for covariate of interest. In the second topic, we develop an updating method in two-phase ODS design as an alternative to the estimated likelihood approach in Weaver and Zhou (2005). Our approach requires less computational effort than the estimated likelihood method and can easily handle continuous auxiliary variable. The third topic considers two-phase OADS design with updating estimators from sample at the second stage. We extends OADS in Wang and Zhou (2006) to two-stage design to gain efficiency.

The layout of following sections in this chapter is as follows. In Section 1.2, we review ODS and OADS schemes according to data structures and methodologies in ODS and OADS problems. In Section 1.3, terms and notations to measurement error and auxiliary information are introduced. Section 1.4 contains several well-known approaches to measurement error problems. In Section 1.5, we introduce proposed methods that correspond to two ODS designs.

1.2 Outcome Dependent Sampling (ODS) / Outcome Auxiliary Dependent Sampling designs (OADS)

In this section, we review Outcome Dependent Sampling (ODS) and Outcome Auxiliary Dependent Sampling (OADS) designs that correspond to types of outcome variable.

1.2.1 ODS-type designs with discrete outcome variable

Two-component design with discrete outcome

Outcome and auxiliary-dependent sampling design (hereafter, OADS) for a categorical outcome variable was proposed by Wang and Zhou (2006). Let Y denote the categorical outcome where $\{Y = j : j = 1, \cdots, J\}$, X denote the covariates of interest, Z denote a confounding variable, and W denote a discrete auxiliary variable for X where $\{W = k : k = 1, 2, \cdots, K\}$. For the SRS component, $\{Y_i, X_i, Z_i, W_i\}$ for $i = 1, \cdots, m$ are observed. For the supplementary ODS part, $\{X_i | Y_i = j, W_i = k\}$ for $i = 1, 2, \cdots, n_{jk}$ with given j and k. Thus, the total sample size of OADS is $n = m + \sum_{k=1}^{K} \sum_{j=1}^{J} n_{jk}$.
The likelihood function from two-component OADS design is derived as

$$\begin{aligned} L(\beta, \{p_{ik}\}, \{\pi_{jk}\}) &= \prod_{k=1}^{K} \prod_{i \in U_k} P(Y_i | X_i; \beta) dG(X_i | W_i = k) \\ &\times \prod_{k=1}^{K} \prod_{j=1}^{J} \prod_{i \in V_{jk}} \frac{P(Y_i = j | X_i; \beta)}{P(Y_i = j | W_i = k)} dG(X_i | W_i = k) \\ &= \prod_{k=1}^{K} \prod_{i \in U_k + V_k} P(Y_i | X_i; \beta) \times \prod_{k=1}^{K} \prod_{i \in U_k + V_k} p_{ik} \times \pi_{jk}^{-n_{jk}} \end{aligned}$$

where U_k is the set for observations in SRS with $W = k$, V_{jk} is the set for observations in the supplementary set with $Y = j$ and $W = k$, $p_{ik} = g(X_i | W_i = k)$ and $\pi_{jk} = P(Y_i = j | W_i = k) = \int P(Y = j | X; \beta) dG(X | W = k)$. They developed a semiparametric empirical likelihood method to avoid the problem of infinite dimension parameter. Concisely, to profile $L(\beta, \{p_{ik}\}, \{\pi_{jk}\})$, $(\beta', \{\pi_{jk}\}')$ are fixed and, under the constraints, $\{p_{ik} \geq 0, \sum_{i \in U_k + V_k} p_{ik} = 1, \sum_{i \in U_k + V_k} p_{ik} P(y =$

$1|x_i;\beta) = \pi_{1k}\}$ for $k = 1,\cdots,K$, $\hat{p}_{ik}$'s are obtained by using the Lagrangian multipliers. Then, $(\beta', \{\pi_{jk}\}')$ can be estimated by the Newton-Raphson method. In particular, this method works well when it comes to the rare event studies. In addition, by incorporating SRS sample, OADS sample could take account of information about underlying population.

Two-stage OADS including non-validation dataset with discrete outcome

Wang and Zhou (2010) extended the OADS to a two-stage sampling to use more information from incomplete observations with auxiliary information for covariates. The main idea is to achieve higher efficiency by extracting information from the invalid set. Let Y be a categorical outcome with possible values from 1 to K. Denote X as covariates of interest and let W be the auxiliary variable for X. Assume that $P(Y|X,W) = P(Y|X) = h(\beta_0 + \beta_1 X)$, where $h^{-1}(\cdot)$ is a known link function. Let $\{c_r\}$ be real numbers where $-\infty = c_0 < c_1 < \cdots < c_{R-1} < c_R = \infty$ on W and W, $\{(c_{r-1}, c_r]\}$ for $r = 1,\cdots,R$ are mutually exclusive intervals. The combination of $Y \times C$ partitions the study cohort into a total $K \times R$ strata. Notations for the dataset can be given as follows : N is size of population dataset, N_{rk} is size of the stratum $\{Y = k, C = r\}$ in study cohort, V_{rk} means set of the stratum $\{Y = k, C = r\}$ in validation sample, n_{rk} is size of OADS subsample from the stratum $\{Y = k, C = r\}$, $\overline{n}_{rk}$ denotes size of remaining set of subjects in the stratum $\{Y = k, C = r\}$ excluding n_{rk}, and $\overline{V}_{rk}$ is the set of the stratum $\{Y = k, C = r\}$ in non-validation sample. Note that $N = \sum_{r=1}^{R}\sum_{k=1}^{K} N_{rk}$, $\overline{n}_{rk} = N_{rk} - n_{rk}$, $V = \sum_{r=1}^{R}\sum_{k=1}^{K} V_{rk}$ and $\overline{V} = \sum_{r=1}^{R}\sum_{k=1}^{K} \overline{V}_{rk}$. Thus, the data structure is summarized as followings : subjects in $V + \overline{V}$ have data structure as $\{Y_i, W_i\}$, and subjects in V as $\{Y_i, X_i, W_i\}$. The likelihood from the population dataset is written as

$$L(\beta) = \prod_{r=1}^{R}\prod_{k=1}^{K}\prod_{i\in V_{rk}} P_\beta(Y_i|X_i,Z_i)g(X_i|Z_i,W_i) \times \prod_{r=1}^{R}\prod_{k=1}^{K}\prod_{i\in \overline{V}_{rk}} P_\beta(Y_j|Z_j,W_j)$$

where $P_\beta(Y_j|W_j) = \int P_\beta(Y_j|x,W_j)dG(x|W_j)$. Since $g(X_i|W_i)$ wasn't specified with any parametric form, a nonparametric estimator for $g(X_i|W_i)$ is considered in a nonparametric way.

Note that $G(X|W) = \sum_s \sum_l \pi_{sl}(W) \times G_{sl}(X|W)$ where $\pi_{sl}(W) = P(Y = l, C = s|W)$ and $G_{sl}(X|W) = G(X|W, Y = l, C = s)$. Wang and Zhou (2010) used a kernel approach for the case of a continuous W and an empirical method for the case of a discrete W. By substituting $\pi_{sl}(W)$, $G_{sl}(X|W)$ with $\hat{\pi}_{sl}(W), \hat{G}_{sl}(X|W)$ into $L(\beta)$ and using the Newton-Raphson algorithm, $\hat{\beta}$ can be obtained and asymptotic properties of this method were developed as well. This method could obtain efficiency gain by incorporating auxiliary information. Especially, they used the kernel density estimation for a continuous auxiliary variable and the empirical method for a discrete auxiliary variable case. Thus, Wang and Zhou (2010)'s method could have smaller standard error than compared methods.

1.2.2 ODS design for continuous response variable

An outcome-dependent sampling (ODS) design is a retrospective sampling scheme and a branch of stratified sampling. When one observes the covariates of interest with a probability that depends on the observed value of the outcome variable, one can draw samples with ODS scheme. From ODS, researchers can gain efficiency in estimating parameters than efficiency from simple random sampling design and Horvitz-Thompson type methods. In this section, we outline outcome-dependent sampling (ODS) design and outcome-auxiliary dependent sampling (OADS) design for a continuos response variable.

ODS design with continuous outcome variable

ODS for continuous outcome variable was proposed by Zhou et al. (2002) with the semiparametric empirical likelihood method for estimating parameters. Suppose that the response variable, Y, can be partitioned $-\infty = a_0 < a_1 < \cdots < a_{K-1} < a_K = \infty$ where $a_i's$ are fixed constant. Let $C_k = (a_{k-1}, a_k], k = 1, \cdots, K$. Note that ODS consists of SRS component and supplementary components. For SRS components, researchers draw n_0 sample from a population dataset from the simple random sampling scheme. To draw supplemental ODS samples from the *kth* stratum, we draw n_k of samples with the given interval, $\{C_k\}$. Hence, the

total sample size of ODS is $n = n_0 + \sum_{k=1}^{K} n_k$. Data structure of ODS is described as follows : SRS,

$$\{Y_{0i}, X_{0i}\}, i = 1, \cdots, n_0;$$

Supplemental ODS,

$$\{Y_{kj}, X_{kj} | Y_{jk} \in C_k\}, j = 1, \cdots, n_k, k = 1, \cdots, K.$$

With the ODS sample, the likelihood function of ODS dataset is written as

$$\begin{aligned} L(\theta, G_X) &= \left\{ \prod_{i=1}^{n_0} f(Y_i|X_i;\theta) g_X(x_i) \right\} \left\{ \prod_{k=1}^{K} \prod_{i \in C_k} f(Y_i, X_i | Y_i \in C_k; \theta) \right\} \\ &= \left\{ \prod_{i=1}^{n_0} f_\theta(y_{0i}|x_{0i}) \times \prod_{k=1}^{K} \prod_{j=0}^{n_k} \frac{f_\theta(y_{kj}|x_{kj})}{F(a_k|x_{kj}) - F(a_{k-1}|x_{kj})} \right\} \\ &\quad \times \left\{ \prod_{i=1}^{n_0} g_X(x_i) \times \prod_{k=1}^{K} \prod_{j=0}^{n_k} \frac{F(a_k|x_{kj}) - F(a_{k-1}|x_{kj})}{F(a_k|x_{kj}) - F(a_{k-1}|x_{kj})} \right\} \\ &= L_1(\theta) \times L_2(\theta, G_X). \end{aligned}$$

In the likelihood, no parametric model is specified for the marginal distribution of X. Thus, $G_X(X)$ has infinite dimensions from the perspective of the semiparametric literature. Qin and Lawless (1994) proposed the semiparametric empirical likelihood method for the estimating equations problem by extending Owen (1988) and Owen (2001). Zhou et al. (2002) developed a maximum semiparametric empirical likelihood estimator (hereafter, MSELE) for ODS design under continuos response. By fixing θ, $L(\theta, G_X)$ is profiled with the following constraints :

$$\left\{ p_i \geq 0, \sum_{i \in V} p_i = 1, \sum_{i \in V} p_i \{P_k(X_i;\theta) - \pi_k(\theta)\} = 0 \quad \text{for} \quad k = 1, \cdots, K \right\},$$

where $p_i = g_X(x_i)$ and $\pi_k = P(Y \in C_k)$. By using the the Lagrange multiplier method, p_i is estimated as $\hat{p}_i = \frac{1}{n} \frac{1}{1+\sum_{k=1}^{K} \lambda_k \{F(a_1|x_i) - \pi_k\}}$ where $\{\lambda_k\}$ are Lagrangian multipliers. After profiling out the $L(\theta, G_X)$, parameters of interest is estimated using the Newton-Rhapson method; the asymptotic results of the MSELE were given Zhou et al. (2002). Their approach could enhance efficiency from drawing oversamples from certain segments of a population that are believed to have more information. From the simulation study, MSELE in Zhou et al. (2002) has smaller SE than other compared methods.

Partial Linear Model approach for ODS design

Zhou and et al. (2011) proposed a partial linear model (PLM) under an ODS setting. Instead of the linear model in Zhou et al. (2002) described in the previous section, they considered a Partial Linear Model (PLM) as follows :

$$E(Y|X, z) = g(X) + Z^{'}\gamma, \quad \text{where} \quad g(\cdot) \quad \text{is an unknown smooth function.}$$

The ODS sampling scheme is the same as one in Zhou et al. (2002). With the total number of ODS sample size at $n = \sum_{k=0}^{K} n_k$ and the data structure for the ODS design at $\{y_{kj}, x_{kj}, z_{kj}\}$ for $k = 0, \cdots, K, \quad j = 1, \cdots, n_k$ where $k = 0$ means the SRS sample. Following Yu and Ruppert (2002), P-spline was used to estimate $g(\cdot)$. Technically, under the assumption that $g(\cdot)$ is a r-degree spline function with T fixed knots $t_1, \cdots, t_T$, then we have $g(x) = M^T(x)\alpha$ where $M(x) = \{1, x, x^2, \cdots, x^r, (x - t_1)_+^r, \cdots, (x - t_T)_+^r\}$ is an r-degree truncated power spline basis with knots $\{t_i\}_{i=1}^{T}$, $(x)_+^r = x^r 1_{x \geq 0}$ and α is a $r + T + 1$ dimensional vector. Hence, $E(Y|x, Z) = g(X) + Z^T\gamma = M^T(X)\alpha + Z^T\gamma = D^T\theta$ where $D = \{M^T(X), Z^T\}^T$ and $\theta = (\alpha^T, \gamma^T)^T$. Because of this P-spline, one should incorporate the penalty when the likelihood is constructed. Other processes are similar to the processes in Zhou et al. (2002). In particular, novelty of this method is founded in PCB study analysis. In previous ODS studies that have been conducted assuming the linear model, $E(Y|X, Z) = \beta X$, didn't find relationship between IQ and PCB. However, in

Zhou, You, Qin and Longnecker (2011) and Qin and Zhou (2010), a relationship between IQ and PCB was detected by this approach through the PLS appraoch.

Including non-validation dataset

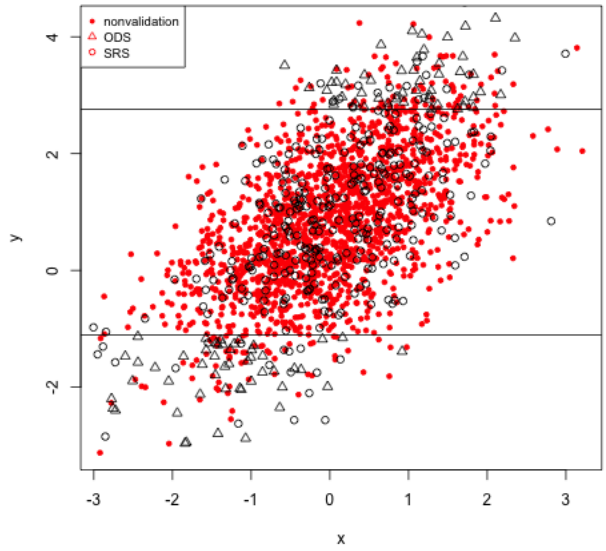

Figure 1.1: Illustration for the two-stage ODS with continuous outcome

Weaver and Zhou (2005) developed a two-phase ODS design that consists of complete ODS samples and incomplete samples in a cohort study. Borrowing terms from the measurement error study, V is defined as the validation dataset that consists of complete observations and $\overline{V}$ as the non-validation dataset that has missing covariates. Weaver and Zhou (2005) proposed methods that include non-validation observations into the validation dataset. Figure 1.1 provides a graphical a graphical intuition about ODS in a cohort study. In Figure 1.1, empty circles and empty triangles are complete observations that are drawn under an ODS scheme. The rest of the observations in a cohort study are presented as colored circles. By incorporating the non-validation dataset, efficient estimates can be achieved through a population dataset. Let the SRS and supplemental ODS components from the previous section be referred to as the validation dataset. From the population dataset, the dataset that is not sampled by ODS design is assumed

as non-validation set, $\overline{V}$. Note that all of the response variable values are known but the values of covariate of interest, X, can be obtained only for ODS samples. That is, the i_{th} observation in V, has the form of (Y_i, X_i) or $(Y_i, X_i|Y_i \in C_k)$, and the j_{th} observation in $\overline{V}$ has Y_j only. The likelihood for the validation set can be written as

$$L_V(\theta, G_X) = \prod_{i \in V} f(Y_i|X_i, \theta) \prod g_X(X_i) \prod_{k=1}^{K} \pi_k(\theta, G_x)^{-n_k}, \tag{1.1}$$

where $\pi_k(\theta, G_x) = \int_X P_k(x : \theta) dG_X(x)$ and $P_k(x : \theta) = \int_{C_k} f(y|x : \theta) dy$. $P_k(X : \theta)$ and $\pi_k(\theta, G_x)$ are the conditional and marginal probabilities that Y is in the k_{th} stratum and n_k denotes the size of supplemental ODS samples from the k_{th} stratum. Recall that N denotes the size of the study population, N_k is used for the size of the study population in the *k*th stratum, $n_{0,k}$ means the size of observations in n_0 that belongs in the *k*th stratum, and n_k is the size of *k*th supplemental ODS sample. At the second stage, along with the SRS samples, the supplemental ODS samples are drawn within strata at the two tails with the idea that these combinations of the extreme values of Y contain more information for the relationship of interest between outcome Y and covariate X than the simple random sampling. Define $n_{\overline{V},k} = N_k - n_{0,k} - n_k$ as the size of the non-validation sample in the kth stratum. Note that N_k, $n_{0,k}$ and $n_{\overline{V}_k}$ are random variables but n_k is fixed for each k. Thus, the distribution of $\{n_{\overline{V},k}\}$ is the same with $\{N_k - n_{0,k}\}$. Thus, $\{n_{\overline{V}_k}\}$ follows a multinomial law

$$P(\{n_{\overline{V}_k}\}) = \frac{(N - n_0)!}{\prod_{k=1}^{K}(N_k - n_{0,k})!} \prod \{\pi_k(\theta, G_X)\}^{N_k - n_{0,k}}. \tag{1.2}$$

Thus, combining (1.1) and (1.2), we can derive the likelihood function which is proportional to

$$L_{full} = \prod_{i \in V} f(Y_i|X_i, \theta) \prod_{i \in V} dG_X(X_i) \prod_{j \in \overline{V}} \int_X f(Y_j : \theta). \tag{1.3}$$

Weaver and Zhou (2005) developed the estimated likelihood approach to ODS sampling under a cohort study that has a continuous response variable. Recalling the likelihood (1.3),

estimating $G_X(X)$ and replacing $G_X(x)$ with $\hat{G}_X(X)$ are key steps to estimate parameters of interest. By the probability law, $G_X(x)$ could be written as

$$G_X(x) = P(X \leq x) = \sum_{k=1}^{K} = Pr(Y \in C_k)Pr(X \leq x|Y \in C_k).$$

$G_X(x)$ was estimated by $\hat{G}_X(x) = \sum_{k=1}^{K} \frac{N_k}{N} \sum_{i \in V_k} \frac{I\{X_i \leq x\}}{n_k + n_{0k}} = \sum_{k=1}^{K} \frac{N_k}{N(n_k+n_{0k})} \sum_{i \in V_k} I\{X_i \leq x\}$ as the empirical cumulative distribution, and, by plugging it into the likelihood (1.3) and using the Newton-Rhapson method, the consistent and unbiased estimates could be obtained. This method could incorporate non-validation sample into full information and it brings efficiency gain. In Chapter 3, we extend the two-stage ODS design in Weaver and Zhou (2005) to a two-stage ODS design including an auxiliary variable.

Two-phase OADS design with continuous outcome

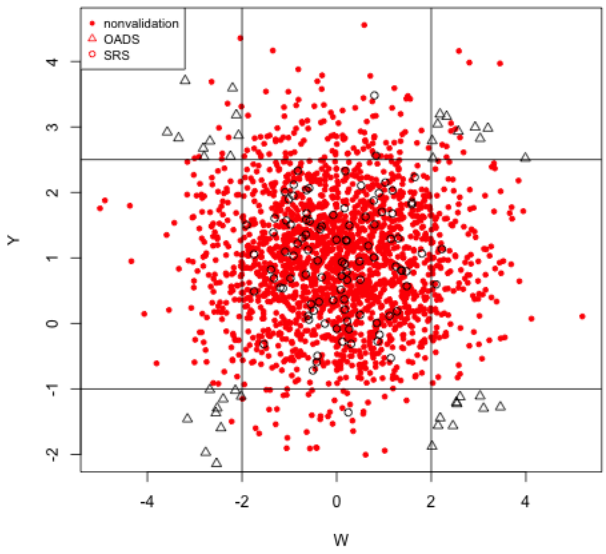

Figure 1.2: Illustration for the two-stage OADS with continuous outcome

As an extension of two-phase OADS design for a discrete outcome variable developed in Wang and Zhou (2010) to a continuos outcome response variable, Zhou, Wu, Liu and Cai

(2011) developed the two-phase OADS design for continuous outcome variable. In the two-stage OADS with continuous response variable, the key idea is to gain efficiency by extracting more information from the OADS design than in the ODS design. The OADS uses categorization of variables not only with the response variable Y but also with the auxiliary variable W as well. X is the measurement which can be observed only for the subjects in the OADS subsample and Z is the vector of all covariates that are observed for all subjects in the study cohort. To fix notation and data structure, let Y be a continuous outcome variable, (Z, X) be a covariate vector, and W be a continuous auxiliary variable for X. Suppose that Y can be partitioned into J mutually exclusive strata by know constants $-\infty = a_0 < a_1 < \cdots < a_{J-1} < a_J = \infty$ and W also can be partitioned into T mutually exclusive strata by know constants $-\infty = b_0 < b_1 < \cdots < b_{T-1} < b_T = \infty$. Then, we have a product domain of $Y \times W$ that can be partitioned into $J \times T$ mutually exclusive rectangles $A_j \times B_t$, for $j = 1, \cdots, J$ and $t = 1, \cdots, T$. For simple notation, we write these rectangles as $\{\Delta_k = A_j \times B_t : k = 1, \cdots, K, \quad$ for each combination of j and t, separately $\}$. From this setting, we describe how to draw a two-stage sample as the following steps : 1. At the first stage, N subjects are drawn at random from a population dataset, 2. At the second stage, two-component OADS sampling design is applied. That is, (1) SRS sample of size n_0 and (2) supplemental OADS sample of size n_k from the k_{th} stratum for $k = 1, \cdots, K$. Thus, the observed data structure for the two-stage OADS design with continuous outcome can be summarized as follows : the first stage : $\{Y_i, Z_i, W_i\}$ for $i = 1, \cdots, N$ where N is the population size; a SRS sample of size n_0 at the second stage : $\{Y_i, X_i, W_i\}$, $i = 1, \cdots, n_0$; a supplementary OADS sample of size n_k from the *k*th stratum in the second stage : $\{Y_i, X_i, W_i | (Y_i, W_i) \in \Delta_k\}$, $i = 1, \cdots, n_k$. To describe the two-phase OADS design, we present Figure 1.2 wfor $J = T = 3$. At the second stage, the supplemental OADS samples are drawn within strata at the four corners with the idea that these combinations of the extreme values of Y and W contain more information about the relationship of interest between outcome Y and covariate X than one with the simple random sampling. Details about the data structure of OADS are given in Zhou, Wu, Liu and Cai (2011). The full likelihood based on the observations under the two-stage OADS design can be derived

as

$$L(\beta) = \prod_{k=0}^{K} \prod_{i \in \widetilde{V}_k} f(Y_i|Z_i, X_i; \beta) g(X_i|Z_i, W_i) \prod_{k=1}^{K} \prod_{i \in \overline{V}_k} \int_X f(Y_i|Z_i, x; \beta) dG(x|Z_i, W_i)$$

where $\widetilde{V}_k$ represent the supplemental OADS samples in the stratum Δ_k and $\overline{V}_k$ represents non-validation set in the stratum Δ_k. By exploiting an estimated likelihood method, to account for the two-stage OADS mechanism, $G(x|w) = \sum_{k=1}^{K} \pi_k(w) G_k(x|w)$ where $\pi_k(w) = P\{(Y, W) \in \Delta_k | w\}$ and $G_k(x|w) = G(x|(Y, W) \in \Delta_k)$. Hence, $\pi_k(w)$ and $G_k(x|w)$ can be estimated by $\hat{\pi}_k(w) = \frac{\sum_{i=1}^{N} I((Y_i, W_i) \in \Delta_k) \phi_{h_N}(W_i - w)}{\sum_{i=1}^{N} \phi_{h_N}(W_i - w)}$ and $\hat{G}_k(x|w) = \frac{\sum_{i \in V_k} I((Y_i, W_i) \in \Delta_k) \phi_{h_N}(W_i - w)}{\sum_{i \in V_k} \phi_{h_N}(W_i - w)}$ where $\phi_{h_N}(\cdot) = \phi(\frac{\cdot}{h_N})$ is a d-dimensional kernel function with the bandwidth h_N. By replacing $\pi_k(w)$ and $G_k(x|w)$ in the full likelihood with $\hat{\pi}_k(w)$ and $\hat{G}_k(x|w)$ and using Newton-Rhapson method, one can estimate β. This 2-stage OADS design suggests that greater efficiency could be gained over other estimators compared. However, when the dimension of auxiliary variables is moderately large, this method would not work well due to the curse of high dimensionality.

Two-stage Probability-Dependent Sampling scheme

Zhou et al. (2014) proposed the Probability Dependent Sampling (PDS) in two-phase study. It can be described as follows. Let Y be the continuous response variable, X be the covariate of interest, and Z be other covariates. At the first phase, an SRS sample is drawn and they have data structure as (Y, X, Z). Before drawing samples at the second phase, by using observations in the first phase, we fit a model for $E(X|Y, Z)$. Without a loss of generality, assume the domain of covariate, X, can be partitioned into three mutually exclusive intervals : $(-\infty, x_L] \bigcup (x_L, x_U] \bigcup (x_U, \infty)$ where x_L and x_U are some fixed constant to partition the domain of X. The chances of that a new subject's X will be in $(-\infty, x_L]$ and (x_U, ∞) can be predicted by $\hat{\phi}_1(y, z) = \hat{P}(X \leq x_L|Y, Z)$ and $\hat{\phi}_3(y, z) = \hat{P}(X \geq x_U|Y, Z)$, respectively. Then, supplemental samples are drawn at the second phase by obtaining an SRS from those who are

likely to have high or low X values. To be more specific, we describe the procedure for the PDS scheme as following steps : (1) At the first stage, the size of n_0 samples are drawn by SRS scheme; (2) A model of $E(X|Y,Z)$ is fitted with n_0 of SRS samples; (3) $\phi_1(y,z) = P(X \leq x_L|Y,Z)$ and $\phi_3(y,z) = P(X \geq x_U|Y,Z)$ are estimated; (4) At the second stage, draw supplemental random samples from those in the study population whose predicted probability ϕ_1 and ϕ_3 are higher than some fixed criteria, eg. 80%, with fixed sizes, n_1 and n_3. Thus, the data structure for the PDS design is as follows : SRS samples, $\{Y_{0i}, X_{0i}, Z_{0i}\}$, $i = 1, \cdots, n_0$, supplemental PDS samples, $\{(Y_{1i}, X_{1i}, Z_{1i}) : P(X_{1i} \in A_1|Y_{1i}, Z_{1i}) \geq c_1\}$, for $i = 1, \cdots, n_1$ and $\{(Y_{3i}, X_{3i}, Z_{3i}) : P(X_{3i} \in A_3|Y_{1i}, Z_{1i}) \geq c_3\}$, for $i = 1, \cdots, n_3$. Zhou et al. (2014) introduced several ways of estimating ϕ_1 and ϕ_3 according to assumed model between X and (Y,Z). Assuming that $f_\beta(Y|X,Z)$ is $Y = \beta_0 + \beta_1 X + \beta_2 X + e$ where $e \sim N(0, \sigma_e^2)$, let $G(X,Z)$ and $g(x,z)$ be the joint cumulative and probability distribution function, respectively. According to the PDS design described above, the likelihood function for PDS samples is represented as

$$L(\beta, G) = \left\{ \prod_{i=1}^{n_0} f_\beta(Y_{0i}|x_{0i}, Z_{0i}) g(X_{0i}, Z_{0i}) \right\} \prod_{k=1,3} \prod_{j=1}^{n_k} f_\beta\{Y_{kj}, X_{kj}, Z_{kj} | \phi_k(Y_{kj}, Z_{kj}) \geq c_k\}$$

where c_k is pre-fixed constant in $(0,1)$. From the likelihood function, we obtain the log-likelihood,

$$\begin{aligned} l(\beta, \{p_i\}, \{\pi_k\}) &= \sum_{i=1}^{n} log\{f_\beta(Y_i|X_i, Z_i)\} + \sum_{i=1}^{n} log(p_i) - \sum_{k=1,3} n_k log_(\pi_k) \\ &= l_1(\beta) + l_2(\{p_k\}, \{\pi_k\}), \end{aligned}$$

where $p_k = g(X_i, Z_i)$, $l_1(\beta) = \sum_{i=1}^{n} log\{f_\beta(Y_i|X_i, Z_i)\}$ and $l_2(\{p_k\}, \{\pi_k\}) = \sum_{i=1}^{n} log(p_i) - \sum_{k=1,3} n_k log_(\pi_k)$. With fixed (β, π_1, π_3) and obtaining the empirical likelihood function of p_i over all distributions whose support contains the observed values of X and Z, one can derive the profile likelihood in the same way with the given constraints in Zhou et al. (2014). After

profiling $l_2(\{p_k\}, \{\pi_k\})$ above, estimate of $(\beta', \{\pi_k\}')$ can be obtained by Newton-Rhapson method. Zhou et al. (2014) also described asymptotic properties and simulation study results showing that PDS performs better than compared estimators in terms of efficiency. This method allows for a continuous variable and a vector of available covariates to be exploited in selecting a more informative second-phase data set. In simulation study, it is shown that PDS method is more efficient than other competitors : IPW method for two-phase design and SRS design with the same size of sample.

Pepe and Fleming (1991) proposed an estimated likelihood approach for binary outcome variable under SRS design with categorical auxiliary variable. Carroll and Wand (1991) also developed an estimated likelihood method for estimating parameter of interest in a logistic regression model for the continuous auxiliary variable for X. Zhou and Pepe (1995) proposed an estimated likelihood to estimate the induced relative risk functions, which were proposed by Prentice (1982), using the validation sample. Chatterjee et al. (2003) proposed the pseudoscore method that used the postulated parametric regression model to improve the efficiency of estimates. To implement their method to ODS problem in Weaver and Zhou (2005), by using the observed sampling fractions, substitution for $\frac{N_k}{n_k+n_{0k}}$ with $q_\theta(X_i) = \sum_{k=1}^{K} \frac{N_k}{n_k+n_{0k}} P_k(X_i;\theta)$ where $P_k(X;\theta)$
$= \int_{C_k} f(y|x;\theta)dy$ is needed. Song et al. (2009) proposed the restricted maximum likelihood estimator under the two-phase ODS design in Weaver and Zhou (2005). Along with the likelihood function, (1.5), they considered the Lagrangian Multiplier :

$$H(\theta, g_i, \theta) = \sum_{i\in V} log f(y_i|x_i\theta) + \sum_{j\in\overline{V}} log\{\sum_{i\in V} g_i f(y_j|x_i;\theta)\} - \lambda(\sum_{i\in V} g_i - 1),$$

where λ is the Lagrangian multiplier. By taking derivatives $\{g_i\}$, $\{\hat{g}_i\}$ were estimated as $\hat{g}_i = \left\{ n - \sum_{j\in\overline{V}} \frac{f(y_j|x_i;\theta)}{\sum_{k\in V} \hat{g}_k f(y_j|x_k;\theta)} \right\}^{-1}$. Different from the two-component ODS design in Zhou et al. (2002), it concerns two-stage ODS design, the number of constraints increases as the sample size increases. Thus, Song et al. (2009) proposed the mixed Newton method to circumvent the

difficulty of estimation with large dimensions of parameters, and also provided the asymptotic properties of the proposed estimator and simulation results with different cutoff points and supplementary sample proportions to show efficiency gains.

1.3 Missing data and Measurement error problems

Missing data can exist over a variety of epidemiological studies. In sample surveys, when researchers send out questionnaires, non-responses could be obtained. In clinical trials, some subjects drop out of the study at certain points or stop taking their treatments. For some studies, covariates of interest may be very expensive or difficult to obtain. For example, if researchers are interested in observing daily fat intake of subjects over a long period, it will be difficult to track down fat intake for every subjects in the study. Instead of providing an accurate fat intake, it is possible for study participants to recall the food that they consume each day. In this case, the recorded food intake can be used to measure fat intake roughly, and that information is called surrogate. Thus, in most of studies involving human subjects, some important covariates may be missed or measured from a limited number of subjects. Alternatively, surrogate information can be obtained for all subjects and it creates an advantage to enhance the efficiency of estimates obtained from studies.

1.3.1 Missing data mechanism

Little and Rubin (2014) classified missing mechanisms into Missing Completely at Random (MCAR), Missing at Random (MAR), and Non-Missing at Random (NMAR). Data is said to be missing completely at random if the failure to observe a value does not depend on any data, either observed or missing. Data is said to be missing at random if the failure to observe a value does not depend on the data which are unobserved. However, the missingness may depend on any observed data. The missing mechanism is said to be non-ignorable if the failure to observe a value depends on the value that would have been observed. This dissertation concerns only MCAR and MAR. Consider a simple logistic regression model with observations (y_i, x_i), where

y_i is a response, and x_i is the corresponding covariate.

$$P(y_i = 1|x_i; \beta) = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}.$$

Assume that there is no missingness in the response variable, Y, and some of $\{x_i\}$ are missing. Define $r_i = 1$ as observed x_i, and $r_i = 0$ otherwise. Then, r_i is a Bernoulli random variable and can be modeled as a function of the covariate and/or the response variable. For example, we can take $P(r_i = 1|\eta) = \frac{exp(\eta_0)}{1+exp(\eta_0)}$ where $\eta = \eta_0$. The missing data mechanism is MCAR because the failure to observe a value does not depend on any data. If we consider $P(r_i = 1|\eta) = \frac{exp(\eta_0+\eta_1 y_i)}{1+exp(\eta_0+\eta_1 y_i)}$ where $\eta = (\eta_0, \eta_1)'$. In this case, the missing data mechanism is MAR since the failure to observe a value does not depend on the data which are unobserved. However, the missingness may depend on any observed data. For the last case, we consider $P(r_i = 1|\eta) = \frac{exp(\eta_0+\eta_1 y_i+\eta_2 x_i)}{1+exp(\eta_0+\eta_1 y_i+\eta_2 x_i)}$ where $\eta = (\eta_0, \eta_1, \eta_2)'$ and it has NMAR mechanism.

1.3.2 Existing methods using auxiliary information

Let Y_i be a response variable, X_i be a covariate of interest, W_i be an auxiliary variable for X_i, and R_i is a missing indicator variable. Assuming that Y_i and W_i are observed for all N observations in a population. The data structure in missing data can be classified as two different data forms of either (Y_i, X_i, W_i, R_i) for complete observations, or (Y_i, W_i, R_i) for incomplete observations, respectively. In the context of measurement error study, we borrow the terminologies, validation sample for the data that has the form of $(Y_i, W_i, R_i = 0)$, and invalidation sample for the data that has the form of $(Y_i, W_i, R_i = 1)$, respectively. Furthermore, in ODS designs, the probability that any individual in the population is selected into the validation sample does not depend on X, that is, the probability of that the i_{th} observation falls into k_{th} strata validation sample set is defined as $\tilde{p}_i = n_k/N_k$. Thus, one can assume that X in the non-validation set is MAR(Missing At Random). Now, a likelihood function for N population dataset along with

the MAR assumption is proportional to

$$\prod_{i=1}^{N} [f(Y_i|X_i)]^{R_i} \times \left[\int f(Y_i|W_i, x; \theta) dG_{X|W}(x|W_i) \right]^{1-R_i}. \tag{1.4}$$

Note that (1.4) cannot be solved by the parametric EM algorithm Dempster et al. (1977) unless a parametric assumption is given on $X|W$ and W. To circumvent this difficulty, Pepe and Fleming (1991), and, Carroll and Wand (1991) developed estimated likelihood methods for binary response variable. Reilly and Pepe (1995) proposed the mean score method for the binary response case. Furthermore, Weaver and Zhou (2005) developed an estimated likelihood method for the continuous outcome regression model under ODS sampling designs in a cohort study. Chen and Chen (2000) proposed a unified approach for two-stage design that includes validation set and non-validation set with auxiliary information for all observations under the SRS scheme. To give a mathematical meaning to a surrogate variable, it is important to understand the difference between differential and non-differential measurement error. When W dosen't have any information about response variable, Y, other than what is available in X and Z, there is non-differential error. In mathematical notation, if $P(Y|X = x, W = w) = P(Y|X = x)$, then W is defined as a surrogate variable. Unless measurement error is non-differential, measurement error is differential. Carroll et al. (1993) developed the pseudo likelihood method to incorporate differential and non-differential error together. For some cases, there would exist coarsened data, which consists of (Y, W, Z) without any exact observation of covariate of interest, X. Tsiatis and Ma (2004) developed a semiparametric approach by using efficient score function under coarsened data.

The measurement error problem that might be interesting in a cohort study surrounds how to use the surrogate information for missing covariates to obtain more efficient estimators. To be more specific, let Y be the response variable, X be the covariates of interest, which is expensive or hard to observe for every study subject, and Z represent another predictor which is measured for all subjects without error. Additionally, we can observe W that is related to X and call it

the auxiliary variable for X. Two types of auxiliary variables are considered in this dissertation. One is $W = I(X + \varepsilon > c)$ where c is some constant and another is $W = X + \varepsilon$ where $\varepsilon \sim (0, \sigma_W^2)$ for some distribution. Since excluding incomplete covariate observations can result in a loss of efficiency, using auxiliary information from a study population would lead to more efficient results. To handle missing covariates, we consider the measurement error approach that uses auxiliary information for missing covariates. We will propose a method that uses binary surrogate information under ODS design and a continuous auxiliary information in a two-stage ODS/OADS design.

As mentioned earlier, we focus on measurement error problems with auxiliary information. Zhou et al. (2002) did not address missing data problem in the ODS design. However, it is likely that researchers are confronted with a variety of missing data problems in ODS. When one draw an SRS sample or supplemental ODS sample, observations from SRS or ODS can have missingness in covariates. Along with incomplete observations, suppose that auxiliary information, W, for covariate,X, is available over all subjects in the whole sample. Let V be the set which consists of complete data and $\overline{V}$ is the set which consists of incomplete data. Observations in missing data with auxiliary information have data structures as follows : Complete observations, $\{Y_i, X_i, W_i\}$ for $i \in V$; incomplete observations, $\{Y_j, W_j\}$ for $j \in \overline{V}$. The underlying data structure explained above would vary according to sampling schemes. Likelihood functions for different study designs would be derived differently as well. In this dissertation, we focus on different measurement error problems according to different ODS/OADS designs : One is missing data with a discrete auxiliary variable; another is missing data with a continuous auxiliary variable under two-stage sampling design. In the following section, we review existing methods to solve measurement error problems, in particular, non-differential measurement error.

1.4 Well-known approaches to the measurement error problems

In this section, we review existing approaches to measurement error problems. In particular, the updating method for two-phase design in 1.4.4 will be used in our proposed method to enhance the efficiency of estimates in a two-stage ODS setting.

1.4.1 EM algorithm and Mean Score method

The EM algorithm in Dempster et al. (1977) is a general iterative algorithm that is used to find maximum likelihood estimates in incomplete data problems. EM is also most useful when maximization from the complete data likelihood is straightforward while maximization based on the observed data likelihood is difficult. Suppose that we are given with the likelihood function,

$$L(\beta,\theta) = \prod_{i\epsilon V} P_\beta(Y_i|X_i)P_\theta(X_i|Z_i)\prod_{j\epsilon\overline{V}} P_{\beta,\theta}(Y_j|Z_j),$$

where $P_{\beta,\theta}(Y|Z) = \int P_\beta(Y|X)P_\theta(X|Z)dX$, and V and $\overline{V}$ denote the complete dataset and the incomplete dataset, respectively. If $P_\theta(X|Z)$ were a completely known model, one could use the EM algorithm for finding the MLE by factoring $L(\beta,\theta)$ above and taking the log-transformation as

$$Q(\beta|\beta^c) = \sum_{i\epsilon V} logP_\beta(Y_i|x_i) + \sum_{j\epsilon\overline{V}} E\bigg(log\{P_\beta(Y_j|X)\}|\beta^c, Y_j, Z_j\bigg).$$

However, in the absence of knowing $P_\theta(X|Z)$, EM algorithm can not be applied directly. To resolve this situation, Reilly and Pepe (1995) proposed the mean score method that estimates $E\bigg[log\{P_\beta(Y_j|X)\}|\beta^c, Y_j, Z_j\bigg]$ as $\sum_{i\epsilon V^{Z_j,Y_j}} \frac{logP_\beta(Y_j|X_i)}{n^V_{Z_j,Y_j}}$ where V^{Z_j,Y_j} is the subset of the complete dataset with $Z = Z_j$and $Y = Y_j$. The mean score method is valid when Z and Y are discrete and complete observations are in each stratum defined by (Y,Z). The EM algorithm described in this section needs to specify an underlying parametric distribution assumption. Reilly and Pepe (1995) developed the mean score method to overcome this issue in the discrete response case.

However, there has not been a mean score method developed for the biased sampling and bias issues arise when the mean score method is used directly for ODS or OADS problems.

1.4.2 Probability weighted likelihoods method

The weighted likelihood function method has been considered useful for some complex sampling mechanisms. Horvitz and Thompson (1952) developed a method that assigns weights to account for a given biased-sampling under probability sampling designs. Zhao and Lipsitz (1992) developed similar methods for the two-stage studies that have a binary outcome. Holt et al. (1980) and Lawless et al. (1999) applied this approach to linear regression models for continuous outcomes with data obtained from a complex survey design. Weaver and Zhou (2005) described the weighted likelihood for ODS data as

$$l_W(\beta) = \sum_{k=1}^{K} \frac{1}{\tilde{p}_k} \sum_{i \epsilon V_k} \ln f(Y_i|X_i;\beta) \tag{1.5}$$

where $\tilde{p}_k$ is the selection probability for all individuals in the k_{th} strata. The $\tilde{p}_k$ can be estimated for different data structures, respectively, as a solution to the weighted score equations,

$$S_W(\beta) = \sum_{k=1}^{K} \frac{1}{\tilde{p}_k} \sum_{i \epsilon V_k} \frac{\partial \ln f(Y_i|X_i;\beta)}{\partial \beta}.$$

Although the probability weighted estimators are generally unbiased, they can be inefficient compared to the other estimators. To be specific, when the outcomes are oversampled from tails of a study distribution, because the most weight would be given to the intermediate, less informative measurements would be obtained from such a sample. Despite such inefficiencies, it is widely used, since it is easy to implement and provides unbiased estimates with a properly stratified population.

1.4.3 General Unbiased Estimating Functions

We say the estimating equations are unbiased if

$$E_\theta\left\{\sum_{i=1}^{N} g_i(Y_i, X_i, W_i)\right\} = 0, \quad \text{for all} \quad \theta.$$

To drag the concept of the unbiased estimating equation into semiparametric approaches, Robins et al. (1994) proposed the unbiased estimating equations based on inverse probability weighted method. Define π_i as either a known or parameterized probability function satisfying $\pi(W) = P(R = 1|W) = P(R = 1|W, X)$ so that X is missing at random (MAR) in the sense of Little and Rubin (2014). Let β be $p \times 1$ parameter vector and $G\Big(\beta, h(X, W)\Big)$ be any unbiased estimating function. Then, the function

$$\frac{R \times G\{\beta, h(X, W)\}}{\pi(W)} - \frac{\{R - \pi(W)\} \times \phi(Y, Z)}{\pi(W)},$$

where h is q by t function and $\phi(Y, Z)$ is an arbitrary $q \times 1$ function, satisfying $E(\phi(Y, Z, W)\phi(Y, Z, W)') < \infty$. is a estimating equation function in the class of the unbiased estimating functions. In the class of unbiased estimating functions in Robins et al. (1994), it was shown that, for the fixed h, the asymptotic variance of $\hat{\beta}$ is minimized at $\phi(Y, Z, W) = E\Big[G(\beta, h(X, W))|Y, Z, W\Big]$. Robins et al. (1994) proposed a class of semiparametric estimators when the data are missing at random assuming the missingness is either known or parametrically estimated. From a theoritical view, they demonstrated that their proposed method is the most efficient estimator and asymptotically normal in the class that they defined. Moreover, it could be applied to the case in which both response and covariates are missing. However, when response is continuous, computation is challenging. Tsiatis and Ma (2004) developed the most efficient estimator under the coarsened dataset and MAR assumption but only logistic regression simulation was conducted and biased sampling was not considered.

1.4.4 Updating method for two-phase design

In many observational studies, it is a main interest for researchers to investigate a relationship between response variable and covariate(s) of interest. However, due to budget limitations, researchers could be confronted with a missing in covariates problem. In this case, a cheap surrogate variable could be considered considered in order to use more information to complement information about covariate(s). There have been several research initiatives that have used auxiliary information to address the missing in covariates problem. Pepe and Fleming (1991) and Carroll and Wand (1991) developed an estimated likelihood method for a discrete and continuous auxiliary variable, respectively. Robins et al. (1994) proposed an estimator in a general class of estimating equations for incomplete covariate data but it proved computational challenging, in particular, in cases of continuous response. Chen and Chen (2000) proposed a unified approach under two-stage sampling design. Since this approach updates the estimates from complete samples by using auxiliary information from sample at the first stage, we call it as "updating method". Jiang and Haibo (2007) extended this method to the additive hazard model for updating pseudoscore estimation by using information from all data available when some of the true covariates are measured only on a randomly selected validation set, whereas auxiliary covariates are observed for all study subjects.

To fix notation, let Y be a response variable, X be a covariate vector and W is an auxiliary variable that is a proxy measure measure of X. To see how it works, the data structure of this method is described as follows : observations from the first-stage, $\{Y_i, W_i\},\ i = 1, \cdots, N$; observations from the second-phase, $\{Y_i, X_i, W_i\},\ i = 1, \cdots, n$. Note that samples from second stage are drawn using the SRS scheme from the primary samples in first stage.

The underlying regression model for the conditional mean of Y given X is defined by $E(Y|X) = g(x_i; \beta)$ where $g(\cdot)$ is a known function and β is a vector of unknown regression parameters. Define $\hat{\beta}$ as the solution of an estimating equation, $0 = \sum_{i \epsilon V} S_i(\beta)$, where $S_i(\beta)$is the score function of i_{th} observation and V is the validation set that consists of complete observations, $\{Y_i, X_i, W_i\},\ i = 1, \cdots, n$. Since W contains information about the covariate, X,

we can also consider a similar underlying regression model Y given W, $E(Y|W) = h(w_i;\gamma)$, and accordingly, define $\hat{\gamma}$ as the solution of the estimating equation, $0 = \sum_{i\epsilon V} \widetilde{S}_i(\gamma)$. where $\widetilde{S}_i$ is a score function for i_{th} observation. Thus, two linear models are considered as follows : For (Y_i, X_i) in the validation set,

$$Y \quad = \quad X\beta + e_x \quad \text{for observations in a validation set, } V;$$

For (Y_i, W_i) in the validation set,

$$Y \quad = \quad W\gamma + e_w \quad \text{for observations in a validation set, } V,$$

where e_x and e_w are error terms that have mean zero and some variances.

Under regularity conditions, $\hat{\beta}$ and $\hat{\gamma}$ are consistent for (β^*, γ^*) that are true value of (β, γ). They derived a multivariate normal distribution of $\sqrt{n}\{(\hat{\beta} - \beta^*)^T, (\hat{\gamma} - \gamma^*)\}^T$,

$$\sqrt{n}\{(\hat{\beta} - \beta^*)^T, (\hat{\gamma} - \gamma^*)^T\}^T \sim N(0, \Sigma) \tag{1.6}$$

where $\Sigma = D^{-1}CD^{-1}$ and $D = diag(D_1, D_2)$ with $D_1 \equiv E\{\partial S(\beta^*)/\partial\beta\}$, $D_2 \equiv E\{\partial\widetilde{S}(\gamma^*)/\partial\gamma\}$, and $C \equiv E\Big[\{S(\beta^*)^T, \widetilde{S}(\gamma^*)^T\}^T, \{S(\beta^*)^T, \widetilde{S}(\gamma^*)^T\}\Big]$. Note that C can be expressed with its components, $C_{11} \equiv E\{S(\beta^*)S(\beta^*)^T\}$, $C_{12} \equiv E\{S(\beta^*)\widetilde{S}(\gamma^*)^T\}$, and $C_{22} \equiv E\{\widetilde{S}(\gamma^*)\widetilde{S}(\gamma^*)^T\}$. By the property of multivariate normal distribution theory, the conditional distribution of $\sqrt{n}(\hat{\beta} - \beta^*)'$ given $\sqrt{n}(\hat{\gamma} - \gamma^*)'$ is asymptotically normal $\sqrt{n}D^{-1}C_{12}D_2(\hat{\gamma} - \gamma^*)'$. To estimate γ^*, a linear model that investigates a relationship between Y and W,

$$Y = W\gamma + e$$

for all observations in a population dataset is considered. Let $\overline{\gamma}$ be the estimate of γ^* by solving a score equation, $0 = \sum_{i=1}^{N} \widetilde{S}_i(\gamma)$. Thus, by equating $\sqrt{n}(\hat{\beta} - \beta^*)'$ with its estimated conditional

mean, $\sqrt{n}\hat{D}_1^{-1}\hat{C}_{12}\hat{C}_{22}^{-1}\hat{D}_2(\hat{\gamma}-\overline{\gamma})$, and solving it with respect to β^*, the updated estimate, $\overline{\beta}$, is

$$\overline{\beta} = \hat{\beta} - D_1^{-1}C_{12}C_{22}D_2(\hat{\gamma}-\overline{\gamma}). \tag{1.7}$$

Chen and Chen (2000) also derived the large sample properties of $\overline{\beta}$. Along with estimate $\overline{\beta}$, the covariance of $\sqrt{n}(\overline{\beta}-\beta^*)$, is given by $\Omega = D_1^{-1}C_{11}D_1^{-1} - (1-\rho)D_1^{-1}C_{12}C_{12}^TD_1^{-1}$. One can observe that the first term of Ω is the asymptotical variance for $\hat{\beta}$. Thus, the updated estimator, $\overline{\beta}$, has smaller variance than that of $\hat{\beta}$.

The advantages of this method are that (1) it can deal with auxiliary information in a varied class of regression models and (2) it is computationally convenient. Compared to other methods used for two-stage design, the updating method can deal with multiple auxiliary surrogates for multiple covariates of interest. However, since their method did not account for the biased sampling, it could not be directly applied to biased sampling. In Chapter 3, we propose a method that combines the updating method under a two-stage ODS design.

After reviewing existing methods in the previous sections, we can see the pros and cons of each method when we estimate parameters of interest under ODS or OADS designs. In the next section we briefly preview three proposed methods that will be covered in Chapter 2, 3 and 4. We will propose three methods : (1) an estimated likelihood method to a missing in covariate in ODS design in Chapter 2; (2) an updating method in two-phase ODS design in Chapter 3; (3) an updating method in two-phase OADS design in Chapter 4. In all methods, we show how to use auxiliary information for a covariate of interest.

1.5 Preview of Proposed Research

1.5.1 An updating method with Auxiliary Information under two-phase Outcome Dependent Sampling

In Chapter 2, we consider a two-phase ODS design in a cohort study. A two-phase ODS sample consists of complete observations under ODS scheme in the second phase and

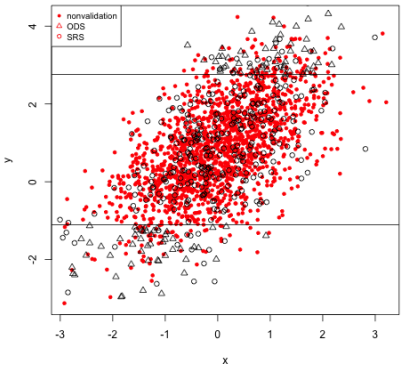

Figure 1.3: Illustration for a two-phase ODS under a linear regression model

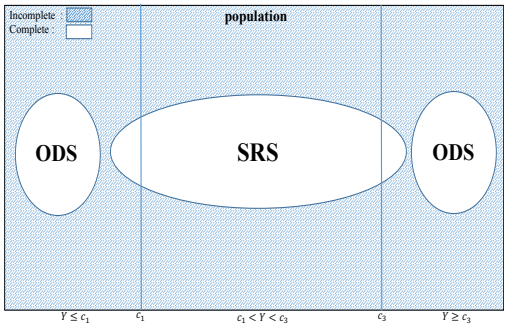

Figure 1.4: Conceptual illustration for a general two-phase ODS design

First stage :		$\{Y_i, W_i : i = 1, \cdots, N\}$;
Second stage :	SRS	$\{Y_i, X_i, W_i : i = 1, \cdots, n_0\}$;
	ODS from the left tail	$\{Y_i, X_i, W_i \vert Y_i \le c_1 : i = 1, \cdots, n_1\}$;
	ODS from the right tail	$\{Y_i, X_i, W_i \vert Y_i \ge c_3 : i = 1, \cdots, n_3\}$.

observations in the first phase. To fix notation let Y denote a continuous outcome variable, X be a covariate vector, and W be a proxy measure for X. Figure 1.3 shows the two-phase ODS design in a cohort study under a linear model. In terms of the type of auxiliary information, our proposed method considers a continuous auxiliary variable, W, for a covariate of interest while Weaver and Zhou (2005) considered a categorical auxiliary variable in their discussion section. We assume that there are independent and identically distributed population samples of size N in the first phase. The domain of Y consists of 3 mutually exclusive intervals : $C_1 \cup C_2 \cup C_3 = (-\infty, c_1] \cup (c_1, c_3] \cup (c_3, \infty)$ where c_1 and c_3 are fixed constants. In the second phase, the ODS sample of size n consists of three parts, SRS sample of size n_0, a supplemental ODS sample of size n_1 from C_1 and another supplemental ODS sample of size n_3 from C_3. Thus, a two-stage ODS design in our study has the data structure as follows : The ODS sample in the second phase is a complete sample but that the rest of the observations in the population are incomplete observations that have missing in covariate. From the measurement error terminology, V denotes the validation sample set and $\overline{V}$ denotes the nonvalidation sample set.

Let n_V be the total sample size of ODS that consists of complete observations, and $n_{\overline{V}} = N - n_V$, is the number of incomplete observations. $n_V = n_0 + n_1 + n_3$ where n_0 is the number of SRS sample and n_k denotes the number of supplemental ODS samples from the k_{th} interval. Figure 1.4 is depicted to give a graphical understanding of two-phase ODS design. The ellipses parts represent V of size n_V, and the shaded area represents $\overline{V}$ of size $n_{\overline{v}}$, respectively. We incorporate two methods : (1) a semiparametric empirical likelihood method for complete observations; (2) an updating method in Chen & Chen (2000) and Jiang & Zhou (2007) to update estimates from the ODS sample. With complete ODS observations from the second phase, We consider two regression models, a regression model that represents a relationship between the response and covariates of interest and one about a relationship between the response and auxiliary variable. Without loss of generality, we consider a regression model for a covariate of interest and continuous response variable,

$$Y = X\beta + e_x, \tag{1.8}$$

where β's denote regression parameters and $e_x \sim N(0, \sigma_x^2)$. On the other hand, a regression model for the auxiliary variable,

$$Y = W\gamma + e_w, \tag{1.9}$$

where γ's denote regression parameters and $e_w \sim N(0, \sigma_w^2)$. By applying the likelihood in Zhou et al. (2002) to two regression models with respect to $\beta = (\beta_0, \beta_1)'$ and $\gamma = (\gamma_0, \gamma_1)'$, respectively, we have two likelihoods for complete observations in the second stage :

For the linear model in (1.8) with ODS samples that have the data structure of $\{Y_i, X_i\}$, $i = 1, \cdots, n_v$,

$$L(\beta, G_X) = L_{SRS}(\beta, G_X) \cdot L_{ODS}(\beta, G_X)$$

$$= \left\{ \prod_{i=1}^{n_0} f_\beta(y_{0i}|x_{0i}) g_X(x_{0i}) \right\} \times \left\{ \prod_{k=1,3} \prod_{i=1}^{n_k} P(y_{ki}, x_{ki}|Y_i \in C_k) \right\},$$

where G_X and g_X denote the cumulative distribution and density function of X. For the linear model in (1.9) with ODS samples that have the data structure of $\{Y_i, W_i\}$, $i = 1, \cdots, n_v$,

$$\begin{aligned} L(\gamma, H_W) &= L_{SRS}(\gamma, H_W) \cdot L_{ODS}(\gamma, H_W) \\ &= \left\{ \prod_{i=1}^{n_0} f_\gamma(y_{0i}|w_{0i}) h_W(w_{0i}) \right\} \times \left\{ \prod_{k=1,3} \prod_{i=1}^{n_k} P(y_{ki}, w_{ki}|Y_i \in C_k) \right\}, \end{aligned}$$

where H_W and h_W denote the cumulative distribution and density function of W. By the semiparametric empirical likelihood method, we obtain $(\hat{\beta}, \hat{\gamma})$ for true value of (β^*, γ^*) with some constraints that will be given in Chapter 3. The multivariate normal distribution theory provides the asymptotic distribution of $\sqrt{n_v}(\hat{\beta} - \beta, \hat{\gamma} - \gamma)$. Since we assumed that all values of auxiliary variable and response in the study population, a regression model for the population dataset is given as

$$Y = W\gamma + e,$$

where γ's denote regression parameters and $e \sim N(0, \sigma^2)$. The estimate of γ is obtained by using maximum likelihood under SRS scheme for the population sample. We will study how to update $\hat{\beta}$ by using the updating algorithm in Chen & Chen (2000) and Jiang & Zhou (2007) under two-phase ODS design. This approach has advantages of using more information in two-phase sampling and more efficient estimators than those in Weaver and Zhou (2005) and computational ease for multiple covariates and auxiliary variables.

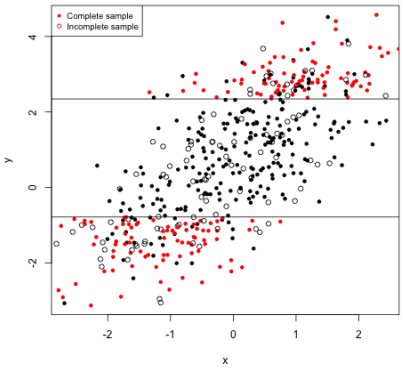

Figure 1.5: Illustration for ODS with missing data under a linear regression model

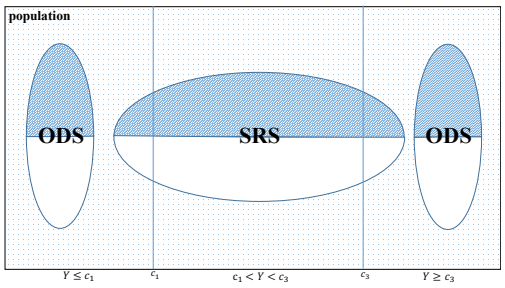

Figure 1.6: Conceptual illustration for the general ODS with missing data

1.5.2 An Estimated likelihood approach to a missing data under an Outcome-dependent Sampling

In Chapter 3, we consider an ODS design that includes missing observations in the covariate of interest with a binary auxiliary information for the covariate. To fix notation, let Y denote a continuous outcome variable, (X, Z) denote the vector of covaritates with X being an expensive scalar variable of interest and Z being an easily obtainable variable. We consider a linear regressions model,

$$Y = X\beta + \varepsilon,$$

where β denotes the unknown regression parameters and $\varepsilon \sim N(0, \sigma^2)$ is the random error. Assume that we also have a binary auxiliary variable, W, for X, and W is obtained for all observations. Compared with Zhou et al. (2002), our study is more realistic since researchers are confronted with missing covariates data in reality. The domain of Y consists of 3 mutually exclusive intervals : $C_1 \cup C_2 \cup C_3 = (-\infty, c_1] \cup (c_1, c_3] \cup (c_3, \infty)$ where c_1 and c_3 are fixed constants. The ODS sample of size N consists of three parts : SRS sample of size N_0; a supplemental ODS sample of size N_1 from C_1; another supplemental ODS sample of size N_3 from C_3. Note

that $N = N_0 + N_1 + N_3$ where $N_0 = n_0 + \overline{n}_{0,w=0} + \overline{n}_{0,w=1}$ and $N_k = n_k + \overline{n}_{k,w=0} + \overline{n}_{k,w=1}$ for $k = 1, 3$, where n_0 denotes the number of complete observations in SRS, $\overline{n}_{0,w=l}$ denotes the number of incomplete observations having $W = l$ in SRS; n_k denotes the number of complete observations drawn from C_k, and $\overline{n}_{k,w=l}$ denotes the number of incomplete observations having $W = l$ drawn from C_k. Recall that we are interested in using a binary auxiliary variable, W. Denote V as a validations set that consists of complete observations and $\overline{V}$ as a non-validation set that consists of incomplete observations that have missing in covariate. Thus, we can define V_0 as the validation set from SRS sample, $\overline{V}_0$ as the non-validation set from SRS sample, V_1 as the validation set of supplemental ODS sample from the left tail, $\overline{V}_1$ as the non-validation set of supplemental ODS sample from the left tail, V_3 as the validation set of supplemental ODS sample from the right tail, $\overline{V}_3$ as the non-validation set of supplemental ODS sample from the right tail. We have different data structures according to where samples are drawn and the missingness of covariate as follows :

SRS sample :	Complete	$\{Y_i, X_i, W_i : i \in V_0, i = 1, \cdots, n_0\}$;
	Incomplete	$\{Y_j, W_j : j \in \overline{V}_0, j = 1, \cdots, \overline{n}_0\}$;
ODS sample from the left tail :	Complete	$\{Y_i, X_i, W_i \vert Y_i \le c_1 : i \in V_1, i = 1, \cdots, n_1\}$;
	Incomplete	$\{Y_j, W_j \vert Y_j \le c_1 : j \in \overline{V}_1, j = 1, \cdots, \overline{n}_1\}$;
ODS sample from the right tail :	Complete	$\{Y_i, X_i, W_i \vert Y_i \ge c_3 : i \in V_3, i = 1, \cdots, n_3\}$;
	Incomplete	$\{Y_j, W_j \vert Y_j \ge c_3 : j \in \overline{V}_3, j = 1, \cdots, \overline{n}_3\}$.

To describe the missing data in ODS design in a graphical way, we present Figure 1.5 and 1.6. We can see that both the supplemental ODS parts and the SRS part contain missing covariates. Figure 1.6 provides a general graphical illustration of the ODS design with missing covariates. The shaded areas in ellipses represent incomplete observations and the other parts in ellipses excluding shaded area denote complete observations. One can see that our study considers missing data in covariate in every component under the ODS design.

Based on data structure above, we can derive the full likelihood as

$$
\begin{aligned}
L(\beta) &= L_{SRS}(\beta)\cdot\overline{L}_{SRS}(\beta)\cdot L_{ODS}(\beta)\cdot\overline{L}_{ODS}(\beta) \\
&= L_{SRS}(\beta)\cdot\overline{L}_{SRS}(\beta)\cdot L_{ODS_{Left}}(\beta)\cdot\overline{L}_{ODS_{Left}}(\beta)\cdot L_{ODS_{Right}}(\beta)\cdot\overline{L}_{ODS_{Right}}(\beta) \\
&= \left\{\prod_{i=1}^{n_0} P(y_i,x_i)\right\}\left\{\prod_{j=1}^{\overline{n}_0} P(y_j,w_j)\right\}\times\left\{\prod_{i=1}^{n_1} P(y_i,x_i|y_i\le c_1)\right\}\left\{\prod_{j=1}^{\overline{n}_1} P(y_j,w_j|y_j\le c_1)\right\}\times \\
&\quad \left\{\prod_{i=1}^{n_3} P(y_i,x_i|y_i\ge c_3)\right\}\left\{\prod_{j=1}^{\overline{n}_3} P(y_j,w_j|y_j\ge c_3)\right\}
\end{aligned}
$$

In Chapter 3, we will decompose the likelihood, $L(\beta)$, accordring to stratum, missingness, and auxiliary information. To obtain estimates of parameters, we propose an estimated likelihood method. Asymptotic properties, simulation results, and real data application to CPP data are conducted in the following chapters.

1.5.3 Auxiliary Covariate Stratified Sampling (ACCS) Design

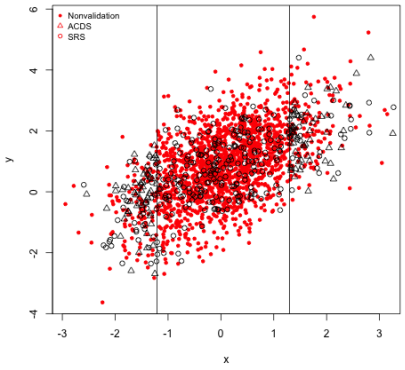

Figure 1.7: Illustration for a two-phase ACSS under a linear regression model

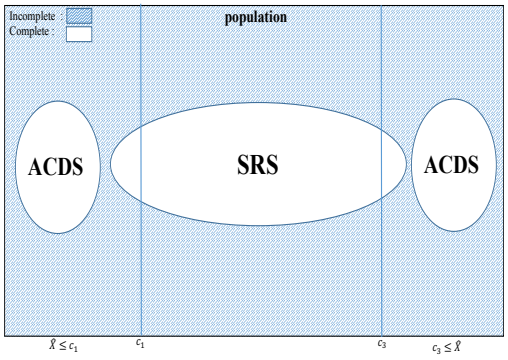

Figure 1.8: Conceptual illustration for a general two-phase ACSS design

In Chapter 4, we consider a method to estimate parameters under a linear regression model under an auxiliary covariate stratified sampling (ACSS). To fix notation, Y is denoted as a continuous outcome variable, X is a covariate of interest, and W is a continuous auxiliary variable for X. We assume that the underlying data $\{(Y_i, X_i, W_i), i = 1, \cdots, N\}$, where N is the

size of sample in study population, are identically and independent distributed. The underlying regression model is assumed as follows:

$$Y = X\beta + \varepsilon$$

where β is the vector of unknown regression parameter and ε follows a normal distribution with mean zero. We assume that the relationship between Y and X, and, the relationship between X and (Y, W) follow parametric model $f(Y \mid X; \beta)$ and $k(X \mid W; \gamma)$, where β and γ are the regression parameter, respectively. The proposed two-phase ACSS scheme with a continuous auxiliary covariate is as follows : Given with a population dataset of size N in the first stage, we can only observe (Y, W). With the population data, we choose cut-points $(c_1, c_3) = (\mu_{\hat{X}} - a * \sigma_{\hat{X}}, \mu_{\hat{X}} + a * \sigma_{\hat{X}})$ where a is a fixed constant and $\hat{X}$ is the prediction of X for all subjects in the population, based on $k(X \mid W; \gamma)$. Hence, here is the summarized data structure as follows :

In the 1st stage,

SRS : $\{Y_i, X_i, W_i, \hat{X}_i\}$

In the 2nd stage,

ACSS Left : $\{Y_i, X_i, W_i, \hat{X}_i \mid \hat{X}_i \leq c_1\}$,

ACSS Right : $\{Y_i, X_i, W_i, \hat{X}_i \mid \hat{X}_i \geq c_3\}$

Graphical illustrations are provided in Figure 1.7 and 1.8 and note that we used $\hat{X}$ based on the working model $k(X \mid W; \gamma)$ to draw supplemental ACSS samples.

Let $g(x)$ and $G(x)$ denote the probability distribution function and the cumulative distribution function of X. Then, based on the data structure above, we can construct an likelihood as follows :

$$L(\beta, G_X(x)) = \prod_{i=1}^{n_0} f_\beta(y_{0i}, x_{0i}) \cdot \prod_{j=1}^{n_1} f_\beta(y_{1j}, x_{1j} \mid \hat{X}_{1j} \leq c_1) \cdot \prod_{j=1}^{n_3} f_\beta(y_{3j}, x_{3j} \mid \hat{X}_{3j} \geq c_3)$$

$$\prod_{i=1}^{n_0} f_\beta(y_{0i} \mid x_{0i}) g(x_{0i}) \cdot \prod_{j=1}^{n_1} \frac{f_\beta(y_{1j} \mid x_{1j}) g(x_{1j})}{P(\hat{X}_{1j} \le c_1)} \cdot \prod_{j=1}^{n_3} \frac{f_\beta(y_{3j} \mid x_{3j}) g(x_{3j})}{P(\hat{X}_{3j} \ge c_3)}$$

$$\propto \left[\prod_{i=1}^{n} f_\beta(y_i \mid x_i) g(x_i) \right]$$

Based on $L(\beta, G_X(x))$, we will use the maximum likelihood to estimate β.

1.5.4 Outline of Proposed Research

In Chapter 2, an updating method in two-phase Outcome Dependent Sampling is proposed. This is an extension of the unified approach in Jiang and Zhou (2007) to ODS design and is also an extension of Weaver and Zhou (2005) to continuous auxiliary covariates information data.

In Chapter 3, we propose an estimated likelihood method under ODS design with missingness in a covariate of interest. Compared to existing ODS designs in Zhou et al. (2002) and Weaver and Zhou (2005), our proposed method considers missingness in covariates in ODS sample and to use auxiliary information about missing covariates whereas existing approaches didn't take account for missing covariate in ODS sample.

In Chapter 4, we develop an auxiliary covariate stratified sampling(ACSS) design under a linear regression model. An estimated likelihood will be derived to reflect ACSS design. It is a new approach in the sense of that auxiliary covariate is used to draw supplemental samples.

In each chapter, asymptotic results about consistency and covariate estimator are given and results of simulation studies and real data analysis are described for all methods.

CHAPTER 2: AN UPDATING METHOD WITH AUXILIARY INFORMATION UNDER A TWO -STAGE OUTCOME DEPENDENT SAMPLING DESIGN

2.1 Introduction

In biomedical studies, cost-effective study designs have been important because millions of dollars are spent to pursue biomedical studies. In many cases, it would be expensive or difficult to obtain information about covariates of interest. Depending on the case, there would be limitations to obtaining all information about main covariates because some studies don't have a sufficient budget to observe them, or participants would not give their information about the main interest of studies. For example, in biomarker studies, the cost of assaying blood samples from all subjects in the studies is expensive. A two-stage sampling design in cohort studies is a popular cost-saving approach to expensive studies. Wang and Zhou (2010) introduced the high cost problem in assessing the epidermal growth factor receptor (EGFR) mutations as a predictive biomarker for whether a subject responds to a greater extent to EGFR inhibitor drugs, and proposing an estimated likelihood approach under a two-stage outcome-auxiliary dependent sampling. Zhao and Lipsitz (1992) compared several statistical methods with twelve different two-stage designs under logistic regression models. Wang, Wu and Zhou (2009) developed a likelihood-based method for outcome and auxiliary dependent subsampling under logistic regression models. Weaver and Zhou (2005) developed an estimated likelihood approach for a linear regression model under two-stage ODS design and Zhou et al. (2011) proposed one for a two-stage outcome-auxiliary-dependent sampling design with a continuous outcome. Chen and Chen (2000) developed a unified approach that incorporated incomplete objects using an auxiliary variable and Jiang and Haibo (2007) extended this approach to survival analysis.

There have been several approaches to conducting statistical inference with ODS-type

designs. Zhou et al. (2002) proposed a semiparametric empirical likelihood method, which has been studied by Owen(1988, 1989) and Qin and Lawless (1994), with a complete observed sample from an ODS scheme and the inference results from ODS performed better in terms of efficiency than other existing methods compared in their study. This approach was extended to Probability Dependent Sampling(PDS) by Zhou et al. (2014) in a two-stage design. On the other hand, Weaver and Zhou (2005) developed an estimated likelihood method, which was developed by Pepe and Fleming (1991), to under a two-phase ODS. Wang and Zhou (2010) extended the approach to logistic regression under OADS(Outcome Auxiliary Dependent Sampling) design.

In two-phase ODS designs, along with complete samples from ODS in the second stage, there are incomplete observations in a population dataset. Weaver and Zhou (2005) studied how to exploit those non-validation sets to enhance efficiency of estimators without considering auxiliary information about the covariate of interest. ODS allows the selection probabilities of each observation in ODS sample, depending on the responses and this procedure can enhance the efficiency of estimators at limited cost. In this chapter, we consider ODS with continuous outcomes and the sample from the ODS scheme is complete, which means no missing values in each observation from ODS. The rest of sample in the population dataset has missing observations in covariates of interest. Thus, the data structure in this study is similar with the data structure that used in Weaver and Zhou (2005) except an auxiliary covariate.

The difference between our approach in this article and those taken by Zhou et al. (2002) and Weaver and Zhou (2005) is whether auxiliary information is considered to enhance efficiency of estimates under ODS design. To be specific, in Zhou et al. (2002), only complete ODS observations were used to obtain estimates. However, we use not only ODS complete observations but also take auxiliary information for covariate of interest from the whole cohort study. Intuitively, this approach will bring more efficient estimators since we can include subjects have more observations than the number of complete ODS observations. Compared to the dataset in Weaver and Zhou(2005), we consider continuous auxiliary variable for the

covariate of interest by using the multivariate normal distribution theory. From the perspective of using auxiliary information from a population dataset, Chen and Chen (2000) proposed a unified approach of regression parameters under two-stage sampling designs and Jiang and Haibo (2007) extended it to the additive hazard model. In addition to a sample from the ODS scheme, to enhance the efficiency of the estimators, we use auxiliary variable information from the whole population dataset. Chen and Chen (2000) proposed an approach to the estimation of regression parameters under double-sampling designs. They used multivariate normal distribution theory to derive a multivariate asymptotic distribution of estimator of regression model from covariates of interest and that from auxiliary covariate of main covariates from completely observed sample. After deriving the multivariate normal distribution, the updating estimator was proposed by using an estimator that was obtained through all auxiliary covariates from a population dataset.

In this chapter, we propose a new method that uses information from complete observations of the ODS sample and the population data set in a cohort study to use auxiliary information. We consider the semiparametric empirical likelihood method for complete observations from the ODS sample and, after that, we update the estimates by using the estimates from auxiliary information in the population dataset. For the estimation from ODS data, we use a semiparametric likelihood method in the sense that the marginal distribution of covariates of interest is unspecified. When we update estimators from ODS sample, multivariate normal distribution theory is used under the assumption in the Appendix. The rest of this article is structured as follows. In Section 2.2, we describe the data structure and how the likelihood would be constructed from a complete ODS sample. In Section 2.3, the way how the estimators from ODS sample are updated with auxiliary information will be presented. In Section 2.4, asymptotic properties of the updated estimators and simulation results of the proposed method will be given and compared with other existing methods. In Section 2.5, we apply the proposed method to CPP data.

2.2 Data structure and likelihood construction in the second stage

2.2.1 Notation and Data Structure

We consider an efficient and simple method to estimate regression parameters under the two-stage sampling scheme. We assume that the Outcome- Dependent Sampling (ODS) in Zhou et al. (2002) is drawn in the second stage and that the ODS sample consists of complete observations. Y referes to a continuous response, X is a covariate of interest, and W is an auxiliary variable for X. We assume that the underlying data (Y_i, X_i, W_i) for $i = 1, \cdots, N$, are independent and identically distributed, where N is the size of the sample in the first stage. The underlying regression model is assumed as follows :

$$Y = X\beta + \varepsilon_1, \tag{2.1}$$

where β is $p \times 1$ vector of unknown regression parameters and $\varepsilon_1 \sim N(0, \sigma_1^2)$. Since we assume that complete samples can be drawn only in the second stage, we first consider an ODS design in the second stage as follows : The domain of Y consists of 3 mutually exclusive intervals : $C_1 \cup C_2 \cup C_3 = (-\infty, c_1] \cup (c_1, c_3] \cup (c_3, \infty)$ where c_1 and c_3 are fixed constants. The ODS sample of size n_v consists of three parts : SRS sample of size n_0; a supplemental ODS sample of size n_1 from the left tail of $Y's$ distribution, C_1; another supplemental ODS sample of size n_3 from the right tail of $Y's$ distribution, C_3. Note that $n_v = n_0 + n_1 + n_3$ and define $\rho = n_v/N$. Hence, we can summarized the ODS sample from the second stage and primary dataset in the first stage as follows :

First stage :		$\{Y_i, W_i : i = 1, \cdots, N\}$
Second stage :	SRS	$\{Y_i, X_i, W_i : i = 1, \cdots, n_0\}$
	ODS from the left tail	$\{Y_i, X_i, W_i \mid Y_i \le c_1 : i = 1, \cdots, n_1\}$
	ODS from the right tail	$\{Y_i, X_i, W_i \mid Y_i \ge c_3 : i = 1, \cdots, n_3\}$

Recall that W is an auxiliary variable for X. Thus, (Y_i, W_i) for $i = 1, \cdots, N$ has information about the underlying linear model. Thus, we can posit a linear model between Y and W in the

population dataset as follows :

$$Y = W\gamma + \varepsilon_2, \tag{2.2}$$

where γ is $p \times 1$ vector of unknown regression parameters and $\varepsilon_2 \sim N(0, \sigma_2^2)$.

2.2.2 Likelihood function in the second stage

Based on the data structure given above and following the derivation of ODS likelihood in Zhou et al.(2002), for the model (2.1) we can construct a likelihood function for the ODS sample in the second stage as

$$\begin{aligned}
L(\beta, G_X) &= L_{SRS}(\beta, G_X) \cdot L_{ODS}(\beta, G_X) \\
&= \left\{ \prod_{i=1}^{n_0} f_\beta(y_{0i}|x_{0i}) g_X(x_{0i}) \right\} \times \left\{ \prod_{l=1,3} \prod_{i=1}^{n_l} P(y_{li}, x_{li}|Y_i \in C_l) \right\}, \\
&= \left\{ \prod_{i=1}^{n_0} f_\beta(y_{0i}|x_{0i}) \times \prod_{j=1}^{n_1} \frac{f_\beta(y_{1j}|x_{1j})}{F(c_1|x_{1j})} \times \prod_{k=1}^{n_3} \frac{f_\beta(y_{3k}|x_{3k})}{\overline{F}(c_3|x_{3k})} \right\} \\
&\quad \times \left\{ \prod_{i=1}^{n_0} g_X(x_{0i}) \times \prod_{j=1}^{n_1} \frac{F(c_1|x_{1j}) g_X(x_{1j})}{F(c_1))} \times \prod_{k=1}^{n_3} \frac{\overline{F}(c_3|x_{3k}) g_X(x_{3k})}{\overline{F}(c_3))} \right\} \\
&= L_1(\beta) \times L_2(\beta, G_X),
\end{aligned} \tag{2.3}$$

where G_X, g_X denote the cumulative distribution and density function of X, respectively, and $F(c_k) = Pr(Y \le c_k)$, $\overline{F}(c_k) = 1 - Pr(Y \le c_k)$, $F(c_k \mid x) = Pr(Y \le c_k \mid x)$ and $\overline{F}(c_k \mid x) = 1 - Pr(Y \le c_k \mid x)$.

Following the same logic, we can construct a likelihood function with the regression model (2.2), as follows :

$$\begin{aligned}
L(\gamma, H_W) &= L_{SRS}(\gamma, H_W) \cdot L_{ODS}(\gamma, H_W) \\
&= \left\{ \prod_{i=1}^{n_0} f_\gamma(y_{0i}|w_{0i}) h_W(w_{0i}) \right\} \times \left\{ \prod_{l=1,3} \prod_{i=1}^{n_l} P(y_{li}, w_{li}|Y_i \in C_l) \right\},
\end{aligned}$$

$$
\begin{aligned}
&= \left\{ \prod_{i=1}^{n_0} f_\gamma(y_{0i}|w_{0i}) \times \prod_{j=1}^{n_1} \frac{f_\gamma(y_{1j}|w_{1j})}{F(c_1|w_{1j})} \times \prod_{k=1}^{n_3} \frac{f_\gamma(y_{3k}|w_{3k})}{\overline{F}(c_3|w_{3k})} \right\} \\
&\quad \times \left\{ \prod_{i=1}^{n_0} h_W(w_{0i}) \times \prod_{j=1}^{n_1} \frac{F(c_1|w_{1j})h_W(w_{1j})}{F(c_1)} \times \prod_{k=1}^{n_3} \frac{\overline{F}(c_3|w_{3k})h_W(w_{3k})}{\overline{F}(c_3)} \right\} \\
&= L_1(\gamma) \times L_2(\gamma, H_W), \qquad (2.4)
\end{aligned}
$$

where H_W and h_W denote the cumulative distribution and density function of W, respectively, $F(c_k) = Pr(Y \leq c_k)$, $\overline{F}(c_k) = 1 - Pr(Y \leq c_k)$, $F(c_k|w) = Pr(Y \leq c_k|w)$ and $\overline{F}(c_k|w) = 1 - Pr(Y \leq c_k|w)$.

2.3 Estimation and Updating

2.3.1 Inference with ODS sample from the second stage

Let $\hat{\beta}$ and $\hat{\gamma}$ be the estimate based on (2.3) and (2.4), respectively. To obtain $\hat{\beta}$ and $\hat{\gamma}$, we use the semiparametric empirical likelihood method in Zhou et al. (2002) without specifying G_X and H_W in either likelihood. First, we profile the likelihood function $L(\beta, G_X)$ by fixing β and obtain the empirical likelihood function of G_X over all distributions whose support contains the observed X values. For simplicity's sake, let $g(x_i) = p_i$, $h(w_i) = q_i$, and $F(c_k) = \pi_k$ for $k = 1, 3$. To maximize $L_2(\beta, G_X)$ with fixed β, π, we consider only discrete distributions with jumps at each of the observed points (Owen, 1988, 1990). That is, for fixed (β, π), we need to find p_i that maximize $L_2(\beta, G_X)$ under the following constraints :

$$
\left\{ p_i \geq 0, \sum_{i=1}^{n_v} p_i = 1, \sum_{i=1}^{n_v} p_i\{F(c_1|x_i) - \pi_1\} = 0, \sum_{i=1}^{n_v} p_i\{\overline{F}(c_3|x_i) - \pi_3\} = 0 \right\}. \qquad (2.5)
$$

These constraints reflect the properties of G_X being a discrete distribution with support points at the observed X values. For a fixed β, with an idea similar to that of Qin and Lawless (1994), a unique maximum for p_i in $L_2(\beta, G_X)$ under constraints (3.5) exists if 0 is inside the convex hull of points $F(c_1|x_1) - \pi_1, \cdots, F(c_1|x_{n_v}) - \pi_1$ and $\overline{F}(c_3|x_1) - \pi_3, \cdots, \overline{F}(c_3|x_{n_v}) - \pi_3$ (Owen, 1988,

1990 ; Qin and Lawless, 1994). We use the Lagrange multipliers method to derive the maximum over $\{p_i\}$. To be specific, write

$$\begin{aligned} K &= logL_2(\beta, \{p_i\}) + \rho\left(1 - \sum_{i=1}^{n_v} p_i\right) \\ &+ n_v\lambda_1 \sum_{i=1}^{n_v} p_i \left\{F(c_1|x_i) - \pi_1\right\} + n_v\lambda_3 \sum_{i=1}^{n_v} p_i \left\{\overline{F}(c_3|x_i) - \pi_3\right\}, \end{aligned}$$

where ρ and λ's are Lagrange multipliers. Taking derivatives of K with respect to $\{p_i's\}$ and solving the score equations of K with the constraints (2.5), we can obtain that $\rho = n_v$ and

$$\hat{p}_i = \frac{1}{n_v} \cdot \frac{1}{1 + \lambda_1 \left\{F(c_1|x_i) - \pi_1\right\} + \lambda_3 \left\{\overline{F}(c_3|x_i) - \pi_3\right\}}, \quad i = 1, ..., n_v.$$

From the biased sampling nature of ODS, λ_1 and λ_3 are not centered at zero (Zhou et al. 2002, 2014). To unify the notation, we center them by reparameterizing as follows :

$$\nu_1 = \lambda_1 - \frac{n_1}{n_v\pi_1}, \nu_3 = \lambda_3 - \frac{n_3}{n_v\pi_3}.$$

We define $\eta = (\pi_1, \pi_3, \nu_1, \nu_3)'$ and $\theta = (\beta, \eta)'$. By substituting $\{p_i\}$ in (2.3) with $\{\hat{p}_i\}$, we have an empirical profile likelihood function. By taking log transformation on $L(\beta, \hat{G}_X)$, we obtain the log transformed profile likelihood function as follows :

$$l(\theta) = l_1(\beta) + l_2(\beta, \eta), \tag{2.6}$$

where

$$l_2(\beta,\eta) = -\sum_{i=1}^{n_v} log\left\{1+\nu' h(x_i)\right\} - \sum_{i=1}^{n_v} log\left\{\Delta(x_i)\right\} - n_1 log\pi_1 - n_3 log\pi_3 + \sum_{j=1}^{n_1} logF(a_1|x_{1j}) + \sum_{j=1}^{n_3} log\overline{F}(a_3|x_{1j})$$

with

$$h = (h_1, h_3), h_1 = \frac{F(c_1|x_i)-\pi_1}{\Delta(x_i)}, h_3 = \frac{\overline{F}(c_3|x_i)-\pi_3}{\Delta(x_i)}$$

and

$$\Delta(x_i) = \frac{n_0}{n_v} + \frac{n_1}{n_v\pi_1}F(c_1|x_i) + \frac{n_1}{n_v\pi_1}F(c_1|x_i).$$

We call $\hat{\theta}$ the maximum semiparametric empirical likelihood estimator where $\hat{\theta}$ maximizes $l(\theta)$. The maximum semiparametric empirical likelihood estimator can be obtained with the profiled likelihood by using the Newton-Raphson method.

Following the same logic, we can obtain $\hat{\gamma}$ based on the likelihood, (2.4). Denote $q_i = h_W(w_i)$. Under the likelihood (2.4) for fixed (γ,π)we need to find q_i that maximizes $L_2(\gamma, H_W)$ under the following constraints :

$$\left\{q_i \geq 0, \sum_{i=1}^{n_v} q_i = 1, \sum_{i=1}^{n_v} q_i\left[F(c_1|w_i)-\pi_1\right] = 0, \sum_{i=1}^{n_v} q_i\left[\overline{F}(c_3|w_i)-\pi_3\right] = 0\right\}. \quad (2.7)$$

With the same procedure for model (2.2), we use the Lagrange multipliers method to derive the maximum over q_i. We write

$$
\begin{aligned}
Q &= logL_2(\gamma, \{q_i\}) + \rho\left(1 - \sum_{i=1}^{n_v} q_i\right) \\
&+ n_v\lambda_1 \sum_{i=1}^{n_v} q_i \left\{F(a_1|w_i) - \pi_1\right\} + n_v\lambda_3 \sum_{i=1}^{n_v} q_i \left\{\overline{F}(a_2|w_i) - \pi_3\right\}.
\end{aligned}
$$

Using the Lagrangian multipliers, we can obtain

$$
\hat{q}_i = \frac{1}{n_v} \cdot \frac{1}{1 + \lambda_1 \left\{F(a_1|w_i) - \pi_1\right\} + \lambda_3 \left\{\overline{F}(a_2|w_i) - \pi_3\right\}}, \quad i = 1, ..., n_v.
$$

Define $\eta = (\pi_1, \pi_3, \nu_1, \nu_3)'$ and $\zeta = (\gamma, \eta)'$. By substituting $\{q_i\}$ in (2.4) with $\{\hat{q}_i\}$, a profile likelihood function $L(\gamma, \hat{H}_W)$ is obtained. By taking log transformation on $L(\gamma, \hat{H}_W)$, we obtain a log transformed profile likelihood function as follows :

$$
l(\zeta) = l_1(\gamma) + l_2(\gamma, \eta), \tag{2.8}
$$

where

$$
\begin{aligned}
l_2(\gamma, \eta) &= -\sum_{i=1}^{n_v} log\left\{1 + \nu' g(w_i)\right\} - \sum_{i=1}^{n_v} log\left\{\Delta(w_i)\right\} - n_1 log\pi_1 - n_3 log\pi_3 \\
&+ \sum_{j=1}^{n_1} logF(a_1|w_{1j}) + \sum_{j=1}^{n_3} log\overline{F}(a_3|w_{1j})
\end{aligned}
$$

where

$$
h = (h_1, h_3), h_1 = \frac{F(a_1|w_i) - \pi_1}{\Delta(w_i)}, h_3 = \frac{\overline{F}(a_2|w_i) - \pi_3}{\Delta(w_i)}
$$

and

$$\Delta(w_i) = \frac{n_0}{n_v} + \frac{n_1}{n_v\pi_1}F(a_1|w_i) + \frac{n_1}{n_v\pi_1}F(a_1|w_i).$$

We call $\hat{\zeta}$ the maximum semiparametric empirical likelihood estimator where $\hat{\zeta}$ maximizes $l(\zeta)$. The Newton-Raphson iterative procedure is used to obtain the maximum semiparametric empirical likelihood estimator (MSELE).

2.3.2 Updating MSELE, $\hat{\beta}$

In this section, we propose an updated method incorporating an auxiliary variable that we can enhance the efficiency of the estimators. Note that we can re-express $l(\theta)$ in (2.6) as

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n_0}\left[log f_\beta(y_i|x_i) - log(1+\nu^T h(x_i)) - log\Delta(x_i)\right] + \\ &\quad \sum_{i=1}^{n_1}\left[log f_\beta(y_i|x_i) - log(1+\nu^T h(x_i)) - log\Delta(x_i) - log\pi_1\right] + \\ &\quad \sum_{i=1}^{n_3}\left[log f_\beta(y_i|x_i) - log(1+\nu^T h(x_i)) - log\Delta(x_i) - log\pi_3\right] \equiv \sum_{i=1}^{n_v} l_i(\theta), \end{aligned}$$

With the same logic, $l(\zeta)$ in (2.8) can be written as

$$\begin{aligned} l(\zeta) &= \sum_{i=1}^{n_0}\left[log f_\gamma(y_i|w_i) - log(1+\nu^T h(w_i)) - log\Delta(w_i)\right] + \\ &\quad \sum_{i=1}^{n_1}\left[log f_\beta(y_i|w_i) - log(1+\nu^T h(w_i)) - log\Delta(w_i) - log\pi_1\right] + \\ &\quad \sum_{i=1}^{n_3}\left[log f_\beta(y_i|w_i) - log(1+\nu^T h(w_i)) - log\Delta(w_i) - log\pi_3\right] \equiv \sum_{i=1}^{n_v} l_i(\zeta), \end{aligned}$$

Recall that $\hat{\theta}$ and $\hat{\zeta}$ are defined as estimates of θ and ζ based on the semiparametric empirical likelihood approach.

Now, to update the estimator, $\hat{\theta}$, from the data in the second stage, we use the following proposition in Chen and Chen (2000) and Jiang and Haibo (2007) to obtain the asymptotic distribution of $(\hat{\theta}, \hat{\zeta})$. Let $S_i(\theta) = \frac{\partial l_i(\theta)}{\partial \theta}$ and $S_i(\zeta) = \frac{\partial l_i(\zeta)}{\partial \zeta}, i = 1, \cdots, n_v$.

Proposition 1. Under the regularity conditions in the Appendix, as $n_v \to \infty$, $(\hat{\theta}, \hat{\zeta})$ is consistent for the true parameter vector, (θ_0, ζ_0) and

$$\sqrt{n_v}(\hat{\theta} - \theta_0, \hat{\zeta} - \zeta_0)' \to_d N(0, \Sigma)$$

where

$$\Sigma = D^{-1} F D^{-1}$$

with $D = \text{diag}(D_1, D_2)$, $D_1 \equiv E\{\partial S(\theta_0)/\partial\theta\}$, $D_2 \equiv E\{\partial S(\zeta_0)/\partial\zeta\}$, and
$F \equiv E\left[\left\{S(\theta_0), S'(\zeta_0)\right\}\left\{S(\theta_0), S'(\zeta_0)\right\}'\right]$.

By multivariate normal distribution theory in Shao (2003), the conditional distribution of $\sqrt{n_v}(\hat{\theta} - \theta_0)$ given $\sqrt{n_v}(\hat{\zeta} - \zeta_0)$ is asymptotically normal with mean $\Sigma_{12}\Sigma_{22}^{-1}\sqrt{n_v}(\hat{\zeta} - \zeta_0)$. Each element in D can be estimated as follows :

$$\hat{D}_1 = \frac{1}{n}\sum_{i=1}^{n_v} \frac{\partial S_i(\theta)}{\partial \theta}, \ \hat{D}_2 = \frac{1}{n}\sum_{i=1}^{n_v} \frac{\partial S_i(\zeta)}{\partial \zeta},$$

F can be factorized into

$$F_{11} = E\{S(\theta_0)S^T(\theta_0)\}, F_{12} = E\{S(\theta_0)S^T(\zeta_0)\},$$

$$F_{21} = E\{S(\zeta_0)S^T(\theta_0)\}, F_{22} = E\{S(\zeta_0)S^T(\zeta_0)\}$$

and each partitioned matrix element can be estimated as followings :

$$\hat{F}_{11} = \frac{1}{n_v}\sum_{i=1}^{n_v} S_i(\theta)S_i^T(\theta), \hat{F}_{12} = \frac{1}{n_v}\sum_{i=1}^{n_v} S_i(\theta)S_i^T(\zeta),$$
$$\hat{F}_{21} = \frac{1}{n_v}\sum_{i=1}^{n_v} S_i(\zeta)S_i^T(\theta), \hat{F}_{22} = \frac{1}{n_v}\sum_{i=1}^{n_v} S_i(\zeta)S_i^T(\zeta).$$

Hence, the conditional mean of $\sqrt{n}(\hat{\theta} - \theta_0)$ given $\sqrt{n_v}(\hat{\zeta} - \zeta_0)$ is estimated $\hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}\sqrt{n}(\hat{\zeta} - \zeta_0)$. By equating $\sqrt{n_v}(\hat{\theta} - \theta_0)$ with its conditional mean and solving for θ, we obtain a resonable estimator $\overline{\theta}$ as $\overline{\theta} = \hat{\theta} - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}(\hat{\zeta} - \zeta_0)$. Since we are interested in parameters, β and γ, we can take β and γ from θ and ζ. Thus, we have $\overline{\beta} = \hat{\beta} - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}(\hat{\gamma} - \gamma_0)$ as a updating estimator for β. Let $\overline{\gamma}$ be the estimate of γ_0 under (2.2) on the underlying sample. Since we assumed that the underlying data consists of independently and identically distributed subjects, we can obtain $\overline{\gamma}$, by solving an estimating equation for a linear regression model (2.2),

$$0 = \sum_{i=1}^{N} \widetilde{S}_i(\gamma) \equiv \sum_{i=1}^{N} \partial log f_\gamma(y_i \mid w_i)/\partial\gamma \tag{2.9}$$

By replacing γ with $\overline{\gamma}$, we obtain

$$\overline{\beta} = \hat{\beta} - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}(\hat{\gamma} - \overline{\gamma}) \tag{2.10}$$

as an updating estimator of β. By the proposition in Chen and Chen (2000), $\overline{\beta}$ and $\overline{\gamma}$ are consistent estimator of β and γ, respectively.

2.4 Asymptotic properties

In this section asymptotic properties of the updated estimator, $\overline{\beta}$, are presented. Recall that, to obtain $\hat{\beta}$ and $\hat{\gamma}$ from ODS, since we are interested in parameters, β and γ, we can take β and γ portion from θ and ζ, respectively. Under the regularity conditions in the Appendix, we can derive a multivariate normal distribution of $\sqrt{n_v}(\hat{\beta}-\beta_0, \hat{\gamma}-\gamma_0, \overline{\gamma}-\gamma_0)^T$, where (β_0, γ_0) are true values of (β, γ). Note that we assume $n_v/N \to \rho$ where ρ is a finite constant as $n_v, N \to \infty$.

Theorem 1. *Under the regularity conditions in the Appendix, as* $n_v, N \to \infty$*,*

$$\sqrt{n_v}(\hat{\beta}-\beta_0, \hat{\gamma}-\gamma_0, \overline{\gamma}-\gamma_0)^T \sim MVN(0, \Sigma_* = D_*^{-1} C D_*^{-1}). \tag{2.11}$$

where

$$D^* = \left[diag(D_{11}^*, D_{22}^*, D_{33}^*) \right]^{-1}, \quad C = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{12}^T & C_{22} & C_{23} \\ C_{13}^T & C_{23}^T & C_{33} \end{bmatrix},$$

$$\begin{aligned} D_{11}^* &= -E\left\{\frac{\partial S(\beta_0)}{\partial \beta}\right\}, D_{22}^* = -E\left\{\frac{\partial S(\gamma_0)}{\partial \gamma}\right\}, D_{33}^* = -E\left\{\frac{\partial \widetilde{S}(\gamma_0)}{\partial \gamma}\right\}, \\ C_{11} &= E\left\{S(\beta_0)S(\beta_0)'\right\}, C_{12} = E\left\{S(\beta_0)S(\gamma_0)'\right\}, C_{13} = \frac{1}{\rho}E\left\{S(\beta_0)\widetilde{S}(\gamma_0)'\right\}, \\ C_{22} &= E\left\{S(\gamma_0)S(\gamma_0)'\right\}, C_{23} = \frac{1}{\rho}E\left\{S(\gamma)\widetilde{S}(\gamma_0)'\right\}, C_{33} = \frac{1}{\rho}E\left\{\widetilde{S}(\gamma_0)\widetilde{S}(\gamma_0)'\right\}. \end{aligned}$$

Then, by the law of large numbers in Lehmann (1999), a consistent estimator for the variance Σ_* in Theorem 1 can be estimated as follows :

Theorem 2. *A consistent estimator for the variance matrix in (2.11) can be obtained as*

$$\hat{\Sigma}_* = \hat{D}_*^{-1}\hat{C}\hat{D}_*^{-1}, \tag{2.12}$$

where

$$\hat{I}_{11} = \frac{-1}{n_v}\sum_{i=1}^{n_v}\partial S_i(\hat{\beta})/\partial\beta, \quad \hat{I}_{22} = \frac{-1}{n_v}\sum_{i=1}^{n_v}\partial S_i(\hat{\gamma})/\partial\gamma, \quad \hat{I}_{33} = \frac{-1}{N}\sum_{i=1}^{n_v}\partial S_i(\overline{\gamma})/\partial\gamma,$$

$$\hat{C}_{13} = \frac{1}{\rho}\frac{1}{N}\sum_{i=1}^{N} I(i \in V)S_i(\hat{\beta})\widetilde{S}_i'(\overline{\gamma}), \quad \hat{C}_{23} = \frac{1}{\rho}\frac{1}{N}\sum_{i=1}^{N} I(i \in V)S_i(\hat{\gamma})\widetilde{S}_i'(\overline{\gamma}),$$

V is the set of subjects in the second stage.

The theorems above can be considered as extension of the methods in Chen and Chen (2000) and Jiang and Haibo (2007). However, our method is more general in that it reflects an outcome-dependent sampling design. Lastly, to derive the asymptotic normal distribution of $\overline{\beta}$, we propose the following theorem.

Theorem 3. *Let $A = [I_{p\times p}, -\Sigma_{12}\Sigma_{22}, \Sigma_{12}\Sigma_{22}]$. By multiplying $\sqrt{n_v}(\hat{\beta} - \beta_0, \hat{\gamma} - \gamma_0, \overline{\gamma} - \gamma_0)^T$ by the matrix A and under the regularity conditions in the Appendix, $\sqrt{n_v}(\overline{\beta} - \beta_0)$ is asymptotically normal with mean zero and covariance matrix $A\Sigma_* A^T$.*

Note that Theorem 3 shows the asymptotic distribution that accounts for two-stage ODS design. Detailed proofs of the theorems above are provided in the Appendix.

2.5 Simulation study

In this section, we conduct simulation studies to evaluate the proposed estimator in finite sample situations. The simulations studies are conducted with the statistical software, R version 3.2.2. The data generated under the model :

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon \tag{2.13}$$

where $X \sim N(0,1)$, $Z \sim Bernoulli(0.45)$, $\varepsilon \sim N(0,1)$, and X and Z are independent. Thus, the conditional distribution of Y given X and Z is normal distribution with mean $\beta_0 + \beta_1 X + \beta_2 Z$ and variance 1. To generate the auxiliary variable, W, as $W = X + e$ where $e \sim N(0, \sigma_w^2)$. We fix $\beta_0 = 1$ and $\beta_2 = -0.5$ for all simulations and vary β_1 and σ_w to see performances under different settings. We also implement simulations with varying sample portions of the supplemental ODS sample out of the total sample in the second stage. In a two-stage design, we assume that there are N subjects available in the first stage. We assume X is observed only in the second stage but Y, Z, and W are observed for all subjects in the first stage and n_v subjects are drawn in the second stage.

We compare the proposed method, $\hat{\beta}_p$ with four existing estimators : (a) The first estimator is the maximum likelihood estimator with the sample size of n_v, which is denoted by $\hat{\beta}_S$ based on the SRS; (b) The second estimator is the semiparametric empirical likelihood estimator in Zhou et al. (2002), which is denoted by $\hat{\beta}_Z$, with validation sample of size n_v under ODS design in the second stage; (c) The third estimator is the probability-weighted estimator in Horvitz and Thompson(1952), which is denoted by $\hat{\beta}_{IPW}$, based on the two-stage ODS design; (d) The fourth competitor is an updated estimator, $\hat{\beta}_{SRS_{up}}$, in Chen and Chen(2000) with an SRS sample having the same size of n_v in the second stage and the same size of N in the first stage with the proposed method. Note that comparison between $\hat{\beta}_S$ and $\hat{\beta}_S$ is to see the efficiency gain when one includes auxiliary information and uses ODS design. $\hat{\beta}_Z$ is obtained with complete ODS observations to compare efficiency of $\hat{\beta}_P$ that incorporates observations that have missing-in covariate X. By using auxiliary variable W, we could use those incomplete subjects in $\hat{\beta}_P$. $\hat{\beta}_{IPW}$ is obtained by calculating selection probabilities in the study population and comparison against $\hat{\beta}_{IPW}$ demonstrates efficiency gain when we use missing observations with their auxiliary variable with the same sample size of $\hat{\beta}_P$ under ODS design. Lastly, we compare $\hat{\beta}_P$ to $\hat{\beta}_{SRS_{up}}$ to see effect of ODS design over SRS design when one uses the same sample size and the same size of study population.

The simulation results based on the underlying model (2.13) are depicted in Table 2.1 and 2.2.

Means, standard errors, mean of the variance estimates, and 95% nominal coverage probabilities for each estimators are obtained from 1,000 independent simulation runs. In Table 2.1, we draw 400 of ODS sample under N=5,000 with $\sigma_w = 0$ assuming no measurement error between covariate of interest X and auxiliary variable W. On the other hand, in Table 2.2, we set $\sigma_w = 1$ to compare efficiency gains between five methods under weak relationship between covariate of interest X and auxiliary variable W. We could see efficiency changes varying supplemental ODS sample portion by comparing results in Table 2.1 and 2.2. All estimators for β are unbiased under our simulation settings, β_P is the most efficient among all compared estimators. Based on 1,000 runs, the nominal 95% confidence interval coverage rates the averages of the variance estimator proposed is close to the empirical variance. Efficiency gains are slightly higher when we allocate more samples to the tails under the fixed size of ODS sample. In addition, a stronger relationship between X and W brings smaller SE of $\hat{\beta}_P$ when we compared SE of $\hat{\beta}_P$ between Tables 2.1 and 2.2. In addition, to compare $\hat{\beta}_p$ with the proposed estimator in Weaver and Zhou (2005), we conduct additional simulation study and results are given in Table 2.3 and 2.4. $\hat{\beta}_W$ denotes the proposed method in Weaver and Zhou (2005). Note that the simulation results in Table 2.3 and Table 2.4 are conducted with $\sigma_w^2 = 0$ and $\sigma_w^2 = 1$, respectively. In Table 2.3, under all settings, SE from $\hat{\beta}_p$ is smaller than SE from $\hat{\beta}_W$. In Table 2.4, with $\sigma_w^2 = 1$, under true value of $\beta_1 = 0$, $\hat{\beta}_p$ is more efficient than $\hat{\beta}_W$. We could find that as β_1 is increasing, SE of $\hat{\beta}_p$ increases. By comparing Table 2.3 and 2.4, $\hat{\beta}_p$ would be more efficient than $\hat{\beta}_W$ when there exists a strongly correlated auxiliary covariate.

Table 2.1: Simulation results for asymptotic properties in Section 2.4. Results are based on 1,000 simulations with N = 5000, various (n_0, n_1, n_3), and σ_w^2 = 0

β_1	Method	Mean	SE	$\widehat{\text{SE}}$	CI	β_2	Mean	SE	$\widehat{\text{SE}}$	CI
(n_0, n_1, n_3)=(300,50,50)										
0	$\hat{\beta}_{SRS}$	0.000	0.049	0.050	0.952	-0.5	-0.502	0.101	0.100	0.946
	$\hat{\beta}_{IPW}$	0.000	0.044	0.045	0.951		-0.508	0.087	0.090	0.949
	$\hat{\beta}_Z$	0.001	0.041	0.043	0.954		-0.497	0.088	0.088	0.945
	$\hat{\beta}_{SRS_{up}}$	0.002	0.014	0.014	0.939		-0.500	0.028	0.028	0.945
	$\hat{\beta}_P$	0.000	0.014	0.013	0.942		-0.499	0.029	0.027	0.930
0.5	$\hat{\beta}_{SRS}$	0.501	0.049	0.050	0.952		-0.502	0.101	0.100	0.946
	$\hat{\beta}_{IPW}$	0.501	0.045	0.047	0.956		-0.501	0.089	0.093	0.958
	$\hat{\beta}_Z$	0.501	0.046	0.048	0.951		-0.495	0.090	0.090	0.948
	$\hat{\beta}_{SRS_{up}}$	0.500	0.014	0.014	0.939		-0.500	0.028	0.028	0.945
	$\hat{\beta}_P$	0.500	0.014	0.013	0.941		-0.499	0.029	0.027	0.930
(n_0, n_1, n_3)=(200,100,100)										
0	$\hat{\beta}_{SRS}$	0.001	0.049	0.050	0.951	-0.5	-0.502	0.101	0.100	0.945
	$\hat{\beta}_{IPW}$	0.000	0.045	0.045	0.952		-0.508	0.091	0.091	0.949
	$\hat{\beta}_Z$	0.001	0.038	0.038	0.952		-0.500	0.078	0.080	0.958
	$\hat{\beta}_{SRS_{up}}$	0.002	0.014	0.014	0.939		-0.500	0.028	0.028	0.945
	$\hat{\beta}_p$	0.000	0.014	0.013	0.950		-0.500	0.027	0.027	0.949
0.5	$\hat{\beta}_{SRS}$	0.501	0.049	0.050	0.952		-0.502	0.101	0.100	0.946
	$\hat{\beta}_{IPW}$	0.502	0.047	0.050	0.954		-0.505	0.095	0.098	0.950
	$\hat{\beta}_Z$	0.503	0.045	0.046	0.948		-0.502	0.081	0.084	0.952
	$\hat{\beta}_{SRS_{up}}$	0.500	0.014	0.014	0.939		-0.500	0.028	0.028	0.945
	$\hat{\beta}_p$	0.500	0.014	0.013	0.949		-0.500	0.027	0.027	0.948
(n_0, n_1, n_3)=(100,150,150)										
0	$\hat{\beta}_{SRS}$	0.001	0.049	0.050	0.952	-0.5	-0.502	0.101	0.100	0.946
	$\hat{\beta}_{IPW}$	0.005	0.055	0.054	0.935		-0.515	0.112	0.109	0.941
	$\hat{\beta}_Z$	0.000	0.036	0.035	0.948		-0.500	0.074	0.074	0.949
	$\hat{\beta}_{SRS_{up}}$	0.002	0.014	0.014	0.939		-0.500	0.028	0.028	0.945
	$\hat{\beta}_p$	0.000	0.014	0.013	0.934		-0.500	0.028	0.027	0.945
0.5	$\hat{\beta}_{SRS}$	0.501	0.049	0.050	0.952		-0.502	0.101	0.100	0.946
	$\hat{\beta}_{IPW}$	0.507	0.063	0.064	0.947		-0.509	0.128	0.123	0.924
	$\hat{\beta}_Z$	0.502	0.044	0.044	0.946		-0.501	0.079	0.079	0.951
	$\hat{\beta}_{SRS_{up}}$	0.500	0.014	0.014	0.939		-0.500	0.028	0.028	0.945
	$\hat{\beta}_p$	0.499	0.014	0.013	0.943		-0.500	0.027	0.027	0.958

a. The results are based on the model $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, $X \sim N(0,1)$, $Z \sim Bernoulli(0.45)$

b. $W = X + N(0, \sigma_w^2)$

c. $\hat{\beta}_{SRS}, \hat{\beta}_{IPW}, \hat{\beta}_Z$ and $\hat{\beta}_{SRS_{up}}$ are described in Section 2.5

d. SE, standard error, CI, confidence interval width

Table 2.2: Simulation results for asymptotic properties in Section 2.4. Results are based on 1,000 simulations with N = 5,000, various (n_0, n_1, n_3), and σ_w^2 = 1

β_1	Method	Mean	SE	$\widehat{\text{SE}}$	CI	β_2	Mean	SE	$\widehat{\text{SE}}$	CI
(n_0, n_1, n_3)=(300,50,50)										
0	$\hat{\beta}_{SRS}$	0.001	0.049	0.050	0.952	-0.5	-0.502	0.101	0.100	0.946
	$\hat{\beta}_{IPW}$	0.000	0.045	0.045	0.941		-0.499	0.090	0.090	0.950
	$\hat{\beta}_Z$	0.001	0.041	0.043	0.954		-0.497	0.088	0.088	0.945
	$\hat{\beta}_{SRS_{up}}$	0.000	0.037	0.036	0.947		-0.500	0.030	0.29	0.940
	$\hat{\beta}_p$	-0.000	0.032	0.032	0.941		-0.498	0.028	0.028	0.951
0.5	$\hat{\beta}_{SRS}$	0.501	0.049	0.050	0.952		-0.502	0.101	0.100	0.946
	$\hat{\beta}_{IPW}$	0.501	0.046	0.047	0.949		-0.504	0.092	0.093	0.950
	$\hat{\beta}_Z$	0.501	0.046	0.048	0.951		-0.495	0.090	0.090	0.948
	$\hat{\beta}_{SRS_{up}}$	0.500	0.039	0.038	0.945		-0.501	0.044	0.043	0.939
	$\hat{\beta}_p$	0.499	0.038	0.037	0.941		-0.500	0.044	0.043	0.949
(n_0, n_1, n_3)=(200,100,100)										
0	$\hat{\beta}_{SRS}$	0.001	0.049	0.050	0.951	-0.5	-0.502	0.101	0.100	0.945
	$\hat{\beta}_{IPW}$	-0.000	0.047	0.045	0.949		-0.503	0.092	0.091	0.941
	$\hat{\beta}_Z$	0.001	0.038	0.038	0.952		-0.500	0.078	0.080	0.958
	$\hat{\beta}_{SRS_{up}}$	0.000	0.037	0.036	0.947		-0.500	0.030	0.29	0.940
	$\hat{\beta}_p$	-0.000	0.030	0.029	0.944		-0.500	0.029	0.028	0.944
0.5	$\hat{\beta}_{SRS}$	0.501	0.049	0.050	0.9 52		-0.502	0.101	0.100	0.946
	$\hat{\beta}_{IPW}$	0.503	0.048	0.050	0.954		-0.506	0.103	0.098	0.939
	$\hat{\beta}_Z$	0.503	0.045	0.046	0.948		-0.502	0.081	0.084	0.952
	$\hat{\beta}_{SRS_{up}}$	0.500	0.039	0.038	0.945		-0.501	0.044	0.043	0.939
	$\hat{\beta}_P$	0.502	0.037	0.037	0.953		-0.501	0.044	0.044	0.949
(n_0, n_1, n_3)=(100,150,150)										
0	$\hat{\beta}_{SRS}$	0.001	0.049	0.050	0.952	-0.5	-0.502	0.101	0.100	0.946
	$\hat{\beta}_{IPW}$	0.000	0.056	0.054	0.931		-0.510	0.111	0.109	0.941
	$\hat{\beta}_Z$	0.000	0.036	0.035	0.948		-0.500	0.074	0.074	0.949
	$\hat{\beta}_{SRS_{up}}$	0.000	0.037	0.036	0.947		-0.500	0.030	0.29	0.940
	$\hat{\beta}_p$	0.000	0.028	0.028	0.943		-0.499	0.028	0.028	0.953
0.5	$\hat{\beta}_{SRS}$	0.501	0.049	0.050	0.952		-0.502	0.101	0.100	0.952
	$\hat{\beta}_{IPW}$	0.507	0.065	0.063	0.934		-0.509	0.121	0.122	0.953
	$\hat{\beta}_Z$	0.502	0.044	0.044	0.946		-0.501	0.079	0.079	0.951
	$\hat{\beta}_{SRS_{up}}$	0.500	0.039	0.038	0.945		-0.501	0.044	0.043	0.939
	$\hat{\beta}_p$	0.500	0.035	0.036	0.960		-0.499	0.044	0.044	0.950

a. The results are based on the model $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, $X \sim N(0, 1)$, $Z \sim Bernoulli(0.45)$

b. $W = X + N(0, \sigma_w^2)$

c. $\hat{\beta}_{SRS}, \hat{\beta}_{IPW}, \hat{\beta}_Z$ and $\hat{\beta}_{SRS_{up}}$ are described in Section 2.5

d. SE, standard error, CI, confidence interval width

Table 2.3: Simulation results for asymptotic properties in Section 2.4. Results are based on 1,000 simulations with N = 2,000, various (n_0, n_1, n_3) = $(200, 50, 50)$, and σ_w^2 = 0

β_1	Method	Mean	SE	$\widehat{\text{SE}}$	CI	β_2	Mean	SE	$\widehat{\text{SE}}$	CI
0	$\hat{\beta}_{SRS}$	0.003	0.058	0.058	0.948	-0.5	-0.498	0.116	0.116	0.941
	$\hat{\beta}_{IPW}$	-0.001	0.052	0.051	0.934		-0.503	0.099	0.103	0.962
	$\hat{\beta}_{Z}$	0.000	0.049	0.047	0.935		-0.502	0.100	0.099	0.947
	$\hat{\beta}_{W}$	0.001	0.048	0.048	0.954		-0.500	0.092	0.091	0.944
	$\hat{\beta}_{SRS_{up}}$	0.001	0.024	0.022	0.929		-0.498	0.048	0.044	0.926
	$\hat{\beta}_{p}$	0.000	0.023	0.022	0.941		-0.500	0.046	0.044	0.941
0.5	$\hat{\beta}_{SRS}$	0.502	0.058	0.058	0.949	-0.5	-0.506	0.116	0.116	0.945
	$\hat{\beta}_{IPW}$	0.501	0.054	0.054	0.946		-0.500	0.105	0.108	0.960
	$\hat{\beta}_{Z}$	0.503	0.054	0.054	0.960		-0.499	0.098	0.102	0.949
	$\hat{\beta}_{W}$	0.503	0.047	0.047	0.954		-0.500	0.096	0.098	0.953
	$\hat{\beta}_{SRS_{up}}$	0.498	0.023	0.022	0.936		-0.499	0.047	0.045	0.937
	$\hat{\beta}_{p}$	0.500	0.023	0.022	0.940		-0.500	0.045	0.044	0.942
1	$\hat{\beta}_{SRS}$	1.002	0.058	0.058	0.949	-0.5	-0.506	0.116	0.116	0.945
	$\hat{\beta}_{IPW}$	1.001	0.055	0.056	0.943		-0.499	0.113	0.115	0.955
	$\hat{\beta}_{Z}$	1.002	0.058	0.058	0.942		-0.504	0.107	0.107	0.947
	$\hat{\beta}_{W}$	1.002	0.047	0.047	0.951		-0.495	0.106	0.108	0.948
	$\hat{\beta}_{SRS_{up}}$	1.000	0.024	0.022	0.929		-0.498	0.048	0.044	0.926
	$\hat{\beta}_{p}$	1.000	0.023	0.022	0.939		-0.500	0.046	0.044	0.941

a. The results are based on the model $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, $X \sim N(0, 1)$, $Z \sim Bernoulli(0.45)$

b. $W = X + N(0, \sigma_w^2)$

c. $\hat{\beta}_{SRS}, \hat{\beta}_{IPW}, \hat{\beta}_{Z}$ and $\hat{\beta}_{SRS_{up}}$ are described in Section 2.5

d. $\hat{\beta}_{W}$ is the MELE in Weaver and Zhou (2005)

e. SE, standard error, CI, confidence interval width

Table 2.4: Simulation results for asymptotic properties in Section 2.4. Results are based on 1,000 simulations with N = 2,000, various (n_0, n_1, n_3) = $(200, 50, 50)$, and σ_w^2 = 1

β_1	Method	Mean	SE	$\widehat{\text{SE}}$	CI	β_2	Mean	SE	$\widehat{\text{SE}}$	CI
0	$\hat{\beta}_{SRS}$	0.003	0.058	0.058	0.948	-0.5	-0.498	0.116	0.116	0.941
	$\hat{\beta}_{IPW}$	-0.001	0.052	0.051	0.934		-0.503	0.099	0.103	0.962
	$\hat{\beta}_{Z}$	0.000	0.049	0.047	0.935		-0.502	0.100	0.099	0.947
	$\hat{\beta}_{W}$	0.001	0.048	0.048	0.954		-0.500	0.092	0.091	0.944
	$\hat{\beta}_{SRS_{up}}$	-0.001	0.043	0.043	0.949		-0.498	0.048	0.045	0.933
	$\hat{\beta}_{p}$	0.003	0.038	0.037	0.943		-0.502	0.046	0.045	0.945
0.5	$\hat{\beta}_{SRS}$	0.502	0.058	0.058	0.949	-0.5	-0.506	0.116	0.116	0.945
	$\hat{\beta}_{IPW}$	0.501	0.054	0.054	0.946		-0.500	0.105	0.108	0.960
	$\hat{\beta}_{Z}$	0.503	0.054	0.054	0.960		-0.499	0.098	0.102	0.949
	$\hat{\beta}_{W}$	0.503	0.047	0.047	0.954		-0.500	0.096	0.098	0.953
	$\hat{\beta}_{SRS_{up}}$	0.499	0.047	0.045	0.942		-0.498	0.058	0.057	0.942
	$\hat{\beta}_{p}$	0.503	0.045	0.045	0.944		-0.504	0.059	0.059	0.947
1	$\hat{\beta}_{SRS}$	1.002	0.058	0.058	0.949	-0.5	-0.506	0.116	0.116	0.945
	$\hat{\beta}_{IPW}$	1.001	0.055	0.056	0.943		-0.499	0.113	0.115	0.955
	$\hat{\beta}_{Z}$	1.002	0.058	0.058	0.942		-0.504	0.107	0.107	0.947
	$\hat{\beta}_{W}$	1.002	0.047	0.047	0.951		-0.495	0.106	0.108	0.948
	$\hat{\beta}_{SRS_{up}}$	0.998	0.048	0.048	0.945		-0.499	0.078	0.076	0.941
	$\hat{\beta}_{p}$	1.005	0.048	0.051	0.958		-0.505	0.080	0.078	0.945

a. The results are based on the model $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, $X \sim N(0, 1)$, $Z \sim Bernoulli(0.45)$

b. $W = X + N(0, \sigma_w^2)$

c. $\hat{\beta}_{SRS}, \hat{\beta}_{IPW}, \hat{\beta}_{Z}$ and $\hat{\beta}_{SRS_{up}}$ are described in Section 2.5

d. $\hat{\beta}_{W}$ is the MELE in Weaver and Zhou (2005)

e. SE, standard error, CI, confidence interval width

2.6 Application to the PCB Data

In this section, we apply the proposed method to CPP study in Niswander and Gordon (1972), Gray et al. (2005), and Zhou et al. (2002). In 12 medical centers in 11 cities in the U.S 44,000 women enrolled into the study between 1959 and 1965. 55,908 pregnancies resulted in with multiple pregnancies. Children in the study were followed until age 8 years. Mothers' non-fasting blood was collected at each prenatal visit. Investigators were interested in the effect of mother's maternal pregnancy serum level of polychlorinated biphenyls (PCBs) on cognitive test scores (IQ) at 7 years of age on the Wechsler Intelligence Scale for Children (WISC) since "epidemiologic studies of perinatal exposure to background-level PCBs in relation to cognitive function in children have been giving inconsistent results" (Gray et al., 2005, p. 18). We use the dataset in Zhou et al. (2002). There were two criteria of children eligibility : 1) They were liveborn singletons, 2) a 3-ml third trimester maternal serum specimen was available. Of the children in CPP study, 44,075 satisfied all of eligibility criteria. In addition to PCB levels, other covariates, the socioeconomic status of the child's family (SES), gender (SEX), race (RACE) and the parent's education (EDU) were also collected. To apply the proposed method, we consider the following linear regression model :

$$IQ = \beta_0 + \beta_1 PCB + \beta_2 SES + \beta_3 SEX + \beta_4 RACE + \beta_5 EDU + \varepsilon$$

Out of 44,075 eligible subjects, 38,709 have complete data of all variables except PCB in the linear regression model above. Zhou et al. (2002) drew an ODS sample of size $n_v = 1,038$ that consisted of $n_0 = 849, n_1 = 81,$ and $n_3 = 108$. From the dataset in Zhou et al. (2002), we generate an auxiliary variable for PCB as $W = PCB + \varepsilon_W$, where $\varepsilon_W \sim N(0,1)$ for all 1,038 subjects. To apply the proposed two-stage sampling design to the dataset in Zhou et al. (2002), we draw an ODS sample at the second stage first: we select an SRS with $n_0 = 200$ from the SRS portion of 849 subjects. Next, we draw 81 of supplementary sample from the left tail portion of $n_1 = 81$ and right tail portion of $n_3 = 108$, respectively, as supplementary ODS samples. Thus,

we have n_v = 362, N = $1,038$, and 676($= N - n_v$) subjects do not have PCBs. As we explained data structure in Section 2.2, we assume that 1,038 subjects in the first stage have W for PCB whereas 362 subjects in ODS sample have PCBs.

In Table 2.5, we summarize the comparison results between MSELE and our proposed method. The results in Table 2.5 also show that the parameter estimates are similar for both methods. To be specific, it is notable that PCB and SEX are not significantly related to IQ test performance in both analyses. In both methods, SEX is not statistically significant effect on children's IQ score at 7 years of age. We could confirm that the proposed bring more efficient estimates than MSELE does: especially for PCB, the proposed method is more efficient in the sense that the proposed method has smaller standard error and a narrower 95% confidence interval of $(-0.11, 0.65)$ than the MSELE method of $(-0.29, 0.75)$.

Table 2.5: Analysis results for the CPP data set with N = $1,038$ and n_v = 362

	MSELE			Proposed Method		
	$\hat{\beta}$	$\widehat{SE}$	95% CI	$\hat{\beta}$	$\widehat{SE}$	95% CI
Int	77.93	3.31	(71.43, 84.43)	76.35	2.07	(72.29, 80.41)
PCB	0.22	0.26	(-0.29, 0.75)	0.27	0.19	(-0.11, 0.65)
SES	0.95	0.33	(0.29, 1.61)	1.42	0.25	(0.91, 1.93)
SEX	0.15	1.04	(-1.88, 2.19)	-0.84	0.81	(-2.43, 0.74)
RACE	-9.03	1.23	(-11.45, -6.61)	-10.17	0.91	(-11.95, -8.38)
EDU	1.62	0.31	(1.00, 2.23)	1.61	0.21	(1.20, 2.02)

a. The results of MSELE and proposed method used the cut points of (82, 110)

b. MSELE uses only ODS portion in the second stage and ODS sample consists of n_0 = 200, n_1 = n_3 = 81

c. PCB is observed only in ODS sample with the size of n_v = 362

2.7 Discussion

We have proposed an updated method under a two-stage ODS design. The ODS design in Zhou et al. (2002) is used to draw a complete ODS sample in the second second stage and the unified approach in Chen and Chen (2000) and Jiang and Haibo (2007) is applied to

incorporate auxiliary information in a study population from which we could obtain a more efficient estimator than estimators without using auxiliary information. The proposed estimator has several advantages compared to other existing methods as follows : Under the same the complete sample size in the second stage, the proposed method gains more efficiency than the methods in Zhou et al. (2002) and the inverse probability method do; The proposed method might be extended to cases which have multiple covariates and multiple auxiliary variables of covariates of interests; by using auxiliary information from a population dataset, more efficient estimators can be obtained than methods that use only complete observations; compared to the updated method in Chen and Chen (2000), the novelty is that our proposed method can account for the properties of ODS design. In addition, since Weaver and Zhou (2005) considered similar setting in terms of data size and ODS sample size, it might be interesting to compare their method with the proposed method. Since the method in Weaver and Zhou (2005) didn't use auxiliary information, we expect the proposed method would be more efficient than the estimator in Weaver and Zhou (2005). Although Weaver and Zhou (2005) could use auxiliary information, one advantage of the proposed method is that the proposed method is that it can be implemented along with multiple auxiliary covariates. Furthermore, in terms of computation, the proposed method will be easier than methods in Chatterjee et al. (2003) and Weaver and Zhou (2005) because they considered nonparametric plug-in density method to estimate unknown density.

There is an interesting area remaining for future study. Since the proposal design is a two stage sampling that consists of validation and non-validation sets, one can compare the proposed method with the AIPW estimator in Robins et al.(1994, 1995). Since the proposed method and the AIPW estimator are in different classes of estimators, the comparison would be valuable and interesting.

2.8 Proof of Theorems

2.8.1 Regularity conditions

The regularity conditions required to prove the theorems in Section 2.4 are as followings :

Condition 1. *$\rho = n_v/N, n_1/N$ and n_3/N are finite and fixed constants.*

Condition 2. *The parameter space (Θ) and (Ξ) are compact subspace of $\mathbb{R}^p$, respectively.*

Condition 3. *$(\theta, \zeta)'$ are in the interior of a compact parameter space, $(\Theta, \Xi)'$.*

Condition 4. *The log-densities $l(\theta)$ and $l(\zeta)$ are twice continuously differentiable with respect to θ and ζ, respectively.*

Condition 5. *$E[S(\theta)] \neq 0$ and $E[S(\zeta)] \neq 0$ if $\theta \neq \theta_0$ and $\zeta \neq \zeta_0$, respectively.*

Condition 6. *$E\{(S(\theta), S(\zeta))(S(\theta), S(\zeta))^T\}$ is finite and positive definite.*

Condition 7. *$E\left\{\frac{\partial S(\theta)}{\partial \theta}\right\}$, $E\left\{\frac{\partial S(\zeta)}{\partial \zeta}\right\}$ and $E\left\{\frac{\partial \widetilde{S}(\zeta)}{\partial \zeta}\right\}$ are finite and positive definite.*

Condition 8. *$E\{S(\beta)S^T(\beta)\}, E\{S(\gamma)S^T(\gamma)\}$ and $E\left\{\widetilde{S}(\gamma)\widetilde{S}^T(\gamma)\right\}$ are finite and positive definite.*

2.8.2 Proof

Outline of proof for Theorem 1 is given as follows :

From the ODS sample in the second stage, based on (2.6) and (2.8), by using the Mean Value Theorem in Khuri (2003) in Shao (2003), we have

$$\begin{aligned}
\sqrt{n_v}(\hat{\beta} - \beta_0) &\cong \sqrt{n_v}\frac{\sum_{i=1}^{n_v} S_i(\beta_0)}{\sum_{i=1}^{n_v} \partial S_i(\beta_0)/\partial\beta}, \\
\sqrt{n_v}(\hat{\gamma} - \gamma_0) &\cong \sqrt{n_v}\frac{\sum_{i=1}^{n_v} S_i(\gamma_0)}{\sum_{i=1}^{n_v} \partial S_i(\gamma_0)/\partial\gamma},
\end{aligned}$$

from the sample in the first stage, based on (2.9),

$$\sqrt{N}(\overline{\gamma} - \gamma_0) \cong \sqrt{N}\frac{\sum_{i=1}^{N} \widetilde{S}_i(\gamma_0)}{\sum_{i=1}^{N} \partial \widetilde{S}_i(\gamma_0)/\partial\gamma}.$$

Proof. Note that we can express

$$\sqrt{n_v}\begin{pmatrix}\hat{\beta}-\beta_0\\ \hat{\gamma}-\gamma_0\\ \overline{\gamma}-\gamma_0\end{pmatrix} = \sqrt{n_v}\begin{pmatrix}\frac{\sum_{i=1}^{n_v} S_i(\beta_0)}{\sum_{i=1}^{n_v}\partial S_i(\beta_0)/\partial\beta}\\ \frac{\sum_{i=1}^{n_v} S_i(\gamma_0)}{\sum_{i=1}^{n_v}\partial S_i(\gamma_0)/\partial\gamma}\\ \frac{\sum_{i=1}^{N}\widetilde{S}_i(\gamma_0)}{\sum_{i=1}^{N}\partial\widetilde{S}_i(\gamma_0)/\partial\gamma}\end{pmatrix} = \begin{pmatrix}\frac{n}{\sum_{i=1}^{n_v}\partial S_i(\beta_0)/\partial\beta}\\ \frac{n_v}{\sum_{i=1}^{n_v}\partial S_i(\gamma_0)/\partial\gamma}\\ \frac{n_v}{N}\cdot\frac{N}{\sum_{i=1}^{N}\partial\widetilde{S}_i(\gamma_0)/\partial\gamma}\end{pmatrix}\sqrt{n_v}\begin{bmatrix}\frac{\sum_{i=1}^{n_v} S_i(\beta_0)}{n}\\ \frac{\sum_{i=1}^{n_v} S_i(\gamma_0)}{n}\\ \frac{N}{n}\frac{\sum_{i=1}^{N}\widetilde{S}_i(\gamma_0)}{N}\end{bmatrix} \tag{2.14}$$

Since we assumed that, as $n_v, N \to \infty$, $\frac{n_v}{N} \to_p \rho$, by the law of large numbers in Lehmann (1999), the first term in RHS of (2.14) goes to D_*^{-1} in probability as follows :

$$\begin{pmatrix}\frac{n_v}{\sum_{i=1}^{n_v}\partial S_i(\beta_0)/\partial\beta}\\ \frac{n_v}{\sum_{i=1}^{n_v}\partial S_i(\gamma_0)/\partial\gamma}\\ \frac{n_v}{N}\cdot\frac{N}{\sum_{i=1}^{N}\partial\widetilde{S}_i(\gamma_0)/\partial\gamma}\end{pmatrix}\to_p\begin{pmatrix}I_{11}^{-1} & 0 & 0\\ 0 & I_{22}^{-1} & 0\\ 0 & 0 & \rho\cdot I_{33}^{-1}\end{pmatrix} = D_*^{-1}. \tag{2.15}$$

By the central limit theorem in Lehmann (1999), as $n_v, N \to \infty$, the second term in (2.14) converges in distribution as follows :

$$\sqrt{n_v}\begin{bmatrix}\frac{\sum_{i=1}^{n_v} S_i(\beta_0)}{n_v}\\ \frac{\sum_{i=1}^{n_v} S_i(\gamma_0)}{n_v}\\ \frac{N}{n_v}\frac{\sum_{i=1}^{N}\widetilde{S}_i(\gamma_0)}{N}\end{bmatrix}\to_d MVN(0, C), \tag{2.16}$$

where C is defined in Theorem 1 in section 2.4. By multiplying (2.16) by (2.15) and using the delta method in Shao (2003), we can obtain asymptotic distribution of $(\hat{\beta}, \hat{\gamma}, \overline{\gamma})^T$ in Theorem 1. □

Next, proof for Theorem 2 is given as follows :

Proof. To estimate C_{13}, C_{23}, we cannot directly use the formulas under SRS design since we use the ODS design in Zhou et al. (2002) for the dataset in the second stage. We can rewrite C_{13}, C_{23} as

$$C_{13} \;=\; Cov(S(\beta_0), \widetilde{S}(\gamma_0)) = E\left(S(\beta_0), \widetilde{S}(\gamma_0)\right),$$

$$C_{23} = Cov(S(\gamma_0), \widetilde{S}(\gamma_0)) = E\left(S(\gamma_0), \widetilde{S}(\gamma_0)\right).$$

By the law of large numbers in Lehmann (1999), we can obtain unbiased and consistent estimators of C_{13}, C_{23} as

$$\hat{C}_{13} = \frac{1}{N}\frac{1}{\rho}\sum_{i=1}^{N} I(i \in V) S_i(\hat{\beta}) S_i'(\overline{\gamma}),$$

$$\hat{C}_{23} = \frac{1}{N}\frac{1}{\rho}\sum_{i=1}^{N} I(i \in V) S_i(\hat{\gamma}) \widetilde{S}_i'(\overline{\gamma}).$$

□

Lastly, proof of Theorem 3 is provided below :

Proof. Recall that we have the multivariate normal distribution of $\sqrt{n_v}(\hat{\beta} - \beta_0, \hat{\gamma} - \gamma_0, \overline{\gamma} - \gamma_0)^T$ and $A = [I_{p\times p}, -\Sigma_{12}\Sigma_{22}, \Sigma_{12}\Sigma_{22}]$ in Theorem 3. In addition, note that $\overline{\beta}$ is expressed as $\hat{\beta} - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}(\hat{\gamma} - \overline{\gamma})$. Thus, we need to multiply a matrix that can produce this quantity. By using the delta method in Shao (2003), we can multiply $\sqrt{n_v}(\hat{\beta} - \beta_0, \hat{\gamma} - \gamma_0, \overline{\gamma} - \gamma_0)^T$ by A. Then, the asymptotic distribution of $\overline{\beta}$ is obtained as a normal distribution with mean zero and covariance matrix of $A\Sigma_* A' = AD_*^{-1}CD_*^{-1}A'$. □

CHAPTER 3: AN ESTIMATED LIKELIHOOD APPROACH TO A MISSING DATA UNDER THE OUTCOME-DEPENDENT SAMPLING DESIGN

3.1 Introduction

Outcome-dependent sampling (ODS) is a biased sampling scheme that has been studied by Zhou et al. (2002) and Weaver and Zhou (2005). The main idea of ODS is to draw a sample based on strata that are defined on a response variable. Using stratification yields a biased sampling scheme or choice-based sampling so that one can expect study efficiency from ODS designs. The ODS design for continuous outcomes is comparable to the case-control design for binary outcomes. Among many studies of biased sampling designs, the case-control design is a well-known method in epidemiological observational studies, since it is often preferred for rare case studies because it can yield an equal number of diseased individuals in a much smaller study. Th us, the case-control method is preferred because of its study efficiency, in particular for a binary response cases.

For studies investigating a relationship between a continuous response variable and covariate of interest, the case-control method cannot be applied directly. One approach is to dichotomize a continuous outcome according to some fixed cut points, and then use the case-control method. However, this approach loses the original information of the continuous response and can raise bias issues as well. To solve this problem, Zhou et al. (2002) developed an ODS scheme with a continuous outcome. A sample from ODS scheme consists of two parts : one is the Simple Random Sampling (SRS) part and another is the supplemental ODS part. Since it is a biased sampling for which one cannot apply standard methods for SRS design, Zhou et al. (2002) developed a semiparametric empirical likelihood estimation method.

On the other hand, in many epidemiological observational studies, missingness in covariate

has drawn attention. For some studies, budget limitations can cause missing covariates when a covariate of interest is too expensive to be measured for all participants in a study. In addition, study participants might not want to answer to some sensitive questions. In those cases, researchers use an auxiliary variable, which contains information about covariates of interest and is relatively cheap and available. To address this issue, Pepe and Fleming (1991) and Carroll and Wand (1991) proposed estimated likelihood methods for a binary auxiliary variable and a continuous auxiliary variable for a covariate, respectively. They showed how to use a auxiliary variable when the likelihood function has terms that have to be estimated in a non-parametrical way. They also showed how to use an auxiliary variable when the likelihood function has terms that have to be estimated in a non-parametical way.

In this article, we are interested in not only ODS design but also missing in covariate(s). Compared to the ODS design in Zhou et al. (2002), we consider the missing data problem by using a binary auxiliary variable for a covariate. In many observational studies, researchers might meet missing-in covariate data including auxiliary information. Thus, we consider a more realistic situation than the one posed in Zhou et al. (2002). To solve the problem of missing-in covariates, we assume that there is a binary auxiliary variable for a covariate of interest. Compared to Pepe and Fleming (1991), who proposed an estimated distribution function for a distribution of a covariate given a binary surrogate covariate under SRS scheme, we extend the estimated likelihood method to an ODS design. A motivating example is a cancer biomarker study. In a cancer study, the epidermal growth factor receptor (EGFR) mutations are used as a predictive biomarkers, and a subject response to EGFR inhibitor as a response variable. Since genotyping EGFR genes is a high-cost procedure, to identify the genotype of EGFR genes for all samples is difficiult. However, the likelihood score of EFGR mutations obtained by a designed questionnaire can be used as an auxiliary variable. Similarly, in the CPP study in Zhou et al. (2002) and Zhou, Wu, Liu and Cai (2011) , they applied an ODS scheme and used the socioeconomic status of the child's family (SES) as an auxiliary variable for a covariate of interest, Polychlorinated Biphenyls (PCB).

In this Chapter, with the continuous outcome variable which is assumed to be relatively easy or inexpensive to achieve, we study how to investigate the relationship between continuous responses and covariates that have missingness in covariates. To solve this problem, we use a binary auxiliary variable to extract more information about a covariate of interest. Compared to Zhou et al. (2002), we have a binary auxiliary variable for a covariate of interest. A finite number of strata based on continuous response will be defined. We propose an estimated likelihood approach that accounts for an ODS design and a binary auxiliary variable. To be specific, we focus on how to exploit auxiliary information in case of missing data under ODS design; the proposed method is semi-parametric, since the conditional distribution of covariates given the binary auxiliary variable is unspecified in any parametric model. In Section 3.2, we describe the data structure of a missing data problem under ODS design and derive a likelihood function. In Section 3.3, we propose an estimated likelihood method and Section 3.4 provides asymptotic results from the proposed estimators. Section 3.5 depicts the simulation results under different settings, and Section 3.6 shows a real data application. A brief discussion is given in Section 3.7.

3.2 Data structure and Likelihood Functions

3.2.1 Notations and Data structure

Let Y be the continuous outcome variable, and let X denote the vector of covariates. Assume that we also have a binary auxiliary variable, W, for a covariate in X. We consider an ODS design as follows : the domain of Y consists of 3 mutually exclusive intervals : $C_1 \cup C_2 \cup C_3 = (-\infty, c_1] \cup (c_1, c_3] \cup (c_3, \infty)$, where c_1 and c_3 are fixed constants. The ODS sample of size N consists of three parts : SRS sample of size N_0; a supplemental ODS sample of size N_1 from the left tail of $Y's$ distribution, C_1; and another supplemental ODS sample of size N_3 from the right tail of $Y's$ distribution, C_3. Recall that we are interested in using a binary auxiliary variable, W. Denote V as a validations set that consists of complete observations and $\overline{V}$ as a non-validation set that consists of incomplete observations that have missing-in covariate. Thus, we can define V_0 as the validation set from SRS sample, $\overline{V}_0$ as the non-validation set from

SRS sample, V_1 as the validation set of supplemental ODS sample from the left tail, $\overline{V}_1$ as the non-validation set of supplemental ODS sample from the left tail, V_3 as the validation set of supplemental ODS sample from the right tail, and $\overline{V}_3$ as the non-validation set of supplemental ODS sample from the right tail. Hence, we can summarize the total ODS sample as follows :

SRS sample:	Complete	$\{Y_i, X_i, W_i : i \in V_0\}$
	Incomplete	$\{Y_j, W_j : j \in \overline{V}_0\}$
ODS sample from the left tail:	Complete	$\{Y_i, X_i, W_i \vert Y_i \leq c_1 : i \in V_1\}$
	Incomplete	$\{Y_j, W_j \vert Y_j \leq c_1 : j \in \overline{V}_1\}$
ODS sample from the right tail:	Complete	$\{Y_i, X_i, W_i \vert Y_i \geq c_3 : i \in V_3\}$
	Incomplete	$\{Y_j, W_j \vert Y_j \geq c_3 : j \in \overline{V}_3\}$

Note that the total number of the ODS sample, $N = N_0 + N_1 + N_3$ and $N_0 = n_0 + \overline{n}_0$ and $N_k = n_k + \overline{n}_k$ for $k = 1, 3$ where n_l denotes the size of V_l and $\overline{n}_l$ denotes the size of $\overline{V}_l$ for $l = 0, 1, 3$. Moreover, since every subject has $W = 0$ or 1 as a binary auxiliary variable, we can stratify sets as follows: $V_0 = V_{0,w=0} \cup V_{0,w=1}$; $\overline{V}_0 = \overline{V}_{0,w=0} \cup \overline{V}_{0,w=1}$ where $V_{0,w=l}, l = 0, 1$ denotes the complete SRS with $W = l$. Similarly, $V_k = V_{k,w=0} \cup V_{k,w=0}$ where $V_{k,w=l}$ for $l = 0, 1$ denotes the validation set from *k*th stratum with $W = l$; $\overline{V}_k = \overline{V}_{k,w=0} \cup \overline{V}_{k,w=1}$ for $k = 1, 3$ where $\overline{V}_{k,w=l}$ for $l = 0, 1$ denotes the non-validation set from *k*th stratum with $W = l$.

In the next section, we derive an estimated likelihood method which reflects the ODS design with auxiliary information. To be specific, we focus on how to use auxiliary information on missing data problem with missing covariate values in the ODS design. We also give asymptotic properties about consistency, normality, and consistent covariance matrix estimator of the proposed method.

3.2.2 Likelihood Functions

This section describes how to construct the likelihood for dataset explained in the previous section. Based on the data structure in Section 3.1, we can construct the full likelihood for the ODS sample as

$$\begin{aligned} L(\beta) &= L_{SRS}(\beta)\cdot\overline{L}_{SRS}(\beta)\cdot L_{ODS}(\beta)\cdot\overline{L}_{ODS}(\beta) \\ &= L_{SRS}(\beta)\cdot\overline{L}_{SRS}(\beta)\cdot L_{ODS_{Left}}(\beta)\cdot\overline{L}_{ODS_{Left}}(\beta)\cdot L_{ODS_{Right}}(\beta)\cdot\overline{L}_{ODS_{Right}}(\beta) \\ &= \prod_{i=1}^{n_0} P(y_i,x_i)\prod_{j=1}^{\overline{n}_0} P(y_j,w_j)\times\prod_{i=1}^{n_1} P(y_i,x_i|y_i\le c_1)\prod_{j=1}^{\overline{n}_1} P(y_j,w_j|y_j\le c_1)\times \\ &\quad \prod_{i=1}^{n_3} P(y_i,x_i|y_i\ge c_3)\prod_{j=1}^{\overline{n}_3} P(y_j,w_j|y_j\ge c_3) \end{aligned} \tag{3.1}$$

Note that, in (3.1), $L_{(\cdot)}(\beta)$ is referring to the likelihood function that consists of complete observations, and $\overline{L}_{(\cdot)}(\beta)$ is referring to the one that consists of incomplete observations.

Now, we decompose each term in (3.1) to derive an estimated likelihood for the full likelihood function. First, consider likelihood functions in the SRS part. Let n_0 be complete observations from SRS and let $\overline{n}_{0,w=0}$ be the number of incomplete observations having $w = 0$ and let $\overline{n}_{0,w=1}$ be the number of incomplete observations having $w = 1$ in SRS sample. Thus, we can write the likelihood function for the SRS sample as follows :

$$\begin{aligned} L_{SRS}(\beta) &= \prod_{i=1}^{n_0} P(y_i,x_i) = \prod_{i=1}^{n_0} f_\beta(y_i|x_i)g(x_i) \propto \prod_{i=1}^{n_0} f_\beta(y_i|x_i) \\ \overline{L}_{SRS}(\beta) &= \overline{L}_{SRS_{w=0}}\overline{L}_{SRS_{w=1}} = \prod_{j=1}^{\overline{n}_{0,w=0}} P(y_j,w_j=0)\prod_{j=1}^{\overline{n}_{0,w=1}} P(y_j,w_j=1) \\ &\propto \prod_{j=1}^{\overline{n}_{0,w=0}} g_\beta(y_j|w_j=0)\prod_{j=1}^{\overline{n}_{0,w=1}} g_\beta(y_j|w_j=1), \end{aligned} \tag{3.2}$$

where $g_\beta(y_j|w_j=l) = \int f_\beta(y_i|x,w_j)g(x|w_j)dx$.

Next, consider the likelihood of supplemental ODS samples from left and right tails. Let n_1

be the size of complete ODS supplemental sample from the left tail and let n_3 be the size of complete ODS supplemental sample from the right tail. The likelihood function for the complete ODS supplemental sample from the left and the right tails can be written as follows:

$$\begin{aligned} L_{ODS}(\beta) &= L_{ODS_{Left}}(\beta) \times L_{ODS_{Right}}(\beta) = \prod_{i=1}^{n_1} P(y_i, x_i | y_i \leq c_1) \times \prod_{i=1}^{n_3} P(y_i, x_i | y_i \geq c_3) \\ &\propto \prod_{i=1}^{n_1} \frac{f_\beta(y_i|x_i)}{P(Y_i \leq c_1)} \times \prod_{i=1}^{n_3} \frac{f_\beta(y_i|x_i)}{P(Y_i \geq c_3)} \end{aligned} \tag{3.3}$$

where $P(Y_i \leq c_1)$ is the probability of the *i*th observation is drawn from the left tail, $Y_i \leq c_1$, and $P(Y_i \geq c_3)$ is the probability of the *i*th observation is drawn from the right tail, $Y_i \geq c_3$.

Now, the likelihood for incomplete supplemental ODS sample from the left tail is derived. Let $\overline{n}_{1,w=l}$ for $l = 0, 1$ be the number of incomplete observations from the supplemental sample having $w = l$ from the left tail. Then, we can construct the likelihood function for the ODS supplemental sample having missing-in covariate from the left tail as follows :

$$\overline{L}_{ODS_{left}}(\beta) = \prod_{l=0,1} \prod_{j=1}^{\overline{n}_{1,w=l}} P(y_j, w_j = l | Y_j \leq c_1) \propto \prod_{l=0,1} \prod_{j=1}^{\overline{n}_{1,w=l}} \frac{g(y_j|w_j = l)}{P(y_j \leq c_1)}, \tag{3.4}$$

where $k(w_j = l) = P(W_j = 0)$. Thus, the likelihood function for the ODS supplemental sample having missingness in covariate from the right tail is

$$\overline{L}_{ODS_{right}}(\beta) = \prod_{l=0,1} \prod_{j=1}^{\overline{n}_{3,w=l}} P(y_j, w_j = l | Y_j \geq c_3) \propto \prod_{l=0,1} \prod_{j=1}^{\overline{n}_{3,w=l}} \frac{g(y_j|w_j = l)}{P(y_j \geq c_3)}. \tag{3.5}$$

Combining terms from (3.2) to (3.5), the full likelihood for ODS sample is proportional to

$$L(\beta) \propto \left\{ \prod_{i \in V} f_\beta(y_i|x_i) \right\} \left\{ \prod_{l=0,1} \prod_{j \in \overline{V}_{w=l}} g(y_j|w_j = l) \right\} \left[P(y \leq c_1) \right]^{-N_1} \left[P(y \geq c_3) \right]^{-N_3}$$

where $N_k = n_k + \overline{n}_{k,w=0} + \overline{n}_{k,w=1}$ for $k = 1, 3$. By taking log-transformation on $L(\beta)$, we can

construct the full log-likelihood for ODS sample as

$$l(\beta) \propto \sum_{i \in V} \log f_\beta(y_i|x_i) + \sum_{j \in \overline{V}_{w=0}} \log g(y_j|w_j = 0) + \sum_{j \in \overline{V}_{w=1}} \log g(y_j|w_j = 1)$$
$$- N_1 \cdot \log P(y \le c_1) - N_3 \cdot \log P(y \ge c_3). \tag{3.6}$$

3.3 An Estimated Likelihood Estimator

In this section, we describe an estimated likelihood method for the ODS design with missing in covariate. First, from the log-likelihood in (3.6), we first estimate the density function, $g(y_j|w_j = l)$, for $l = 0, 1$. By Bayes rule, we can write $g(y_j|w_j = l)$ as

$$g(y_j|w_j = l) = \int f_\beta(y_j|x) dG_{x|w_j=l}, \; g(y_j|w_j = l) = \int f_\beta(y_j|x) dG_{x|w_j=l}$$

where $g(x|w)$ and $G_{x|w}$ are the distribution and the distribution function of X given W, respectively.

Thus, we need to estimate the distribution function of $X|W$ while accounting for ODS design and auxiliary information. By the law of total probability and Bayes rule,

$$G_{X|W=l}(x|W = l) = \sum_{k=1}^{3} P(c_{k-1} < Y < c_k|W = l) P(X \le x|c_{k-1} < Y < c_k, W = l).$$

One can estimate $P(c_{k-1} < Y < c_k|W = l)$ as

$$\widehat{P}(c_{k-1} < Y < c_k|W = l) = \frac{\widehat{P}(c_{k-1} < Y \le c_k, W = l)}{\widehat{P}(W = l)} = \frac{N_{0,k,w=l}/N_0}{N_{0,w=l}/N_0} = \frac{N_{0,k,w=l}}{N_{0,w=l}},$$

where $N_{0,w=l}$ is the number of subjects in the set, $\{i : W_i = l \text{ where } i \in V_0 \cup \overline{V}_0\}$, and $N_{0,k,w=l}$ denotes the number of subjects in the set, $\{i : c_{k-1} < Y_i \le c_k, W_i = l \text{ where } i \in V_0 \cup \overline{V}_0\}$.

Next, one can estimate $P(X \le x | c_{k-1} < Y \le c_k, W = l)$ for $k = 1, 2, 3$ as

$$\hat{P}(X \le x | c_{k-1} < Y \le c_k, W = l) = \frac{\hat{P}(X \le x, Y \le c_1, W = l)}{\hat{P}(Y \le c_1, W = l)} = \frac{\sum_{i \in V_{0,k,w=l}} I(X_i \le x)}{n_{0,k,w=l}},$$

where the size of the set $V_{0,k,w=l} = \{i : c_{k-1} < Y_i \le c_k, W_i = l, R_i = 1 \text{ where } i \in V_0 \cup \overline{V}_0\}$ is written as $n_{0,k,w=l}$, and $R_i = 1$ implies the *i*th observation is complete. Hence, the distribution function of $X|W$ can be estimated as

$$\widehat{G}(x|W = l) = \sum_{k=1}^{3} \frac{N_{0,k,w=l}}{N_{0,w=l}} \sum_{i \in V_{0,k,w=l}} \frac{I(X_i \le x)}{n_{0,k,w=l}},$$

where $V_{0,k,w=l}$ is the set of complete SRS observations in *k*th stratum having $W = l$ for $l = 0, 1$ and $n_{0,k,w=l}$ is the size of $V_{0,k,w=l}$. Thus, we can propose an estimated p.d.f for $g(y_j|w_j = l)$ for $k = 1, 2, 3$ and $l = 0, 1$ as follows :

$$\hat{g}(y_j|w_j = l) = \int f_\beta(y_j|x) d\hat{G}_{x|w_j=l} = \sum_{k=1}^{3} \frac{N_{0,k,w=l}}{N_{0,w=l}} \frac{\sum_{i \in V_{0,k,w=l}} f_\beta(y_j|x_i)}{n_{0,k,w=l}}$$

On the other hand, $P(y \le c_k)$ can be written

$$P(y \le c_k) = \int F_\beta(c_1|x) dG(x),$$

where $F_\beta(c_1|x) = \int_{-\infty}^{c_1} f_\beta(y|x) dy$. By the law of total probability and Bayes rule,

$$G_X(x) = \sum_{k=1}^{3} P(c_{k-1} < Y \le c_k) P(X \le x | c_{k-1} < Y \le c_k).$$

One can estimate each component in $G_X(x)$ as

$$\widehat{P}(c_{k-1} < Y \le c_k) = \frac{N_{0,k}}{N_0}, \quad \widehat{P}(X \le x | c_{k-1} < Y \le c_k) = \frac{\sum_{i \in V_{0,k}} I(X_i \le x)}{n_{0,k}}.$$

By substituting $\widehat{P}(c_{k-1} < Y \leq c_k)$ and $\widehat{P}(X \leq x | c_{k-1} < Y \leq c_k)$ into $G_X(x)$, we could estimate $G(x)$ as

$$\widehat{G}_X(x) = \sum_{l=1}^{3} \frac{N_{0,l}}{N_0} \frac{\sum_{i \in V_{0,l}} I(X_i \leq x)}{n_{0,l}}.$$

Hence,

$$\hat{P}(y \leq c_k) = \int F_\beta(c_1|x) d\hat{G}(x) = \sum_{l=1}^{3} \frac{N_{0,l}}{N_0} \frac{\sum\limits_{i \in V_{0,l}} F_\beta(a_1|x_i)}{n_{0,l}}, \text{ for } k = 1, 3.$$

By substituting $\hat{g}(y_j|w_j = l)$ for $l = 0, 1$, $\hat{P}(Y \leq c_1)$, and $\hat{P}(Y \geq c_3)$ into $lnL(\beta)$, we obtain the following estimated log-likelihood function for β :

$$\begin{aligned} \ln \widehat{L}(\beta) = & \sum_{i \in V} \log f_\beta(y_i|x_i) + \sum_{j \in \overline{V}_{w=0}} \log \hat{g}(y_j|w_j = 0) \\ & + \sum_{j \in \overline{V}_{w=1}} \log \hat{g}(y_j|w_j = 1) - \sum_{k=1,3} N_k \cdot \log \widehat{P}(y \in C_k), \end{aligned}$$

where $N_k = (n_k + \overline{n}_{k,w=0} + \overline{n}_{k,w=1})$ for $k = 1, 3$. Note that in $\ln \widehat{L}(\beta)$, the index i corresponds to the complete data and index j corresponds to incomplete data, respectively. Now, we construct the score equation to obtain a Maximum Estimated Likelihood Estimator (MELE). Taking derivatives on equation $\ln \widehat{L}(\beta)$ with respect to β, we could obtain the score equation for the estimated likelihood,

$$\begin{aligned} 0 = \hat{U}(\beta) = & \sum_{i \in V} \frac{\partial f_\beta(y_i|x_i)/\partial \beta}{f_\beta(y_i|x_i)} - \sum_{k=1,3} N_k \cdot \frac{\partial \hat{P}(y \in C_k)/\partial \beta}{\hat{P}(y \in C_k)} \\ & + \sum_{j \in \overline{V}_{w=0}} \frac{\partial \hat{g}(y_j|w_j = 0)/\partial \beta}{\hat{g}(y_j|w_j = 0)} + \sum_{j \in \overline{V}_{w=1}} \frac{\partial \hat{g}(y_j|w_j = 1)/\partial \beta}{\hat{g}(y_j|w_j = 1)}. \end{aligned} \tag{3.7}$$

The equation (3.7) can be solved by using the Newton-Raphson algorithm and $\hat{\beta}_p$ is defined as a solution to the equation above.

3.4 Asymptotic Results

We assume that, as $N \to \infty$, $N_{0,k,w=0}/N_{0,w=0} \to_p P(Y \in C_k|W = 0) = a_{0k}$, $N_{0,k,w=1}/N_{0,w=1} \to_p P(Y \in C_k|W = 1) = a_{1k}$, $n_{0,k,w=0}/N \to_p b_{0k}$ $n_{0,k,w=1}/N \to_p b_{1k}$ $\overline{n}_{T,k,w=0}/N_{k,w=0} \to_p c_{0k}$ $\overline{n}_{T,k,w=1}/N_{k,w=1} \to_p c_{1k}$. Define $\overline{n}_{T,w=0} = \overline{n}_{0,w=0} + \overline{n}_{1,w=0} + \overline{n}_{3,w=0}$ and $\overline{n}_{T,w=1} = \overline{n}_{0,w=1} + \overline{n}_{1,w=1} + \overline{n}_{3,w=1}$. β^* denotes the true parameter value. Theorem 4 provides the consistency and Theorem 5 gives the asymptotic normality of $\hat{\beta}_p$. Theorem 6 establishes a consistent estimator for the asymptotic covariance matrix in Theorem 5.

Theorem 4. *(Consistency of $\hat{\beta}_p$) Under the regular conditions in the Appendix, as $N \to \infty$, a sequence of $\{\hat{\beta}_p\}$ of solutions to the estimated score equations, $\hat{U}(\beta)$, converges to β^* with probability 1. If another sequence of $\{\tilde{\beta}\}$ of solutions to the equation, (3.7), such that $\tilde{\beta} \to_p \beta^*$, then $\tilde{\beta} = \hat{\beta}_p$ with probability 1 as $N \to \infty$.*

Theorem 5. *(Asymptotic Normality of $\hat{\beta}_p$) Under the regular conditions in the Appendix, as $N \to \infty$, $\sqrt{N}\left(\hat{\beta}_p - \beta^*\right)$ converges weakly to a normal distribution with mean zero and covariance $\Sigma(\beta^*)$, where*

$$\Sigma(\beta^*) = I_{\beta^*}^{-1} + \sum_{l=1}^{3}\left\{\frac{(a_{0l})^2}{b_{0l}} I_{\beta^*}^{-1}\Sigma_{l0}I_{\beta^*}^{-1}\right\} + \sum_{l=1}^{3}\left\{\frac{(a_{1l})^2}{b_{1l}} I_{\beta^*}^{-1}\Sigma_{l1}I_{\beta^*}^{-1}\right\}, \tag{3.8}$$

where

$$\begin{aligned} I(\beta)^{-1} &= -E\left[\frac{\partial^2 lnL(\beta)}{\partial\beta\partial\beta^T}\right], \\ \Sigma_{l,w=l} &= \mathrm{Var}_{X|Y\in C_{l,W=l}}\left\{\sum_{k=1}^{3} c_{lk} \times E_{Y|Y\in C_{k,w=l}}[M_X(Y{:}\beta)]\right\} \end{aligned}$$

with

$$M_X(Y:\beta) = \left(\frac{\partial f(Y|X;\beta)}{g(Y|W=0;\beta)} - \frac{\partial g(Y|W=0;\beta)/\partial\beta}{\{g(Y|W=0;\beta)\}^2} f(Y|X;\beta)\right).$$

The proofs of Theorem 4 and 5 are provided in the Appendix. The consistent variance

estimator is provided in the following theorem.

Theorem 6. *A consistent estimator for the covariance matrix in* $\Sigma(\beta^*)$ *is*

$$\widehat{\sum}(\hat{\beta}_p) = \hat{I}^{-1}(\hat{\beta}_p) + \sum_{l=1}^{3} \frac{(N_{0,k,w=0}/N_{0,w=0})^2}{(n_{0,k,w=0}/N)} \hat{I}^{-1}(\hat{\beta}_p)\hat{\Sigma}_{l0}\hat{I}^{-1}(\hat{\beta}_p) + \sum_{l=1}^{3} \frac{(N_{0,k,w=1}/N_{0,w=1})^2}{(n_{0,k,w=1}/N)} \hat{I}^{-1}(\hat{\beta}_p)\hat{\Sigma}_{l1}\hat{I}^{-1}(\hat{\beta}_p), \tag{3.9}$$

where

$$\hat{I}^{-1}(\hat{\beta}_p) = \frac{-\partial\hat{U}(\beta)/\partial\beta^T}{N},$$

$$\widehat{\Sigma}_{l,w=l} = \widehat{\mathrm{Var}}_{\{X_i:i\in V_{0,l,w=l}\}}\left\{\sum_{k=1}^{K} \frac{\overline{n}_{T,k,w=l}}{N}\widehat{\overline{M}}_{k,w=l}(X_i:\beta)\right\},$$

$$\widehat{\overline{M}}_{k,w=l}(X_i:\beta) = \sum_{j\in\overline{V}_{T,k,w=l}}\left[\left(\frac{\partial f(y_j|x_i:\beta)}{\hat{g}(y_j|w_j=l)} - \frac{f(y_j|x_i:\beta)\partial\hat{g}(y_j|w_j=l)/\partial\beta}{\{\hat{g}(y_j|w_j=l)\}^2}\right)/\overline{n}_{T,k,w=l}\right],$$

for $l = 0, 1$*, such that* $\widehat{\Sigma}(\beta) \to \Sigma(\beta)$ *in parobability as* $N \to \infty$*.*

$\widehat{\sum}(\hat{\beta}_p)$ needs to be calculated once after final step of iteration in Newton-Raphson method.

3.5 Simulation study

Simulation is conducted to evaluate the performance of the proposed method in finite samples. We generated data from a linear regression model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon, \tag{3.10}$$

where $X \sim N(0,1)$, $Z \sim Bernoulli(0.45)$, and $\varepsilon \sim N(0,\sigma^2)$.

To generate the auxiliary variable W for X, we generate $W = I(X + e > c)$ as an auxiliary variable for X, where $e \sim N(0, \sigma_e^2)$ and c is fixed constant. Mutual independence between X, ε and e is assumed. Thus, the conditional distribution Y given X and Z is normal with mean $\beta_0 + \beta_1 X + \beta_2 Z$ and variance σ^2. On the other hand, we also generate missing index, R, where $R \sim Bernoulli(p)$, which is independent of other variables. In the following tables, we assume that Y and W are observed for all N subjects in a study population but X can be observed only for the validation portions according to missing rate, p. We fix $\beta_0 = 1$, $\beta_2 = -0.5$, $c = 0$ and $\sigma^2 = 1$, and vary $\beta_1 = 0.1$ and 0.5, $(c_1, c_3) = (\mu_Y - \sigma_Y,\ c_3 = \mu_Y + \sigma_Y)$ and (1st quartile of Y, 3rd quartile of Y). The simulation results are depicted in Table 3.1 and 3.2. Under $N = 10,000$ and $n_v = 800$, the estimate means, standard errors, mean of the variance estimated, and 95% nominal confidence intervals (CIs) for each estimator are obtained from 1,000 independent simulation runs. Under the fixed n_v, we vary portions of SRS and supplementary ODS sample to see efficiencies according to different portions between SRS and supplementary ODS samples. Simulation studies are conducted under 30% and 50% missingness in ODS sample.

We compare the proposed method, $\hat{\beta}_p$ with three existing estimators : (a) The first estimator is the maximum likelihood estimator, which is denoted by $\hat{\beta}_S$ based on the SRS with the same sample size with the proposed method; (b) The second estimator is the semiparametric empirical likelihood estimator in Zhou et al. (2002), which is denoted by $\hat{\beta}_Z$, only with a validation sample from ODS design in the proposed method; (c) The third estimator is the probability weighted estimator in Horvitz and Thompson (1952), which is denote by $\hat{\beta}_{IPW}$, using the observed sampling weight in a study population. Note that $\hat{\beta}_S$ is a benchmark to investigate the performance of the proposed method with the sample size, n_v, assuming no missing observations. $\hat{\beta}_Z$ is conducted with complete observations from ODS sample to compare efficiencies of $\hat{\beta}_P$ that uses missing covariate observations. For example, if we assume 50% of missingness in a sample with the size of 1,200, roughly, $\hat{\beta}_Z$ uses 600 subjects and $\hat{\beta}_P$ include 600 more subjects having missing covariate observations. $\hat{\beta}_{IPW}$ is obtained by calculating selection probabilities in the study population. The comparison against $\hat{\beta}_{IPW}$

demonstrates an efficiency gain when we use missing observations with their auxiliary variable with the same sample size of $\hat{\beta}_P$.

Through simulation studies, we could see interesting findings. First, all four methods are unbiased; no noticeable bias is observed in the means of $\hat{\beta}_S, \hat{\beta}_Z, \hat{\beta}_{IPW}$, or $\hat{\beta}_{P_1}$. In terms of efficiency, $\hat{\beta}_{P_1}$ is more efficient than $\hat{\beta}_Z$ and $\hat{\beta}_{IPW}$. This fact shows that including more missing covariate observations with auxiliary information can enhance efficiency. Comparing this against $\hat{\beta}_S$, $\hat{\beta}_{P_1}$ is more efficient under missing rate=0.3, that means even with some missing covariate observations, ODS design can be a more efficient option than simple random sampling. Note that even under missing rate=0.5, $\hat{\beta}_{p_1}$ has standard errors very close to those of $\hat{\beta}_S$. For example, in Table 3.1, under the setting of $\beta = 0.1$ and missing rate=0.3 with $(N_0, N_1, N_3) = (400, 200, 200)$, SE($\hat{\beta}_S$) / SE($\hat{\beta}_P$)=1.13. Under the setting of $\beta = 0.5$ and missing rate=0.3 with with $(N_0, N_1, N_3) = (400, 200, 200)$, SE($\hat{\beta}_S$) / SE($\hat{\beta}_{P_1}$)=1.03. Those comparisons show that with 30% of missing in covariate, $\hat{\beta}_P$ has very close efficiency to the one from the complete SRS of the same total sample size when $\beta = 0$. Lastly, we can see that the strength of relationship between a covariate of interest and auxiliary variable affects on efficiency. $\hat{\beta}_{p_3}$ denotes the estimator with the data set that has less stronger relationship between X and W. In Table 3.1 and 3.2, SE of $\hat{\beta}_{p_3}$ is larger than $\hat{\beta}_{p_2}$. This is the expected result since $\hat{\beta}_{p_3}$ uses weaker W than $\hat{\beta}_{p_2}$ does.

In addition, we apply the proposed estimator for datasets missing X assuming MAR(Missing At Random) defined by Rubin (1976). The same underlying model, (3.10), with $\beta_0 = 1$, is used for the study population. To generate missing covariate X under MAR, we use the following logistic regression model :

$$P(R = 1 \mid y) = \frac{\exp(\gamma_0 + \gamma_1 y)}{1 + \exp(\gamma_0 + \gamma_1 y)}, \tag{3.11}$$

where $R = 1$ if X is missing, $R = 0$, otherwise. The value of the intercept, γ_0, in (3.11) is chosen to simulate the situation of missing X such that the overall event rate depends on γ_0.

We show the results in Table 3.3 and 3.4 with varying allocations of ODS sample, (β_1, β_2), and missing rates. $\hat{\beta}_C$ uses complete observations from the simple random sample and the least square method is used to estimate β. $\hat{\beta}_{IPW}$ uses complete observations from the ODS sample and is estimated by the inverse probability weighted method in Weaver and Zhou (2005). $\hat{\beta}_p$ is the proposed estimator described in the previous sections. Note that overall missing rates are about 20% and 82% in Table 3.3 and 3.4, respectively. In Table 3.3, under $\beta_1 = 0$ and $\beta_2 = 0$, $\hat{\beta}_C$, $\hat{\beta}_{IPW}$, and $\hat{\beta}_p$ are also unbiased. However, under $\beta_1 = 0.5$ and $\beta_2 = -0.5$, $\hat{\beta}_C$ and $\hat{\beta}_{IPW}$ bring biased estimators since they do not reflect MAR mechanism. Under about 20% missing rate, $\hat{\beta}_p$ is unbiased, satisfies the asymptotic properties for the estimates for β_1 and β_2, and has smaller SE of estimates than those of $\hat{\beta}_C$ and $\hat{\beta}_{IPW}$. However, with about 82% missing rate, $\hat{\beta}_p$ is also biased and does not satisfy the asymptotic normality properties under MCAR mechanism.

Table 3.1: Simulation results for asymptotic properties in Section 3.4. Results are based on 1,000 simulations with (N_0, N_1, N_3) = $(400, 200, 200)$ and various missing rates in X

Missing %	β_1	Method	Mean	SE	$\widehat{\text{SE}}$	CI	β_2	Mean	SE	$\widehat{\text{SE}}$	CI
0.3	0.1	$\hat{\beta}_s$	0.098	0.034	0.035	0.954	-0.5	-0.500	0.073	0.071	0.950
		$\hat{\beta}_{IPW}$	0.101	0.037	0.038	0.956		-0.499	0.077	0.077	0.954
		$\hat{\beta}_z$	0.100	0.032	0.033	0.951		-0.501	0.066	0.068	0.956
		$\hat{\beta}_{p1}$	0.099	0.032	0.032	0.955		-0.501	0.068	0.071	0.960
		$\hat{\beta}_{p2}$	0.101	0.030	0.029	0.944		-0.498	0.067	0.065	0.943
		$\hat{\beta}_{p3}$	0.102	0.032	0.031	0.945		-0.502	0.067	0.065	0.939
0.5		$\hat{\beta}_s$	0.098	0.034	0.035	0.954		-0.500	0.073	0.071	0.950
		$\hat{\beta}_{IPW}$	0.101	0.047	0.046	0.941		-0.508	0.096	0.092	0.941
		$\hat{\beta}_z$	0.098	0.040	0.039	0.943		-0.499	0.083	0.080	0.947
		$\hat{\beta}_{p1}$	0.099	0.035	0.035	0.952		-0.501	0.077	0.083	0.963
		$\hat{\beta}_{p2}$	0.101	0.033	0.032	0.942		-0.497	0.077	0.077	0.952
		$\hat{\beta}_{p3}$	0.103	0.037	0.035	0.942		-0.502	0.077	0.077	0.956
0.3	0.5	$\hat{\beta}_S$	0.501	0.035	0.035	0.947		-0.501	0.072	0.071	0.942
		$\hat{\beta}_{IPW}$	0.501	0.041	0.042	0.948		-0.499	0.081	0.083	0.953
		$\hat{\beta}_{p1}$	0.499	0.034	0.035	0.951		-0.500	0.072	0.073	0.954
		$\hat{\beta}_{p2}$	0.500	0.034	0.033	0.944		-0.502	0.071	0.068	0.946
		$\hat{\beta}_{p3}$	0.502	0.036	0.034	0.936		-0.501	0.070	0.068	0.944
0.5		$\hat{\beta}_S$	0.501	0.035	0.035	0.947		-0.501	0.072	0.071	0.942
		$\hat{\beta}_{IPW}$	0.504	0.051	0.050	0.936		-0.507	0.102	0.098	0.933
		$\hat{\beta}_Z$	0.501	0.044	0.046	0.953		-0.500	0.083	0.084	0.947
		$\hat{\beta}_{p1}$	0.503	0.039	0.041	0.959		-0.499	0.086	0.086	0.952
		$\hat{\beta}_{p2}$	0.502	0.038	0.038	0.946		-0.499	0.083	0.080	0.948
		$\hat{\beta}_{p3}$	0.502	0.042	0.040	0.936		-0.500	0.082	0.081	0.945

$\hat{\beta}_{SRS}$, $\hat{\beta}_Z$, $\hat{\beta}_{IPW}$ are described in section 3.5

$\hat{\beta}_{p1}$: the proposed estimator using cut points of (1st quartile of Y, 3rd quartile of Y) with $W = I(X > 0)$

$\hat{\beta}_{p2}$: the proposed estimator using cut points of $(\mu_Y - \sigma_Y, \mu_Y + \sigma_Y)$ with $W = I(X > 0)$

$\hat{\beta}_{p3}$: the proposed estimator using cut points of $(\mu_Y - \sigma_Y, \mu_Y + \sigma_Y)$ with $W = I(X + N(0,1) > 0)$

Mean : mean of estimated quantities for β

SE : statndard error of estimated quantities for β

$\widehat{\text{SE}}$: estimate of SE

CI : 95% confidence interval

Table 3.2: Simulation results for asymptotic properties in Section 3.4. Results are based on 1,000 simulations with (N_0, N_1, N_3) = $(640, 80, 80)$ and various missing rates in X

Missing %	β_1	Method	Mean	SE	$\widehat{\text{SE}}$	CI	β_2	Mean	SE	$\widehat{\text{SE}}$	CI
0.3	0.1	$\widehat{\beta}_S$	0.098	0.034	0.035	0.954		-0.500	0.073	0.071	0.950
		$\widehat{\beta}_{IPW}$	0.099	0.037	0.038	0.966		-0.500	0.076	0.077	0.951
		$\widehat{\beta}_Z$	0.100	0.039	0.037	0.939		-0.500	0.075	0.075	0.960
		$\widehat{\beta}_{p1}$	0.100	0.035	0.035	0.949		-0.500	0.076	0.077	0.949
		$\widehat{\beta}_{p2}$	0.100	0.033	0.033	0.953		-0.500	0.072	0.074	0.962
		$\widehat{\beta}_{p3}$	0.099	0.035	0.035	0.950		-0.500	0.077	0.074	0.943
0.5		$\widehat{\beta}_S$	0.098	0.034	0.035	0.954		-0.500	0.073	0.071	0.950
		$\widehat{\beta}_{IPW}$	0.102	0.048	0.045	0.936		-0.502	0.093	0.091	0.946
		$\widehat{\beta}_Z$	0.099	0.045	0.044	0.948		-0.501	0.092	0.090	0.942
		$\widehat{\beta}_{p1}$	0.099	0.037	0.038	0.951		-0.499	0.089	0.090	0.954
		$\widehat{\beta}_{p2}$	0.100	0.036	0.036	0.949		-0.498	0.085	0.086	0.949
		$\widehat{\beta}_{p3}$	0.099	0.039	0.040	0.952		-0.501	0.089	0.086	0.940
0.3	0.5	$\widehat{\beta}_S$	0.501	0.035	0.035	0.947		-0.501	0.072	0.071	0.942
		$\widehat{\beta}_{IPW}$	0.500	0.040	0.040	0.948		-0.501	0.079	0.079	0.950
		$\widehat{\beta}_Z$	0.500	0.040	0.040	0.954		-0.498	0.078	0.078	0.946
		$\widehat{\beta}_{p1}$	0.499	0.035	0.037	0.952		-0.502	0.077	0.079	0.957
		$\widehat{\beta}_{p2}$	0.500	0.035	0.036	0.959		-0.500	0.072	0.076	0.961
		$\widehat{\beta}_{p3}$	0.500	0.035	0.037	0.957		-0.498	0.078	0.076	0.954
0.5		$\widehat{\beta}_S$	0.501	0.035	0.035	0.947		-0.501	0.072	0.071	0.942
		$\widehat{\beta}_{IPW}$	0.502	0.048	0.047	0.943		-0.503	0.093	0.093	0.948
		$\widehat{\beta}_Z$	0.502	0.046	0.048	0.958		-0.499	0.094	0.092	0.939
		$\widehat{\beta}_{p1}$	0.500	0.038	0.041	0.964		-0.500	0.089	0.092	0.958
		$\widehat{\beta}_{p2}$	0.501	0.038	0.040	0.964		-0.499	0.085	0.088	0.953
		$\widehat{\beta}_{p3}$	0.500	0.040	0.042	0.958		-0.499	0.091	0.089	0.946

$\hat{\beta}_{SRS}$, $\hat{\beta}_Z$, $\hat{\beta}_{IPW}$ are described in section 3.5

$\hat{\beta}_{p1}$: the proposed estimator using cut points of (1st quartile of Y, 3rd quartile of Y) with $W = I(X > 0)$

$\hat{\beta}_{p2}$: the proposed estimator using cut points of $(\mu_Y - \sigma_Y, \mu_Y + \sigma_Y)$ with $W = I(X > 0)$

$\hat{\beta}_{p3}$: the proposed estimator using cut points of $(\mu_Y - \sigma_Y, \mu_Y + \sigma_Y)$ with $W = I(X + N(0,1) > 0)$

Mean : mean of estimated quantities for β

SE : statndard error of estimated quantities for β

$\widehat{\text{SE}}$: estimate of SE

CI : 95% confidence interval

Table 3.3: Simulation results for asymptotic properties in Section 3.4. Results are based on 1,000 simulations with (N_0, N_1, N_3) under MAR assumption on missing X with about 20% missing rate

(N_0, N_1, N_3)	β_1	Method	Mean	SE	$\widehat{\text{SE}}$	CI	β_2	Mean	SE	$\widehat{\text{SE}}$	CI
$(800, 200, 200)$	0	$\widehat{\beta}_C$	0.001	0.030	0.031	0.959	0	-0.000	0.067	0.064	0.937
		$\widehat{\beta}_{IPW}$	-0.000	0.036	0.035	0.930		0.003	0.073	0.071	0.950
		$\widehat{\beta}_p$	0.000	0.027	0.026	0.947		0.003	0.054	0.058	0.958
	0.5	$\widehat{\beta}_C$	0.469	0.031	0.031	0.837	-0.5	-0.467	0.066	0.063	0.910
		$\widehat{\beta}_{IPW}$	0.553	0.033	0.033	0.646		-0.553	0.071	0.069	0.873
		$\widehat{\beta}_p$	0.498	0.030	0.029	0.943		-0.492	0.060	0.059	0.945
$(600, 300, 300)$	0	$\widehat{\beta}_C$	0.001	0.030	0.031	0.959	0	-0.000	0.067	0.064	0.937
		$\widehat{\beta}_{IPW}$	-0.000	0.039	0.038	0.945		0.003	0.079	0.077	0.950
		$\widehat{\beta}_p$	0.001	0.025	0.025	0.942		0.002	0.050	0.054	0.963
	0.5	$\widehat{\beta}_C$	0.469	0.031	0.031	0.837	-0.5	-0.467	0.066	0.063	0.910
		$\widehat{\beta}_{IPW}$	0.569	0.037	0.036	0.525		-0.571	0.076	0.076	0.843
		$\widehat{\beta}_p$	0.500	0.030	0.028	0.930		-0.495	0.056	0.057	0.947

$\hat{\beta}_C$, $\hat{\beta}_{IPW}$ are described in Section 3.5

$\hat{\beta}_p$ and $\hat{\beta}_{IPW}$ use cut-off points of $(\mu_Y - \sigma_Y, \mu_Y + \sigma_Y)$

Missing in X is generated with the model, $P(R = 1 \mid y) = \frac{\exp(-2+0.7y)}{1+\exp(-2+0.7y)}$

Mean : mean of estimated quantities for β

SE : standard error of estimated quantities for β

$\widehat{\text{SE}}$: estimate of SE

CI : 95% confidence interval

Table 3.4: Simulation results for asymptotic properties in Section 3.4. Results are based on 1,000 simulations with (N_0, N_1, N_3) under MAR assumption on missing X with about 82% missing rate

(N_0, N_1, N_3)	β_1	Method	Mean	SE	$\widehat{\text{SE}}$	CI	β_2	Mean	SE	$\widehat{\text{SE}}$	CI
$(800, 200, 200)$	0	$\widehat{\beta}_C$	0.003	0.065	0.066	0.963		-0.009	0.136	0.133	0.940
		$\widehat{\beta}_{IPW}$	0.001	0.067	0.068	0.957		0.008	0.138	0.138	0.946
		$\widehat{\beta}_p$	0.001	0.036	0.043	0.988		0.001	0.106	0.154	0.989
	0.5	$\widehat{\beta}_C$	0.462	0.061	0.062	0.914		-0.467	0.125	0.123	0.933
		$\widehat{\beta}_{IPW}$	0.495	0.060	0.061	0.951		-0.490	0.128	0.125	0.946
		$\widehat{\beta}_p$	0.483	0.051	0.064	0.980		-0.420	0.113	0.169	0.981
$(600, 300, 300)$	0	$\widehat{\beta}_C$	0.003	0.065	0.066	0.963		-0.009	0.136	0.133	0.940
		$\widehat{\beta}_{IPW}$	0.001	0.072	0.071	0.934		0.009	0.144	0.144	0.949
		$\widehat{\beta}_p$	0.000	0.034	0.041	0.979		0.004	0.114	0.146	0.975
	0.5	$\widehat{\beta}_C$	0.462	0.061	0.062	0.914		-0.467	0.125	0.123	0.933
		$\widehat{\beta}_{IPW}$	0.490	0.065	0.064	0.942		-0.489	0.134	0.130	0.942
		$\widehat{\beta}_p$	0.486	0.055	0.066	0.969		-0.417	0.122	0.174	0.972

$\hat{\beta}_C$, $\hat{\beta}_{IPW}$ are described in Section 3.5

$\hat{\beta}_p$ and $\hat{\beta}_{IPW}$ use cut-off points of $(\mu_Y - \sigma_Y, \mu_Y + \sigma_Y)$

Missing in X is generated with the model, $P(R = 1 \mid y) = \frac{\exp(1+0.7y)}{1+\exp(1+0.7y)}$

Mean : mean of estimated quantities for β

SE : standard error of estimated quantities for β

$\widehat{\text{SE}}$: estimate of SE

CI : 95% confidence interval

3.6 Application to the PCB Data

We apply the proposed method to fit a data set from the Collaborative Perinatal Project(CPP, Niswander and Gordon, 1972). From 1959 to 1965, more than 44,000 women were enrolled into the study in 12 medical centers in 11 cities in the U.S. 55,908 pregnancies resulted in with multiple pregnancies. Researchers collected data at each prenatal visit of mother and delivery. The children born into the study were followed with several outcomes up to 8 years. Mothers' non-fasting blood was collected at each prenatal visit and delivery, and sera were collected and stored for later analyses.

In the environmental epidemiological studies in Gray et al. (2005) and Zhou et al. (2002), investigators studied the effect of mothers' maternal pregnancy serum level of polychlorinated biphenyls (PCBs) on cognitive test scores (IQ) at 7 years of age on the Wechsler Intelligence Scale for Children (WISC). Participants were enrolled through university-affiliated medical clinics and data were collected from each participant at each visit. One of the more interesting findings of was that the PCB levels have a relationship to IQ test performance. To investigate the in utero exposure of PCBs in relation to neurodevelopmental abnormalities, the PCBs levels were measured by analyzing the third-trimester blood serum specimens that had been preserved from mothers in the CPP study. In addition to PCB levels, other covariates, the socioeconomic status of the child's family (SES), gender (SEX), race (RACE) and the parent's education (EDU), were also collected.

To apply the proposed method to CPP study, we consider the following linear regression model:

$$IQ = \beta_0 + \beta_1 PCB + \beta_2 EDU + \beta_3 SES + \beta_4 RACE + \beta_5 SEX + \epsilon$$

In Zhou et al. (2002), there were 44,075 eligible children from the CPP. Among them, 38,709 had complete data for variables in the regression model above except PCBs. Originally, Zhou et al. (2002) tried to draw an ODS sample with $(N_0, N_1, N_3) = (1200, 200, 200)$ but obtained

(N_0, N_1, N_3) = $(849, 81, 108)$. From the dataset in Zhou et al. (2002), we take the simple random sample with size of 849 as the population in our real data application. PCB level is the covariate of interest and a binary auxiliary variable for PCB is generated as W = I(PCB > median of PCB). To generate missing data artificially, 849 of missing indicators are generated with probability 0.5: that is, we have roughly 50% of missing observations in the population data set with the size of 849. To apply the proposed ODS scheme to the population data set, the cut-off value for the left tail is set as 1 SD below the mean of IQ and the cut-off value for the right tail is 1 SD above the mean of IQ. We draw 300 of the SRS sample first, and then draw 50 of supplementary samples from the left and right tails given cut-off values, (81.44, 109.46), from the distribution of IQ score in the population data set.

The results with the complete portion of the ODS sample using MSELE in Zhou et al. (2002) are presented in Table 3.5. We compare results from our proposed method to those from MSELE. One benefit of our proposed method is that one can include more subjects that have incomplete data set, especially PCB here : the results from MSELE with complete portion in the ODS sample under roughly 50% of missing rate while the results from the proposed method are based on 210 incomplete observations and 190 complete observations with binary auxiliary information about PCB. Even though the proposed method is used with 50% of missing in PCB, it provides more precise estimates than ones from MSELE with smaller 95% confidence intervals. The point estimates for covariates are similar. It turns out that PCB does not have statistically significant effect on children's IQ score, based on the 95% confidence interval. Both of the two analyses show that children's IQ has a positive linear relationship to PCB and a negative linear relationship to race and sex. The variance estimates for the proposed method are smaller than that of MSELE, as expected, because the increase in sample size and auxiliary information was used in the proposed method.

Table 3.5: Analysis results for the CPP data set with $(N_0 = 300, N_1 = 50, N_3 = 50)$

	MSELE			Proposed Method		
	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	95% CI	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})$	95% CI
Int	78.13	5.03	(68.26, 88.00)	80.18	4.23	(71.87, 88.48)
PCB	0.15	0.39	(-0.62, 0.93)	0.27	0.33	(-0.39, 0.93)
EDU	1.67	0.51	(0.65, 2.68)	1.49	0.43	(0.63, 2.34)
SES	0.55	0.56	(-0.54, 1.66)	0.56	0.52	(-0.46, 1.59)
RACE	-7.18	1.92	(-10.95, -3.41)	-7.84	1.79	(-11.35, -4.33)
SEX	0.64	1.69	(-2.66, 3.96)	0.16	1.62	(-3.02, 3.34)

a. Results with MSELE are obtained by using complete observations in ODS under 50% missingness

b. Results with the proposed method are obtained under 50% missingness and a binary auxiliary variable for PCB is used

c. An auxiliary variable of PCB is defined as W = I(PCB > median of PCB) and generated for all subjects in the study population

3.7 Discussion

We have proposed an estimated likelihood method to a missing-in covariate data with a binary auxiliary variable under outcome dependent sampling. An estimated likelihood function based on the plug-in method is developed to estimate non-parametric components in the likelihood function. The proposed estimator is shown to be consistent and asymptotically normality is also shown through simulation studies. As far as we know, this is the first trial to handle a continuous missing covariate by using a binary auxiliary variable. The proposed method can include subjects that have missing-in covariate into the estimation procedure. Compared to the maximum semiparametric empirical likelihood estimator in Zhou et al. (2002), that helps the proposed method obtain efficiency gain since one can use incomplete observations with auxiliary information.

For practical usage, we would like to give some guidance to draw an ODS sample having missingness in covariate. In Zhou et al. (2002) and Zhou et al. (2014), they could gain efficiency as they draw supplemental samples from more extreme tails. In our study, since we have a binary auxiliary variable and use an estimated likelihood approach, if we draw a supplementary sample from too extreme tails, it can cause insufficient subjects to derive an estimated likelihood. To be more specific, Zhou et al. (2014) could obtain bigger efficiency gains with cut points of

$(\mu_Y - 1.5\sigma_Y, \mu_Y + 1.5\sigma_Y)$ than with cut points of $(\mu_Y - \sigma_Y, \mu_Y + \sigma_Y)$. However, according to our simulation studies, if we choose extreme cut points, then it is difficult to obtain covariance estimator since we could not obtain a sufficient complete SRS sample that has binary cases from too extreme tails. Hence, we recommend to use cut-points up to $(\mu_Y - \sigma_Y, \mu_Y + \sigma_Y)$.

There are two possible interesting future works. Since we have developed an plug-in estimator about a binary auxiliary variable, the continuous auxiliary variable might be an extension of this study. For example, when auxiliary PCB is measured as a continuous case, we can still make a binary auxiliary variable, but it would cause loss of efficiency. To handle a continuous auxiliary variable, Kernel density estimators in Wasserman (2006) might be good options. Secondly, developing a inference method that uses multiple auxiliary variables corresponding to multiple covariates would be a challenging study, since the plug-in type method would require heavy computations with multiple and mixed types of auxiliary variables.

3.8 Proof of Theorems

3.8.1 Regularity conditions

Condition 9. *β is in the interior of a compact parameter space, B, containing β^* as its interior point.*

Condition 10. *$f(y|x:\beta)$ has the second-order continuous derivatives with respect to β.*

Condition 11. $E\left[-\frac{\partial^2 f(y|x;\beta)}{\partial\beta\beta^T}\right]$ *is finite and positive definite at β^*.*

3.8.2 Proof of the Theorem 4 (Consistency)

Proof. Under the compactness assumption in Condition 9, in a neighborhood of the true parameter β^*, one can show that,

$$\frac{1}{N}\left[\frac{\partial\hat{U}(\beta)}{\partial\beta} - \frac{\partial U(\beta)}{\partial\beta}\right] \to_p 0, \tag{3.12}$$

uniformly for $\beta \in B$ where B denotes the parameter space by using the results (Lemma 3.4 and 3.5) in Weaver (2001).

Since $\frac{-1}{N}\frac{\partial U(\beta)}{\partial\beta\partial} \to_p I(\beta)$ and from (3.12),

$$\frac{-1}{N}\frac{\partial\hat{U}(\beta)}{\partial\beta\partial} \to_p I(\beta) = E\left[\frac{-\partial^2 ln(\beta)}{\partial\beta^T\partial\beta}\right].$$

Moreover, to conclude $\hat{\beta}_p \in B$, we can use the theorems in Foutz (1977, p148) and lemma 3.3 in Weaver (2001). Let

$$f_N(\beta) = \frac{1}{N}\frac{\partial\hat{l}(\beta)}{\partial\beta}, \quad f'_N(\beta) = \frac{1}{N}\frac{\partial^2\hat{l}(\beta)}{\partial\beta\partial\beta^T}.$$

Then, by applying the lemma 3.3 in Weaver (2001), we can conclude that $\hat{\beta}_P = f_N^{-1}(0)$ exists in the set B. Hence, consistency of $\hat{\beta}_p$ is proved. □

3.8.3 Proof of the Theorem 5 (Asymptotic Normality)

Proof. We can write $\frac{1}{\sqrt{N}}\hat{U}(\beta)$ as

$$\begin{aligned}
\frac{1}{\sqrt{N}}\hat{U}(\beta) &= \frac{1}{\sqrt{N}}\sum_{i\in V}\frac{\partial f_\beta(y_i|x_i)/\partial\beta}{f_\beta(y_i|x_i)} - \frac{1}{\sqrt{N}}\sum_{k=1,3} N_k \cdot \frac{\partial\hat{\pi}_k/\partial\beta}{\hat{\pi}_k} \\
&+ \frac{1}{\sqrt{N}}\sum_{j\in\overline{V}_{w=0}}\frac{\partial g_\beta(y_j|w_j=0)/\partial\beta}{g_\beta(y_j|w_j=0)} + \frac{1}{\sqrt{N}}\sum_{j\in\overline{V}_{w=1}}\frac{\partial g_\beta(y_j|w_j=1)/\partial\beta}{g_\beta(y_j|w_j=1)} \\
&+ \frac{1}{\sqrt{N}}\sum_{j\in\overline{V}_{w=0}}\left\{\frac{\partial\hat{g}_\beta(y_j|w_j=0)/\partial\beta}{\hat{g}_\beta(y_j|w_j=0)} - \frac{\partial g_\beta(y_j|w_j=0)/\partial\beta}{g_\beta(y_j|w_j=0)}\right\} \\
&+ \frac{1}{\sqrt{N}}\sum_{j\in\overline{V}_{w=1}}\left\{\frac{\partial\hat{g}_\beta(y_j|w_j=1)/\partial\beta}{\hat{g}_\beta(y_j|w_j=1)} - \frac{\partial g_\beta(y_j|w_j=1)/\partial\beta}{g_\beta(y_j|w_j=1)}\right\} \qquad (3.13)
\end{aligned}$$

Following the proof in Weaver and Zhou (2005), for the set that consists of incomplete observations with $W = 0$, the terms in the third line in (3.13) can be re-expressed as

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{k=1}^{3}\sum_{j\in\overline{V}_{T,k,w=0}}\left\{\frac{\partial\hat{g}_\beta(y_j|w_j=0)/\partial\beta}{g_\beta(y_j|w_j=0)} - \frac{\partial g_\beta(y_j|w_j=0)/\partial\beta}{[g_\beta(y_j|w_j=0)]^2}\times\hat{g}_\beta(y_j|w_j=0)\right\} + O_p(\frac{1}{\sqrt{N}}) \\
= &\frac{1}{\sqrt{N}}\sum_{l=1}^{3}\frac{N_{0,l,w=0}}{N_{0,w=0}\cdot n_{0,k,w=0}}\sum_{i\in V_{0,k,w=0}}\sum_{k=1}^{3}\overline{n}_{T,k,0}\sum_{j\in\overline{V}_{T,k,w=0}}\frac{M_{X_i}(Y_j:\beta)}{\overline{n}_{T,k,w=0}} + O_p(\frac{1}{\sqrt{N}}) \\
= &\frac{1}{\sqrt{N}}\sum_{l=1}^{3}\frac{N_{0,l,w=0}/N_{0,w=0}}{n_{0,k,w=0}/N}\sum_{i\in V_{0,k,w=0}}\sum_{k=1}^{3}\frac{\overline{n}_{T,k,0}}{N}\sum_{j\in\overline{V}_{T,k,w=0}}\frac{M_{X_i}(Y_j:\beta)}{\overline{n}_{T,k,w=0}} + O_p(\frac{1}{\sqrt{N}}) \\
= &\frac{1}{\sqrt{N}}\sum_{l=1}^{3}\frac{N_{0,l,w=0}/N_{0,w=0}}{n_{0,k,w=0}/N}\sum_{i\in V_{0,k,w=0}}\sum_{k=1}^{3}\frac{\overline{n}_{T,k,w=0}}{N}\overline{M}_{X_{i,k,w=0}} + O_p(\frac{1}{\sqrt{N}}),
\end{aligned}$$

where

$$M_{X_i}(Y_j;\beta) = \frac{\partial f_\beta(y_j|X_i)/\partial\beta}{g_\beta(y_j|w_j=0)} - \frac{f(Y_j|X_i;\beta)\partial\hat{g}_\beta(y_j|w_j=0)/\partial\beta}{[g_\beta(y_j|w_j=0)]^2}$$

and

$$\overline{M}_{X_{i,k,w=0}} = \sum_{j \in \overline{V}_{T,k,w=0}} \frac{M_{X_i}(Y_j : \beta)}{\overline{n}_{T,k,w=0}}.$$

By the law of large number,

$$\overline{M}_{X_{i,k,w=0}} \rightarrow_p E_{Y \epsilon C_k, w=0}[M_{X_i}(Y;\beta)].$$

Following (3.27) in Weaver (2001), we can show

$$E_X\Big[E_{Y \epsilon C_k, w=0}(M_X(Y;\beta)) = 0\Big].$$

With the same logic, for the case of $W = 1$, we can conclude that

$$\frac{1}{\sqrt{N}} \sum_{k=1}^{3} \sum_{j \in \overline{V}_{T,k,w=1}} \left\{ \frac{\partial \hat{g}_\beta(y_j|w_j = 1)/\partial\beta}{\hat{g}_\beta(y_j|w_j = 1)} - \frac{\partial g_\beta(y_j|w_j = 1)/\partial\beta}{g_\beta(y_j|w_j = 1)} \right\}$$

$$= \quad \frac{1}{\sqrt{N}} \sum_{l=1}^{K} \frac{N_{0,l,w=1}/N_{0,w=1}}{n_{0,k,w=1}/N} \sum_{i \in V_{0,k,w=1}} \sum_{k=1}^{K} \frac{\overline{n}_{T,k,w=1}}{N} \overline{M}_{X_{i,k,w=1}} + O_p(\frac{1}{\sqrt{N}}),$$

where

$$\overline{M}_{X_{i,k,w=1}} = \sum_{j \in \overline{V}_{T,k,w=1}} \frac{M_{X_i}(Y_j : \beta)}{\overline{n}_{T,k,w=1}}.$$

Define

$$\Lambda_{X_i, w=0}(\beta) = \sum_{k=1}^{3} \frac{\overline{n}_{T,k,w=0}}{N} \cdot \overline{M}_{X_{i,k,w=0}}.$$

and

$$T_{N,w=0} = \frac{1}{N}\sum_{l=1}^{3}\sum_{i\in V_{0,l,w=0}} \frac{N_{0,l,w=0}/N_{0,w=0}}{n_{0,l,w=0}/N}\cdot \Lambda_{X_i,w=0}(\beta).$$

Then, we can show that

$$E(T_{N,w=0}) = \frac{n_{0,l,w=0}}{N}\sum_{l=1}^{3}\frac{N_{0,l,w=0}/N_{0,w=0}}{n_{0,l,w=0}/N}\cdot E(\Lambda_{X,w=0}) = 0$$

and

$$\begin{aligned}
Var(T_{N,w=0}) &= \frac{1}{N^2}\sum_{l=1}^{3} n_{0,l,w=0}\left(\frac{N_{0,l,w=0}/N_{0,w=0}}{n_{0,l,w=0}/N}\right)^2 \text{Var}_{X|Y\in C_l,w=0}(\Lambda_{X,w=0}(\beta)) \\
&= \frac{1}{N}\sum_{l=1}^{3}\frac{n_{0,l,w=0}}{N}\left(\frac{N_{0,l,w=0}/N_{0,w=0}}{n_{0,l,w=0}/N}\right)^2 \text{Var}_{X|Y\in C_l,w=0}(\Lambda_{X,w=0}(\beta)) \\
&= \frac{1}{N}\sum_{l=1}^{3}\frac{(N_{0,l,w=0}/N_{0,w=0})^2}{n_{0,l,w=0}/N}\text{Var}_{X|Y\in C_l,w=0}(\Lambda_{X,w=0}(\beta))
\end{aligned}$$

Along with the same logic, one can derive

$$E(T_{N,w=1}) = \frac{n_{0,l,w=1}}{N}\sum_{l=1}^{3}\frac{N_{0,l,w=1}/N_{0,w=1}}{n_{0,l,w=1}/N}\cdot E(\Lambda_{X,w=1}) = 0$$

and

$$\begin{aligned}
Var(T_{N,w=1}) &= \frac{1}{N^2}\sum_{l=1}^{3} n_{0,l,w=1}\left(\frac{N_{0,l,w=1}/N_{0,w=1}}{n_{0,l,w=1}/N}\right)^2 \text{Var}_{X|Y\in C_l,w=1}(\Lambda_{X,w=1}(\beta)) \\
&= \frac{1}{N}\sum_{l=1}^{3}\frac{n_{0,l,w=1}}{N}\left(\frac{N_{0,l,w=1}/N_{0,w=1}}{n_{0,l,w=1}/N}\right)^2 \text{Var}_{X|Y\in C_l,w=1}(\Lambda_{X,w=1}(\beta)) \\
&= \frac{1}{N}\sum_{l=1}^{3}\frac{(N_{0,l,w=1}/N_{0,w=1})^2}{n_{0,l,w=1}/N}\text{Var}_{X|Y\in C_l,w=1}(\Lambda_{X,w=1}(\beta))
\end{aligned}$$

Recall what we need to drive is the asymptotic distribution of $\frac{1}{\sqrt{N}}\widehat{U}(\beta)$. Note that

$$\begin{aligned}
\frac{1}{\sqrt{N}}\widehat{U}(\beta) = \quad & \frac{1}{\sqrt{N}}\Bigg[\sum_{i\in V}\frac{\partial f_\beta(y_i|x_i)/\partial\beta}{f_\beta(y_i|x_i)} - \sum_{k=1,3} N_k\cdot\frac{\partial\hat{\pi}_k/\partial\beta}{\hat{\pi}_k} + \\
& \sum_{j\in\overline{V}_{w=0}}\frac{\partial g_\beta(y_j|w_j=0)/\partial\beta}{g_\beta(y_j|w_j=0)} + \sum_{j\in\overline{V}_{w=1}}\frac{\partial g_\beta(y_j|w_j=1)/\partial\beta}{g_\beta(y_j|w_j=1)}\Bigg] + \\
& \left[\frac{1}{\sqrt{N}}\sum_{j\in\overline{V}_{w=0}}\left\{\frac{\partial\hat{g}_\beta(y_j|w_j=0)/\partial\beta}{\hat{g}_\beta(y_j|w_j=0)} - \frac{\partial g_\beta(y_j|w_j=0)/\partial\beta}{g_\beta(y_j|w_j=0)}\right\}\right] + \\
& \left[\frac{1}{\sqrt{N}}\sum_{j\in\overline{V}_{w=1}}\left\{\frac{\partial\hat{g}_\beta(y_j|w_j=1)/\partial\beta}{\hat{g}_\beta(y_j|w_j=1)} - \frac{\partial g_\beta(y_j|w_j=1)/\partial\beta}{g_\beta(y_j|w_j=1)}\right\}\right] \\
\equiv \quad & T_1 + T_2 + T_3 \qquad (3.14)
\end{aligned}$$

The terms in the first bracket denotes the true-full score function and, the second and third are differences between the estimated and true score functions. Note that those three parts are mutually independent. Thus, we can derive the asymptotic distribution for each term in (3.14) and by summing them together, the asymptotic distribution of $\frac{1}{\sqrt{N}}\widehat{U}(\beta)$ as follows :

Since we assume that $\frac{N_{0,k,w=0}}{N_{0,w=0}} \to_p P(Y\in C_k|W=0) = a_{0k}$, $\frac{N_{0,k,w=1}}{N_{0,w=1}} \to_p P(Y\in C_k|W=1) = a_{1k}$, $\frac{n_{0,k,w=0}}{N} \to_p b_{0k}$, $\frac{n_{0,k,w=1}}{N} \to_p b_{1k}$, $\frac{\overline{n}_{T,k,w=0}}{N_{k,w=0}} \to_p c_{0k}$, and $\frac{\overline{n}_{T,k,w=1}}{N_{k,w=1}} \to_p c_{1k}$ for all k's and we already showed that $\frac{-1}{N}\frac{\partial\hat{U}(\beta)}{\partial\beta\partial} \to_p I(\beta)$,

$$\begin{aligned}
T_1 &\sim N\Big[E(T_1), I(\beta)\Big];\\
T_2 &\sim N\Big[E(T_2), \sum_{l=1}^{K}\{\frac{(a_{0l})^2}{b_{0l}}\Sigma_{l0}\}\Big];\\
T_3 &\sim N\Big[E(T_3), \sum_{l=1}^{K}\{\frac{(a_{1l})^2}{b_{1l}}\Sigma_{l1}\}\Big]
\end{aligned}$$

By summing asymptotic distributions of $T_1, T_2,$ and T_3 together, we can derive the asymptotic

distribution of

$$\frac{-1}{\sqrt{N}}\widehat{U}(\beta) \rightarrow_D N(0, V(\beta)),$$

where

$$V(\beta) = I_\beta + \sum_{l=1}^{K}\left\{\frac{(a_{0l})^2}{b_{0l}}\Sigma_{l0}\right\} + \sum_{l=1}^{K}\left\{\frac{(a_{1l})^2}{b_{1l}}\Sigma_{l1}\right\}.$$

Finally, by using Slutsky's theorem in Lehmann (1999), we can conclude that $\sqrt{N}\left(\hat{\beta} - \beta_*\right) \xrightarrow{D} N_p\left(0, \Sigma\left(\beta_*\right)\right)$. □

CHAPTER 4: AN AUXILIARY COVARIATE STRATIFIED SAMPLING DESIGN

4.1 Introduction

In observational studies, the study sizes are often limited by the cost of assessments to the exposure. Biased sampling design is a well-known approach to improve study efficiency and save costs. Case-control study and Outcome Dependent Sampling design are especially well-known approaches to improve study efficiency by drawing an oversample of subjects that are believed to have more information about the relationship between response and covariates of interest. However, in reality, there would be missing in covariate(s) when it comes to prospective observational studies. In the CPP study described in Niswander and Gordon (1972), to assay PCB from blood samples was too expensive, and researchers were not allowed to have PCB exposure from all participants in the study. Thus, only a limited number of subjects in the cohort population were used to assay blood samples. In similar situations, incorporating an auxiliary variable for a covariate of interest can be considered to extract related information. In this chapter, we describe how to use an auxiliary variable when one dose not have the covariate of interest in prospective cohort studies.

To obtain more efficient estimates under the fixed sample size, Zhou et al. (2002) proposed an Outcome-Dependent Sampling (hereafter, ODS) design and compared its efficiency with the inverse probability estimator and maximum likelihood estimators with a simple random sample scheme. They showed that estimates under ODS design are more efficient and derived asymptotic properties reflecting the ODS design. The main idea of ODS design is to enhance the efficiency of the estimator by drawing a supplemental sample from two tails of response distribution. In addition, Zhou et al. (2014) developed a new two-stage sampling design, Probability-Dependent

Sampling (PDS), to draw a sample by using covariate information from a simple random sample in the first stage. In this chapter, we consider a measurement error problem on a covariate of interest and propose a new sampling scheme based on an auxiliary covariate. We will show how to use an auxiliary covariate to draw supplemental sample when the covariate of interest is not obtainable in the first stage. In previous research, the estimated likelihood method under ODS design was developed by Weaver and Zhou (2005). Zhou et al. (2002) and Zhou et al. (2014) used a semiparametric empirical likelihood approach along with Owen (1988) and Qin and Lawless (1994). We use an estimated likelihood approach to estimate selection probabilities in the likelihood function.

In measurement error problems, auxiliary variables for some expensive (or missing) variables can be included in inferences to gain more efficient estimators. In this chapter we assume that a continuous exposure is too expensive to obtain in a study population but an auxiliary variable for the expensive exposure is easy to obtain. For example, researchers are interested in the relationship between BMI and food intake. However, to measure the exact amount of weekly food intake for each participant is quite difficult. As an alternative, self-assessed food intake, breakfast menu, or most frequent dishes might be approximate measurements for weekly food intake. Under the assumption that response variable and covariate of interest are observed in a study population, an auxiliary variable for the covariate of interest is measured for all subjects in the study. With the auxiliary covariate for the exposure, X^*, we draw a simple random sample having measurements of response, Y, and exposure of interest, X. With this SRS, we estimate X under a working model between X and X^* for all subjects in the study. Then, $\hat{X}$ is assumed to be partitioned into three mutually exclusive intervals : $(-\infty, c_1] \cup (c_1, c_3] \cup (c_3, \infty)$ where c_1, c_3 are some =fixed constants. With these cutoff points, we draw two subsamples from $(-\infty, c_1]$ and (c_3, ∞), respectively to implement a new sampling scheme.

Thus, we propose a new sampling design, an Auxiliary Covariate Stratified Sampling(ACSS) scheme and use an estimated likelihood approach for inference. Compared to the outcome dependent sampling design in Zhou et al. (2002) and Weaver and Zhou (2005), the ACSS is

as follows : (a) ACSS assumes that one does not have information about response variable and covariate of interest in a study population ; (b) auxiliary information for covariate of interest is available for all subjects in study ; (c) a biased sampling scheme is applied to a cohort population based on a working model between an auxiliary covariate and covariate of interest ;

In this chapter, we proposed a new two-stage sampling scheme with auxiliary variable for covariate of interest. To be specific, we consider a linear regression model in a prospective observational study and assume that a continuous auxiliary variable, W, for covariate of interest, can be observed in all subjects in a study but we can draw a sample of complete observations only in the ACSS sample. In Section 4.2, we will describe the data structure under the ACSS, and Section 4.3 will show the likelihood and estimation methods. Sections 4.4 and 4.5 will show asymptotic properties of the proposed method and simulation study results, respectively. A real data application by comparing the proposed method with other existing methods will be given in Section 4.6 and a brief discussion will follow in section 4.7.

4.2 Data structure and Inference for two-stage ACSS

4.2.1 Data structure for ACSS

To fix notation, let Y be a continuous response, X be a covariate of interest, and W be a continuous auxiliary variable for X. We assume that the relationship between Y and X, and, the relationship between X and W will follow parametric model $f(Y \mid X;\beta)$ and $k(X \mid W;\gamma)$, where β and γ are the regression parameter for $f(\cdot)$ and $k(\cdot)$, respectively. The proposed two-phase ACSS scheme with a continuous auxiliary covariate is as follows: Given a population dataset of size N in the first stage, we can only observe (Y, W). With a simple random sample from the study population, we estimate γ under $k(X \mid W;\gamma)$. Then, $\{\hat{X}_i\}$ for all subjects in the study is obtained and we choose cut-points $(c_1, c_3) = (\mu_{\hat{X}} - a * \sigma_{\hat{X}}, \mu_{\hat{X}} + a * \sigma_{\hat{X}})$ where a is a fixed constant and $\hat{X}$ is the prediction of X for all subjects in the population. Next, we draw a supplemental ACSS sample of size n_1 from the sample having $\hat{X}$ below c_1 and another supplemental sample of size n_3 from the sample having $\hat{X}$ above c_3. Thus, we can summarize

the data structure as follows :

$$
\begin{aligned}
&\text{In the 1st stage,} \\
&\text{SRS :} \quad \{Y_i, X_i, W_i, \hat{X}_i\} \\
&\text{In the 2nd stage,} \\
&\text{ACSS Left :} \quad \{Y_i, X_i, W_i, \hat{X}_i \mid \hat{X}_i \le c_1\}, \\
&\text{ACSS Right :} \quad \{Y_i, X_i, W_i, \hat{X}_i \mid \hat{X}_i \ge c_3\}.
\end{aligned} \tag{4.1}
$$

In the next section, we derive a likelihood function that reflects (4.1).

4.2.2 Construction of a likelihood function

Let $g(x)$ and $G(x)$ denote the probability distribution function and the cumulative distribution function of X. Then, based on the data structure in (4.1), we can construct a likelihood function as follows:

$$
L(\beta) = \prod_{i=1}^{n_0} f_\beta(y_{0i}, x_{0i}) \cdot \prod_{j=1}^{n_1} f_\beta(y_{1j}, x_{1j} \mid \hat{X}_{1j} \le c_1) \cdot \prod_{j=1}^{n_3} f_\beta(y_{3j}, x_{3j} \mid \hat{X}_{3j} \ge c_3)
$$

Using the Bayes formula, $L(\beta)$ can be expressed as

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{n_0} f_\beta(y_{0i} \mid x_{0i}) g(x_{0i}) \cdot \prod_{j=1}^{n_1} \frac{f_\beta(y_{1j} \mid x_{1j}) g(x_{1j})}{P(\hat{X}_{1j} \le c_1)} \cdot \prod_{j=1}^{n_3} \frac{f_\beta(y_{3j} \mid x_{3j}) g(x_{3j})}{P(\hat{X}_{3j} \ge c_3)} \\
&= \prod_{i=1}^{n} f_\beta(y_i \mid x_i) g(x_i) \cdot \left(\frac{1}{P(\hat{X} \le c_1)} \right)^{n_1} \cdot \left(\frac{1}{P(\hat{X} \ge c_3)} \right)^{n_3} \\
&\propto \prod_{i=1}^{n} f_\beta(y_i \mid x_i)
\end{aligned} \tag{4.2}
$$

In the next section, we show that $P(\hat{X} \le c_1)$ and $P(\hat{X} \ge c_3)$ are not related to β that we can obtain the likelihood function in the last line in (4.2).

4.3 Maximum likelihood approach

4.3.1 Derivation of the likelihood function

We propose a maximum likelihood approach for the likelihood in (4.2). First, we show $P(\hat{X} \leq c_1)$ in (4.2) is not related to β, the main parameter. By using Bayes formula, we can write

$$P(\hat{X} \leq c_1) = \int_W P(\hat{X} \leq c_1 \mid W, \gamma) dP(w)$$

Since we do not specify any assumption about the distribution of W, we can estimate $P(W \leq w)$ by using SRS sample in the first stage as $\hat{P}(W \leq w) = \sum_{i=1}^{n_0} I(W_i \leq w)/n_0$. By using the plug-in estimator $\hat{P}(W \leq w)$ and $k(X \mid W, \gamma)$ that assumes a parametric model for the relationship between X and W, we can estimate $P(\hat{X} \leq c_1)$ as

$$\hat{P}(\hat{X} \leq c_1) = \frac{\sum_{i=1}^{n_0} \left\{ P(\hat{X} \leq c_1 \mid W, \hat{\gamma}) \right\}}{n_0}$$

where $\hat{\gamma}$ is obtained based on SRS sample. $\hat{P}(\hat{X} \geq c_3)$ is obtained in the same way. Thus, $P(\hat{X} \leq c_1)$ and $\hat{P}(\hat{X} \geq c_3)$ are not related to the estimation of β. Hence, $L(\beta)$ can be expressed as follows :

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} f_\beta(y_i \mid x_i) g(x_i) \cdot \left(\frac{1}{\hat{P}(\hat{X} \leq c_1)} \right)^{n_1} \cdot \left(\frac{1}{\hat{P}(\hat{X} \geq c_3)} \right)^{n_3} \\ &\propto \prod_{i=1}^{n} f_\beta(y_i \mid x_i) \end{aligned}$$

4.3.2 Inferences

By taking the log-transformation on $L(\beta))$, we have the log-likelihood function as

$$l(\beta) = \log L(\beta).$$

By taking derivatives with respect to β on $l(\beta)$, we can define a score equation

$$0 = S(\beta) = \frac{\partial l(\beta)}{\partial \beta}, \tag{4.3}$$

and $\hat{\beta}_p$ denotes the solution of (4.3).

Lastly, we use NLopt (Nonlinear optimization) algorithm to find $\hat{\beta}_p$ that maximizes $l(\beta)$. In the next section, we show asymptotic properties of the proposed estimator.

4.4 Asymptotic properties of the proposed estimator

In this section, we present two theorems about asymptotic properties of $\hat{\beta}_p$. A consistent estimator of covariance of $\hat{\beta}_p$ could be obtained by using the maximum likelihood theory in Casella and Berger (2002). Theorem 7 provides the consistency and asymptotic normality results for the proposed estimator, whereas Theorem 8 gives a consistent estimator for the asymptotic variance matrix.

Theorem 7. *(Consistency and asymptotic normality of $\hat{\beta}_p$)*
Under general regularity conditions and as $n \to \infty$, $\hat{\beta} \to \beta_*$ in probability, where β_* is the true value and

$$\sqrt{n}(\hat{\beta}_p - \beta_*) \to_d N(0, I^{-1}(\beta_*)) \tag{4.4}$$

where

$$I(\beta) = -E\left\{\frac{\partial^2 l(\beta)}{\partial \beta^T \beta}\right\}$$

Theorem 8. *(A consistent estimator of Σ)*
Under general regularity conditions and as $n \to \infty$, a consistent estimator of $I(\beta_*)$ in (4.4) can

be obtained as

$$\widehat{I}(\beta_*) = \frac{-1}{n_v}\sum_{i=1}^{n_v}\log f_\beta(y_i \mid x_i).$$

In the next section, we show how well the asymptotic properties work out through simulation studies under different settings.

4.5 Simulation studies

4.5.1 Simulations under the correct working model

In this section, to see the properties of small sample behavior, we compare the performance of the proposed methods with those of other existing methods under different settings. The simulations studies are conducted with the statistical software, R version 3.2.2. To generate the underlying population for $f(Y \mid X : \beta)$, we consider a linear regression model as follows :

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon \tag{4.5}$$

where $X \sim N(0,1)$ in Table 4.1 and 4.2, $X \sim LN(1, 0.6^2)$ in Table 4.3 and 4.4, $Z \sim Bernoulli(0.45)$, $\varepsilon \sim N(0,1)$, and X, Z, and ε are mutually independent. We generate a continuous auxiliary variable W for X as $W = X + e$ where $e \sim N(0, \sigma_w^2)$ in Table 4.1 and 4.2, and $e \sim LN(1, \sigma_w^2)$ in Table 4.3 and 4.4. We consider a working model between X and W as

$$X = \gamma_0 + \gamma_1 W + e, \tag{4.6}$$

where e is a random error term. To choose cut-off points, with a simple random sample with size of n_0, we estimate γ and obtain $\hat{X}$ for all subjects in the study except n_0 subjects in the simple random sample. We choose $(c_1, c_3) = (\mu_{\hat{X}} - a \cdot \sigma_{\hat{X}}, \mu_{\hat{X}} + a \cdot \sigma_{\hat{X}})$ for the symmetric X and $(c_1, c_3) = (\text{1st quartile of } \hat{X}, \text{3rd quartile of } \hat{X})$ for the asymmetric X from the population data. From subjects in the population except the simple random sample, two supplemental

ACDS samples are drawn from the population below c_1 and above c_3 with sizes of n_1 and n_3, respectively.

Four competing estimators are compared and all methods are under the same sample size, $n(= n_0 + n_1 + n_3)$. We describe following competitors : (i) $\hat{\beta}_X$, the estimator ignoring the sampling structure and treats the data as if it were an independent sample; (ii) $\hat{\beta}_{SRS}$, the ordinary linear regression estimator from an SRS with the same sample size as the total sample size in ACDS; (iii) $\hat{\beta}_{IPW}$, the inverse-probability-weighted method in Horvitz and Thompson (1952) from a two-stage ACDS sample assuming one can calculate a sampling fraction; (iv) $\hat{\beta}_P$ is the proposed estimator for a continuous auxiliary variable under ACDS sample design. Simulation studies are conducted by varying β_1 and σ_w. Since σ_w shows the relationship between covariate of interest and auxiliary covariate, σ_w = 0 and 1 are used in particular to see changes in efficiency gain with varying the strength of relationship between X and W.

In Table 4.1, we conducted Monte Carlo studies with $\sigma_w = 0$, assuming there is no measurement error from the main covariate. Simulation studies in Table 4.2 are conducted with σ_w = 1 to see small sample properties under weak relationships between the main covariate and auxiliary variable. We conduct additional simulation studies to compare estimators with a non symmetric continuous covariate following log-normal distributions in Table 4.3 and 4.4. X is generated $LN(1, 0.6^2)$ and W follows $X + LN(1, 1)$ and $X + LN(1, e^2)$ in Table 4.3 and 4.4, respectively. We could make the following observations from simulation studies : (1) All estimators are unbiased except $\hat{\beta}_X$ that does not reflect biased sampling scheme; (2) $\hat{\beta}_P$ is the most efficient under all different settings; (3) With the total fixed sample size, the more subjects from tails we have, the more efficient $\hat{\beta}_P$ we obtain. However, note that SD of $\hat{\beta}_{IPW}$ increases when we draw more subjects from tails under the fixed total sample size; (4) The finding that $\hat{\beta}_P$ is more efficient than $\hat{\beta}_{IPW}$ shows that the estimated likelihood approach can bring efficiency gain under the same sampling design; (5) Based on the 95% coverage probability, the nominal 95% confidence interval coverage rates are close to 95%, indicating that the large sample normal approximation works well in the simulation studies for all estimators; (6) To see

the performance according to strength of the relationship between X and an W, we conduct simulation studies by varying σ_w. With σ_w = 1, SE's of $\hat{\beta}_{IPW}$ and $\hat{\beta}_P$ become larger than SE's with σ_w = 0 since $X's$ with σ_w = 1 have a stronger relationship between X and W. In addition, we conduct additional simulation results to compare the efficiencies of two ACSS with two different cut-off points : $(\mu_{\hat{X}} - \sigma_{\hat{X}}, \mu_{\hat{X}} + \sigma_{\hat{X}})$ and $(\mu_{\hat{X}} - 1.5\sigma_{\hat{X}}, \mu_{\hat{X}} + 1.5\sigma_{\hat{X}})$ in Table 4.5 ; (1st quartile, 3rd quartile) of $\hat{X}$ and (lower 10%, upper 10%) of $\hat{X}$ in Table 4.6. We could find that we have a more efficient estimator when we draw supplemental samples from more extreme tails of the distribution of $\hat{X}$ than estimators from less extreme tails, in general. For example, in Table 4.5, under the true value of β_1 = 0.5, the relative efficiency is greater than 1, which means , the proposed estimator with ACSS from more extreme part of $\hat{X}$ is the more efficient than the one from the more extreme part of $\hat{X}$.

Table 4.1: Simulation results for asymptotic properties in Section 4.4. Results are based on 1,000 simulations with N = $2,000$, $X \sim N(0,1)$, $W = X$, and various (n_0, n_1, n_3) with n_v = 400

β_1	Method	Mean	SE	$\widehat{\text{SE}}$	95% CI
(n_0, n_1, n_3) = $(250, 25, 25)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.056	0.057	0.959
	$\hat{\beta}_{IPW}$	0.000	0.051	0.052	0.954
	$\hat{\beta}_p$	0.000	0.051	0.051	0.948
0.5	$\hat{\beta}_{SRS}$	0.498	0.057	0.058	0.952
	$\hat{\beta}_{IPW}$	0.500	0.053	0.052	0.944
	$\hat{\beta}_p$	0.501	0.050	0.051	0.952
(n_0, n_1, n_3) = $(150, 75, 75)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.056	0.057	0.959
	$\hat{\beta}_{IPW}$	0.001	0.051	0.050	0.951
	$\hat{\beta}_p$	0.001	0.043	0.043	0.942
0.5	$\hat{\beta}_{SRS}$	0.498	0.057	0.058	0.952
	$\hat{\beta}_{IPW}$	0.500	0.049	0.050	0.952
	$\hat{\beta}_p$	0.499	0.042	0.043	0.956
(n_0, n_1, n_3) = $(100, 100, 100)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.056	0.057	0.959
	$\hat{\beta}_{IPW}$	0.000	0.053	0.054	0.959
	$\hat{\beta}_p$	-0.000	0.041	0.040	0.950
0.5	$\hat{\beta}_{SRS}$	0.498	0.057	0.058	0.952
	$\hat{\beta}_{IPW}$	0.501	0.055	0.054	0.953
	$\hat{\beta}_p$	0.501	0.039	0.040	0.948

a. The results are based on $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, where $\varepsilon \sim N(0,1), X \sim N(0,1)$, $Z \sim Bernoulli(0.45)$

b. $W = X$ and a working model is posited as $X = \gamma_0 + \gamma_1 W + e$, where e is a random error term

c. The cut-off points are $(\mu_{\hat{X}} - \sigma_{\hat{X}}, \mu_{\hat{X}} + \sigma_{\hat{X}})$

Table 4.2: Simulation results for asymptotic properties in Section 4.4. Results are based on 1,000 simulations with $N = 2,000$, $X \sim N(0,1)$, $W = X + N(0,1)$, and various (n_0, n_1, n_3) with $n_v = 400$

β_1	Method	Mean	SE	$\widehat{\text{SE}}$	95% CI
$(n_0, n_1, n_3) = (250, 25, 25)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.056	0.057	0.959
	$\hat{\beta}_{IPW}$	-0.000	0.055	0.055	0.955
	$\hat{\beta}_p$	-0.002	0.053	0.054	0.945
0.5	$\hat{\beta}_{SRS}$	0.498	0.057	0.058	0.952
	$\hat{\beta}_{IPW}$	0.501	0.056	0.055	0.947
	$\hat{\beta}_p$	0.500	0.054	0.054	0.946
$(n_0, n_1, n_3) = (150, 75, 75)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.056	0.057	0.959
	$\hat{\beta}_{IPW}$	-0.001	0.062	0.061	0.945
	$\hat{\beta}_p$	-0.002	0.050	0.049	0.942
0.5	$\hat{\beta}_{SRS}$	0.498	0.057	0.058	0.952
	$\hat{\beta}_{IPW}$	0.501	0.063	0060	0.936
	$\hat{\beta}_p$	0.497	0.049	0.048	0.949
$(n_0, n_1, n_3) = (100, 100, 100)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.056	0.057	0.959
	$\hat{\beta}_{IPW}$	0.005	0.073	0.070	0.923
	$\hat{\beta}_p$	-0.001	0.045	0.046	0.952
0.5	$\hat{\beta}_{SRS}$	0.498	0.057	0.058	0.952
	$\hat{\beta}_{IPW}$	0.501	0.069	0.068	0.938
	$\hat{\beta}_p$	0.499	0.047	0.047	0.955

a. The results are based on $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, where $\varepsilon \sim N(0,1), X \sim N(0,1)$, $Z \sim Bernoulli(0.45)$

b. $W = X + N(0,1)$ and a working model is posited as $X = \gamma_0 + \gamma_1 W + e$, where e is a random error term

c. The cut-off points are $(\mu_{\hat{X}} - \sigma_{\hat{X}}, \mu_{\hat{X}} + \sigma_{\hat{X}})$

Table 4.3: Simulation results for asymptotic properties in Section 4.4. Results are based on 1,000 simulations with $N = 2,000$, $X \sim LN(1, 0.6)$, $W = X + LN(1, 1)$, and various (n_0, n_1, n_3) with $n_v = 400$

β_1	Method	Mean	SE	$\widehat{\text{SE}}$	95% CI
$(n_0, n_1, n_3) = (250, 25, 25)$					
0	$\hat{\beta}_{SRS}$	0.000	0.026	0.027	0.962
	$\hat{\beta}_{IPW}$	0.001	0.024	0.025	0.945
	$\hat{\beta}_p$	0.001	0.024	0.024	0.945
0.5	$\hat{\beta}_{SRS}$	0.499	0.026	0.027	0.955
	$\hat{\beta}_{IPW}$	0.499	0.025	0.025	0.953
	$\hat{\beta}_p$	0.499	0.025	0.025	0.954
$(n_0, n_1, n_3) = (150, 75, 75)$					
0	$\hat{\beta}_{SRS}$	0.000	0.026	0.027	0.962
	$\hat{\beta}_{IPW}$	0.000	0.023	0.023	0.951
	$\hat{\beta}_p$	0.000	0.022	0.022	0.947
0.5	$\hat{\beta}_{SRS}$	0.499	0.026	0.027	0.955
	$\hat{\beta}_{IPW}$	0.498	0.024	0.023	0.945
	$\hat{\beta}_p$	0.501	0.023	0.022	0.943
$(n_0, n_1, n_3) = (100, 100, 100)$					
0	$\hat{\beta}_{SRS}$	0.000	0.026	0.027	0.962
	$\hat{\beta}_{IPW}$	-0.000	0.023	0.023	0.941
	$\hat{\beta}_p$	-0.000	0.022	0.021	0.943
0.5	$\hat{\beta}_{SRS}$	0.499	0.026	0.027	0.955
	$\hat{\beta}_{IPW}$	0.499	0.024	0.023	0.931
	$\hat{\beta}_p$	0.500	0.021	0.021	0.948

a. The results are based on $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, where $\varepsilon \sim N(0, 1)$, $X \sim LN(1, 0.6^2)$, $Z \sim Bernoulli(0.45)$

b. $W = X + LN(1, 1)$ and a working model is posited as $\log X = \gamma_0 + \gamma_1 \log W + e$, where e is a random error term

c. The cut-off points are (25% quantile, 75% quantile) of $\hat{X}$

Table 4.4: Simulation results for asymptotic properties in Section 4.4. Results are based on 1,000 simulations with $N = 2,000$, $X \sim LN(1, 0.6)$, $W = X + LN(1, e^2)$, and various (n_0, n_1, n_3) with $n_v = 400$

β_1	Method	Mean	SE	$\widehat{\text{SE}}$	95% CI
$(n_0, n_1, n_3) = (250, 25, 25)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.026	0.027	0.959
	$\hat{\beta}_{IPW}$	-0.000	0.025	0.026	0.943
	$\hat{\beta}_p$	0.000	0.026	0.025	0.934
0.5	$\hat{\beta}_{SRS}$	0.500	0.028	0.027	0.944
	$\hat{\beta}_{IPW}$	0.500	0.027	0.026	0.923
	$\hat{\beta}_p$	0.499	0.025	0.025	0.937
$(n_0, n_1, n_3) = (150, 75, 75)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.026	0.027	0.959
	$\hat{\beta}_{IPW}$	0.000	0.027	0.026	0.942
	$\hat{\beta}_p$	0.000	0.024	0.023	0.938
0.5	$\hat{\beta}_{SRS}$	0.500	0.028	0.027	0.944
	$\hat{\beta}_{IPW}$	0.501	0.026	0.026	0.946
	$\hat{\beta}_p$	0.500	0.024	0.024	0.941
$(n_0, n_1, n_3) = (100, 100, 100)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.026	0.027	0.959
	$\hat{\beta}_{IPW}$	0.000	0.030	0.028	0.932
	$\hat{\beta}_p$	-0.000	0.023	0.023	0.948
0.5	$\hat{\beta}_{SRS}$	0.500	0.028	0.027	0.944
	$\hat{\beta}_{IPW}$	0.499	0.030	0.028	0.924
	$\hat{\beta}_p$	0.499	0.023	0.022	0.933

a. The results are based on $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, where $\varepsilon \sim N(0, 1), X \sim LN(1, e^2), Z \sim Bernoulli(0.45)$

b. $W = X + LN(1, e^2)$ and a working model is posited as $\log X = \gamma_0 + \gamma_1 \log W + e$, where e is a random error term

c. The cut-off points are (25% quantile, 75% quantile) of $\hat{X}$

Table 4.5: Relative efficiency comparison under symmetric $X \sim N(0,1), W = X + N(0, \sigma_w^2)$

	$\sigma_w = 0$		$\sigma_w = 1$	
(n_0, n_1, n_3)	$\beta_1 = 0$	$\beta_1 = 0.5$	$\beta_1 = 0$	$\beta_1 = 0.5$
(250, 25,25)				
RE	1.06	1.15	1.06	1.06
(150, 75,75)				
RE	1.18	1.14	1.14	1.14
(100,100,100)				
RE	1.21	1.23	1.12	1.12

- $RE = SE(\hat{\beta}_{P_1})/SE(\hat{\beta}_{P_2})$
- Cut-off points for $\hat{\beta}_{P_1}$ is $(\mu_{\hat{X}} - \sigma_{\hat{X}},\ \mu_{\hat{X}} + \sigma_{\hat{X}})$
- Cut-off points for $\hat{\beta}_{P_2}$ is $(\mu_{\hat{X}} - 1.5\sigma_{\hat{X}},\ \mu_{\hat{X}} + 1.5\sigma_{\hat{X}})$

Table 4.6: Relative efficiency comparison under asymmetric $X \sim LN(1, 0.6), W = X + LN(1, \sigma_w^2)$

	$\sigma_w = 0$		$\sigma_w = 1$	
(n_0, n_1, n_3)	$\beta_1 = 0$	$\beta_1 = 0.5$	$\beta_1 = 0$	$\beta_1 = 0.5$
(250, 25,25)				
RE	1.10	1.10	1.03	1.03
(150, 75,75)				
RE	1.28	1.28	1.07	1.07
(100,100,100)				
RE	1.35	1.34	1.10	1.15

- $RE = SE(\hat{\beta}_{P_1})/SE(\hat{\beta}_{P_2})$
- Cut-off points for $\hat{\beta}_{P_1}$ are (1st quartile, 3rd quartile) of $\hat{X}$
- Cut-off points for $\hat{\beta}_{P_2}$ are (lower 10%, upper 10%) of $\hat{X}$

4.5.2 Simulation study under the misspecified working model

In this section, we provide simulation studies to see the results under the misspecified working model between $X \sim N(0,1)$ and $W = X + N(0,1)$. We use the same underlying linear regression model in (4.8) and the true working model is $X = \gamma_0 + W\gamma_1 + e$, where $W = X + N(0,1)$. We consider a misspecified model,

$$X = \gamma_0 + \gamma_1 \sin W + \gamma_2 Z + e, \tag{4.7}$$

where e is a random error term. Based on the misspecified working model, we choose $(c_1, c_3) = (\mu_{\hat{X}} - \sigma_{\hat{X}}, \mu_{\hat{X}} + \sigma_{\hat{X}})$ for X from the population data based on a simple random sample with the size of n_0. From subjects in the population except the simple random sample, two supplemental stratified samples are drawn from the population below c_1 and above c_3 of $\hat{X}$ with sizes of n_1 and n_3, respectively. We compare $\hat{\beta}_{SRS}$ and $\hat{\beta}_{IPW}$ with $\hat{\beta}_p$ under the same sample size and with varying β_1 and allocation of the stratified sample. By comparing $\hat{\beta}_{SRS}$ and $\hat{\beta}_p$, we could see that with the misspecified working models' SE from $\hat{\beta}_{SRS}$ and $\hat{\beta}_p$ are about the same. It tells us that we would not expect efficiency gain with the stratified sampling design if we consider the misspecified working model between X and W. In addition, we could see that we obtain unbiased estimators even with the misspecified working model for the relationship between X and W. Compared to the results under the working model (4.6) in the previous section, there are two findings for results in Table 4.7 : (1) Comparing $\hat{\beta}_p$ and $\hat{\beta}_{SRS}$, since $\hat{X}$s with the working model (4.7) are less precise than $\hat{X}$s with (4.6), we could expect that $\hat{X}$s from (4.7) would not be as good as $\hat{X}$s from the working model (4.6). Thus, SE of $\hat{\beta}_p$ and SE of $\hat{\beta}_{SRS}$ about the same, meaning that we might not take advantage of the ACSS with the working model (4.7); (2) $\hat{\beta}_p$ has smaller SE that SE from $\hat{\beta}_{IPW}$ since $\hat{\beta}_p$ uses the maximum likelihood approach whereas $\hat{\beta}_{IPW}$ is obtained with the estimating equations approach. For all three estimators, we could see that all of them satisfy the asymptotic properties based on $SE, \widehat{SE}$, and 95% CIs.

Table 4.7: Simulation results under the misspecified working model between X and W. Results are based on 1,000 simulations with N = $2,000$, $X \sim N(0,1)$, $W = X + N(0,1)$, and varying (n_0, n_1, n_3) under n_v = 400

β_1	Method	Mean	SE	$\widehat{\text{SE}}$	95% CI
(m_0, n_1, n_3) = $(250, 25, 25)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.056	0.057	0.959
	$\hat{\beta}_{IPW}$	-0.001	0.059	0.057	0.939
	$\hat{\beta}_p$	0.001	0.058	0.057	0.939
0.5	$\hat{\beta}_{SRS}$	0.498	0.057	0.058	0.952
	$\hat{\beta}_{IPW}$	0.499	0.059	0.057	0.934
	$\hat{\beta}_p$	0.499	0.057	0.057	0.949
(m_0, n_1, n_3) = $(150, 75, 75)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.056	0.057	0.959
	$\hat{\beta}_{IPW}$	0.002	0.067	0.064	0.931
	$\hat{\beta}_p$	-0.001	0.058	0.056	0.947
0.5	$\hat{\beta}_{SRS}$	0.498	0.057	0.058	0.952
	$\hat{\beta}_{IPW}$	0.499	0.065	0.064	0.945
	$\hat{\beta}_p$	0.499	0.058	0.056	0.941
(m_0, n_1, n_3) = $(100, 100, 100)$					
0	$\hat{\beta}_{SRS}$	-0.000	0.056	0.057	0.959
	$\hat{\beta}_{IPW}$	0.001	0.076	0.073	0.926
	$\hat{\beta}_p$	0.000	0.057	0.056	0.940
0.5	$\hat{\beta}_{SRS}$	0.498	0.057	0.058	0.952
	$\hat{\beta}_{IPW}$	0.496	0.075	0.074	0.936
	$\hat{\beta}_p$	0.498	0.056	0.056	0.949

a. The results are based on $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, where $\varepsilon \sim N(0,1), X \sim N(0,1), Z \sim Bernoulli(0.45)$

b. $W = X + N(0,1)$ and a working model is posited as $X = \gamma_0 + \gamma_1 \sin W + \gamma_2 Z + e$, where e is a random error term

c. The cut-off points are $(\mu_{\hat{X}} - \sigma_{\hat{X}}, \mu_{\hat{X}} + \sigma_{\hat{X}})$ based on the working model, $X = \gamma_0 + \gamma_1 \sin W + \gamma_2 Z + e$

4.6 Real data application

In this section, we apply the proposed method to the CPP study in Niswander and Gordon (1972), Gray et al. (2005), and Zhou et al. (2002). From 1959 to 1966, more than 44,000 pregnant women were enrolled in the study resulting in 55,908 pregnancies including multiple pregnancies. Pregnant mothers were enrolled through twelve U.S medical centers, and blood samples were collected from each mother at each prenatal visit. Children in the study were followed for various *in utero* developmental outcomes for up to 8 years. Two eligibility conditions were 1) liveborn singletons and 2) availability of a 3 ml third trimester maternal serum specimen. Out of 55,908 pregnancies, 44,075 satisfied the eligibility criteria. The main interest was to determine the relationship between the *in utero* exposure of PCBs and children's IQ score. PCBs were measured by analyzing the third-trimester blood serum specimens that had been preserved from mothers. The effect of mother's maternal pregnancy serum level of poly-chlorinated biphenyls (PCBs) on her child's IQ-test performance up to the age of 8 years is studied. In addition, other variables are considered such as the socio-economic status of the children's family (SES), the gender (SEX) and race of the child (RACE), the mother's education (EDU), and age (AGE). Due to the cost of assaying the PCB exposure from all subjects, Gray et al. (2005) and Zhou et al. (2002) used the ODS design with respect to PCB. Zhou et al. (2002) drew an ODS sample with the size of $(n_0 = 849, n_1 = 81, n_3 = 108)$ out of 38,709 subjects having complete data for the variables above except PCB. We use the SRS sample in Zhou et al. (2002) as the study population. Thus, we take 849 subjects in the population set. In addition we assume that we don't have IQ and PCB in the population dataset. However, we assume that we have a continuous auxiliary variable W for PCB, which is generated by W = PCB + $LN(1, 1)$ for all subjects in the population. To apply the proposed method, we draw a simple random sample of 100 subjects out of 849 underlying population that does not have PCB but has W for all subjects in the population. With the simple random sample, we take a log-transformation on PCB to use the following working model below. Based on the following

working model between log(PCB) and W,

$$\log(\text{PCB}) = \alpha_0 + \alpha_1 \log(W) + e,$$

where e follows a normal distribution with zero-mean and unknown variance, we obtain $(\hat{\alpha}_0, \hat{\alpha}_1)$, and estimate PCBs for all subjects in the population. Note that since subjects in the SRS sample already have PCBs, we estimated PCBs only for the remaining subjects in the population. After obtaining $\widehat{\text{PCB}}$, we select two supplemental ACSS samples each of size 100 from the rest of the population with defined cut-off points, $(\mu_{log(\widehat{\text{PCB}})} - \sigma_{log(\widehat{\text{PCB}})}, \mu_{log(\widehat{\text{PCB}})} + \sigma_{log(\widehat{\text{PCB}})}) = (0.88, 1.18)$ where $\mu_{log(\widehat{\text{PCB}})}$ and $\sigma_{log(\widehat{\text{PCB}})}$. The fitted linear model for CPP data analysis is

$$IQ = \beta_0 + \beta_1 \text{PCB} + \beta_2 \text{EDU} + \beta_3 \text{SES} + \beta_4 \text{AGE} + \beta_5 \text{BLACK} + \beta_6 \text{SEX} + \varepsilon,$$

where ε is a random error term.

The results for the CPP data analysis are given in Tables 4.7-9. Note that $\widehat{\beta}_P$ and $\widehat{\beta}_{IPW}$ use the same sample from ACSS scheme. In Tables 4.7 and 4.8 we summarize the descriptive statistics for IQ score and covariates. Note that for all subjects in the study population, PCB is given to see how well our working model estimates PCB. In Table 4.9, we compare the performance of two competing estimators. First, we find that the PCB level of mother's third-trimester blood serum specimen is insignificantly related to the IQ scores for children at 7 years of age. Second, $\widehat{\beta}_P$ has a narrower 95% confidence interval than that of $\widehat{\beta}_{IPW}$. For example, the 95% confidence interval for the estimates of PCB coefficients corresponding to $\widehat{\beta}_P$ is $(-0.70, 0.57)$ and for $\widehat{\beta}_{IPW}$ is $(-1.05, 0.59)$. Third, the estimators for the remaining covariates under two estimators considered have the same directions for all variables except for the AGE variable. $\widehat{\beta}_P$ and $\widehat{\beta}_{IPW}$, two estimators confirm that EDU and SES have positive impacts on the IQ scores of children while the RACE and SEX have negative effect on the IQ scores. In addition, PCB, AGE, and SEX are not statistically significant based on 95% confidence intervals.

Table 4.8: Summary table by Race and Gender in the population

	Gender		
Race	Female	Male	Total
Black	206	208	414
White & Other	206	229	435
Total	412	437	849

Table 4.9: Descriptive statistics for the continuous variables

Number of subjects	Variable	Mean	Std.Dev.	Minimum	Maximum
N = 849	IQ	95.45	14.01	59.00	142.00
	$\widehat{\text{PCB}}$	2.87	1.44	0.27	16.34
	EDU	10.72	2.30	1.00	18.00
	SES	4.69	2.10	0.30	9.30
	AGE	24.41	6.26	13.00	45.00
n_v = 100	IQ	97.1	14.24	62.00	136.00
	PCB	3.21	1.71	0.90	10.64
	EDU	11.04	2.41	1.00	18.00
	SES	4.91	2.08	1.00	9.30
	AGE	23.91	5.91	14.00	41.00

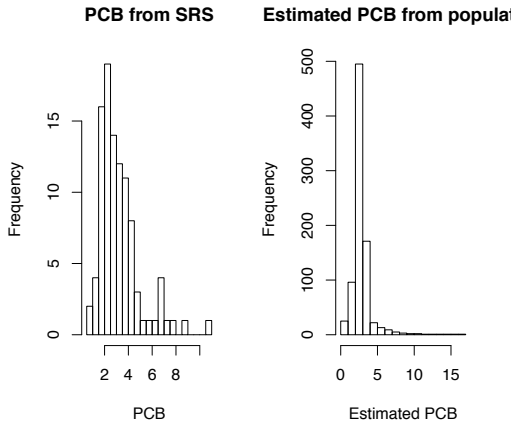

Figure 4.1: Comparison between PCB from SRS and $\widehat{\text{PCB}}$ in the population

Table 4.10: Analysis results for the CPP data set with n_v = 300

	$\hat{\beta}_{IPW}$			$\hat{\beta}_p$		
	$\hat{\beta}$	$\widehat{SE}$	95% CI	$\hat{\beta}$	$\widehat{SE}$	95% CI
Int	88.73	7.51	(73.99, 103.47)	76.56	4.82	(67.10, 86.02)
PCB	-0.22	0.42	(-1.05, 0.59)	-0.06	0.32	(-0.70, 0.57)
EDU	1.30	0.48	(0.35, 2.26)	1.49	0.35	(0.80, 2.18)
SES	1.41	0.59	(0.25, 2.57)	1.38	0.41	(0.56, 2.20)
AGE	-0.30	0.17	(-0.63, 0.03)	0.01	0.13	(-0.25, 0.28)
RACE	-7.87	2.22	(-12.24, -3.50)	-6.12	1.50	(-9.07, -3.18)
SEX	-2.72	1.94	(-6.53, 1.08)	-0.89	1.43	(-3.71, 1.91)

1. The allocation pattern is (n_0, n_1, n_3) = $(100, 100, 100)$.
2. The fitted model is IQ = β_0 + β_1PCB + β_2EDU + β_3SES + β_4AGE + β_5BLACK + β_6SEX + ε, where ε is a zero-mean normal variable with unknown variance
3. $\hat{\beta}_{IPW}$ and $\hat{\beta}_P$ are described in section 4.5

4.7 Discussion

We have proposed an auxiliary covariate dependent sampling design with an estimated likelihood approach. In terms of sampling design, since we assumed that one does not have measurements of the main covariate and response at the beginning of a prospective cohort study, we draw a simple random sample and estimate main covariate for all subjects in the study population. Then, we draw supplemental ACSS samples from the upper tail and lower tail of the estimated covariate distribution. To calculate selection probability from ACSS design, we used a plug-in method with a simple random sample in the first stage under an assumed working model. Then, we substitute the selection probabilities with the estimated selection probabilities in the likelihood function. To maximize the estimated likelihood, we used Nlopt algorithm and could obtain optimized estimators under the considered linear regression model.

The main advantage of the proposed design is that it allows researchers to use an auxiliary variable when there is no available measurement for a covariate of interest or response about a study population. This design is especially useful when one doesn't have enough information

about a prospective cohort population. If a prospective cohort study starts with insufficient information about main covariate and response, one relies on information that can be easily obtained through the population.

For practical use of the proposed method, one must be cautious when drawing samples that are too extreme in the distribution since some data would not have enough observations in two extreme tails. If supplemental samples are drawn from the very extreme tails that we recommend a between 1 and 1.5. In addition, when one estimates covariate of interest from a simple random sample, choosing a working model should be carefully done. When it comes to asymmetric covariates and auxiliary covariate, one need to draw supplemental samples with enough sample sizes from each stratum to take advantage of stratified sampling designs. In addition we conducted simulation studies with the correctly specified $k(X \mid W)$ and the misspecified $k(X \mid W)$. If the the working model is misspecified, we could not take advantage of the stratified sampling design even though we still obtain unbiased estimators.

For future work, we suggest two possible studies. First, after drawing a sample with ACSS, the rest of the subjects still remain in the study population. This implies that the smaller sample size from ACSS one has, the more subjects remain in the study population. Thus, the augmented IPW in Robins et al. (1994) method can be explored to include more information and number of subjects into the inferences. Second, instead of using the plug-in estimator, other approaches to calculate $\hat{P}(X \leq c)$ would be interesting to explore bias-variance tradeoff in comparing them to the proposed method. Lastly, it might be an interesting study to develop

4.8 Proof of Theorems

4.8.1 Regular conditions

The regular conditions required in proving the theorems in Section 4.5 are as followings :

Condition 12. *The parameter space, B, is a compact subset of R^p ; β_*, the true value, lies in the interior of β*

Condition 13. *The log-density $log(Y \mid X : \beta)$ is twice continuously differentiable with respect to β.*

Condition 14. *The following expected value matrix is finite and positive definite at β_0*

$$E(-\frac{\partial^2 log(Y \mid X : \beta)}{\partial\beta\partial\beta^T})$$

Condition 15. *There exists a $\delta > 0$ such that, for the set $M = \{\beta \in B : |\beta - \beta_*| < \delta\}$,*

$$E\left\{sup_B\left|\frac{\partial^2 log(Y \mid X : \beta)}{\partial\beta\partial\beta^T}\right|\right\} < \infty$$

4.8.2 Proof

The outline of Theorem 7 and 8 is given as follows :

Proof. By the mean value theorem in Khuri (2003),

$$S(\hat{\beta}) = S(\beta_*) + \frac{\partial S(\widetilde{\beta})}{\partial\beta}(\hat{\beta} - \beta_*),$$

where $\widetilde{\beta}$ is some value between $\hat{\beta}$ and β_*.

Recall that $S(\hat{\beta}) = 0$. Thus,

$$0 = S(\beta_*) + \frac{\partial S(\widetilde{\beta})}{\partial\beta}(\hat{\beta} - \beta_*)$$

Now consider $\sqrt{n_v}(\hat{\beta} - \beta_*)$. Then, we have

$$\sqrt{n_v}(\hat{\beta} - \beta_*) = \frac{\frac{1}{\sqrt{n_v}} S(\beta_*)}{\frac{-1}{n_v} \frac{\partial S(\widetilde{\beta})}{\partial \beta}} \tag{4.8}$$

Note that

$$\frac{1}{\sqrt{n_v}} S(\beta_*) = \frac{1}{\sqrt{n_v}} \sum_{i=1} \frac{\partial \log f_\beta(y_i \mid x_i)}{\partial \beta}|_{\beta_*}$$

By the Central Limit Theorem in Lehmann (1999), we have

$$\frac{1}{\sqrt{n_v}} \sum_{i=1} \frac{\partial \log f_\beta(y_i \mid x_i)}{\partial \beta}|_{\beta_*} \to_d N(0, \Sigma), \tag{4.9}$$

where

$$\Sigma = \frac{-1}{n_v} \sum_{i=1}^{n_v} E \frac{\partial S(\beta)}{\partial \beta}|_{\beta_*} = I(\beta_*)$$

Now consider the denominator in (4.10). By the Strong Law of Large Number in Lehmann (1999),

$$\frac{-1}{n_v} \frac{\partial S(\widetilde{\beta})}{\partial \beta} = \frac{-1}{n_v} \sum_{i=1}^{n_v} \frac{\partial^2 \log f_\beta(y_i \mid x_i)}{\partial \beta \partial \beta^T|_{\beta_*}} \to_p -I(\beta_*) \tag{4.10}$$

Hence, by multiplying (4.10) by (4.9), we could derieve

$$\sqrt{n_v}(\hat{\beta} - \beta_*) \sim N(0, I(\beta_*))$$

□

CHAPTER 5: SUMMARY AND FUTURE RESEARCH

In many epidemiological studies, efficient sampling design is important to save cost. This dissertation focuses on developing efficient estimators using an auxiliary variable that has information for the main covariate. Since an auxiliary variable is easy to obtain and informative of covariate of interest, subjects with missing covariate(or response) can be included in the statistical inference that it can enhances the efficiency of the estimators. Through this dissertation, we propose three estimators according to different study designs.

In Chapter 2, we have considered an ODS sample which has a missing covariate of interest but a binary auxiliary variable for the covariate. We considered a linear regression model and propose an estimated likelihood estimator reflecting the ODS sample with a missing covariate. In the proposed estimator, we used a binary auxiliary variable to improve statistical efficiency. Based on the likelihood function reflecting ODS design with missing covariate subjects, we derived an estimated likelihood function using the plug-in method to estimate the non-parametric part of the likelihood function. We showed that our estimator was consistent and asymptotically normal. Simulation studies were conducted and showed that the proposed method could produce more efficient estimator than the Inverse Probability Weighted estimator and estimator from the simple random sample with the same sample size. Application to CPP study compared the proposed method to MSELE in Zhou et al. (2002) and showed that the proposed method is more efficient than MSELE since the proposed method could include subjects having missing covariate. For future work, one can think of different missing data mechanism or continuous auxiliary variable cases. Since our proposed method considered missing completely at random (MCAR), we can extend our study to the case of MAR

mechanism. In that case, we need to construct a new likelihood reflecting MAR mechanism and a new estimated likelihood as well. In addition, developing an estimated likelihood for the case of a continuous auxiliary variable might be an interesting. The kernel type of estimators can be considered to handle it and characteristics of ODS design need to be considered.

In Chapter 3, a two-stage ODS design with a continuous auxiliary variable was studied. A similar setting was studied in Weaver and Zhou (2005) without an auxiliary variable for a covariate of interest. We proposed a method that combines the two estimators together: One is the semiparametric empirical likelihood method in Qin and Lawless (1994), Zhou et al. (2002) and Zhou et al. (2014); the other one is the updated method in Chen and Chen (2000) and Jiang and Haibo (2007). We exploited the semiparametric empirical likelihood approach to the dataset from ODS in the 2nd stage. We updated estimators from the semiparametric empirical likelihood approach using an auxiliary variable for the main covariate. Since the auxiliary variable is assumed to be observed for all subjects in the study, we can have more efficient estimators than estimators from complete observations only. Consistency and asymptotic normality of the proposed estimator were provided. From simulation studies, we found that the proposed estimator has smaller SE than SE of other competitors. In addition, the stronger the relationship between covariate and auxiliary variables, the more efficient estimators we obtained. CPP study was used to implement the proposed method and SES(mother's Social Economic Status) was used as the auxiliary variable for PCB. For future research, it might be interesting to compare the proposed method with other classes of estimators. The doubly robust estimator in Robins et al. (1994) would be a good competitor because missing probability can also be calculated through two-stage ODS design in the data structure in Chapter 2.

In Chapter 4, different from ODS, we considered a stratified sampling design using an auxiliary covariate. Since we assume a missing a covariate at the beginning of a prospective study, we estimated covariate of interest for all subjects in a study using the auxiliary variable

from a simple random sample and chose cut off values from the estimated covariate in the study population. With those cut-off values, we draw supplemental samples from the low cut-off point and the high cut-off point, respectively. We use the maximum likelihood approach to obtain estimators under our stratified sampling design. We could show consistency and asymptotic normality of the proposed estimator. Through simulations studies, we found that one could obtain more efficient estimators when one had more observations from supplemental samples under a fixed total sample size. In addition, when there is a stronger relationship between a covariate and auxiliary variable, the more efficient estimator we have. In addition, we could find that with a misspecified working model, we might not expect efficiency gain from the proposed method compared to efficiency gain with a correctly specified model. Development of an augmented inverse probability estimator might be valuable for future research that one can include subjects with a missing covariate to improve efficiency with our proposed auxiliary covariate dependent sampling design.

BIBLIOGRAPHY

Breslow, N. and Cain, K. (1988), 'Logistic regression for two-stage case-control data', *Biometrika* **75**(1), 11–20.

Breslow, N. E., Amorim, G., Pettinger, M. B. and Rossouw, J. (2013), 'Using the whole cohort in the analysis of case-control data', *Statistics in biosciences* **5**(2), 232–249.

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

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.

Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E. and Kulich, M. (2009), 'Using the whole cohort in the analysis of case-cohort data', *American Journal of Epidemiology* p. kwp055.

Breslow, N. E. and Wellner, J. A. (2007), 'Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression', *Scandinavian Journal of Statistics* **34**(1), 86–102.

Breslow, N. and Powers, W. (1978), 'Are there two logistic regressions for retrospective studies?', *Biometrics* pp. 100–105.

Carroll, R., Gail, M. and Lubin, J. (1993), 'Case-control studies with errors in covariates', *Journal of the American Statistical Association* **88**(421), 185–199.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006), *Measurement error in nonlinear models: a modern perspective*, CRC press.

Carroll, R. J. and Wand, M. (1991), 'Semiparametric estimation in logistic measurement error models', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 573–585.

Casella, G. and Berger, R. (2002), *Statistical Inference*, Duxbury advanced series in statistics and decision sciences, Thomson Learning.
URL: *https://books.google.com/books?id=0x_vAAAAMAAJ*

Chatterjee, N. and Carroll, R. J. (2005), 'Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies', *Biometrika* **92**(2), 399–418.

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.

Chen, Y.-H. and Chen, H. (2000), 'A unified approach to regression analysis under double-sampling designs', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(3), 449–460.

Cochran, W. G. (2007), *Sampling techniques*, John Wiley & Sons.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.

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.

Haneuse, S. J. et al. (2008), 'The combination of ecological and case–control data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 73–93.

Holt, D., Smith, T. and Winter, P. (1980), 'Regression analysis of data from complex surveys', *Journal of the Royal Statistical Society. Series A (General)* pp. 474–487.

Horvitz, D. G. and Thompson, D. J. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American statistical Association* **47**(260), 663–685.

Jiang, J. and Haibo, Z. (2007), 'Additive hazard regression with auxiliary covariates', *Biometrika* **94**(2), 359–369.

Khuri, A. (2003), *Advanced Calculus with Applications in Statistics*, Wiley Series in Probability and Statistics, Wiley.
URL: *https://books.google.com/books?id=WLVKP_TZ2YIC*

Landsman, V. and Graubard, B. (2013), 'Efficient analysis of case-control studies with sample weights', *Statistics in medicine* **32**(2), 347–360.

Lawless, J., Kalbfleisch, J. and Wild, C. (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.

Lehmann, E. L. (1999), *Elements of large-sample theory*, Springer Science & Business Media.

Little, R. J. and Rubin, D. B. (2014), *Statistical analysis with missing data*, John Wiley & Sons.

Ma, Y. et al. (2010), 'A semiparametric efficient estimator in case-control studies', *Bernoulli* **16**(2), 585–603.

Mann, C. (2003), 'Observational research methods. research design ii: cohort, cross sectional, and case-control studies', *Emergency medicine journal* **20**(1), 54–60.

Mariani, A. W. and Pego-Fernandes, P. M. (2014), 'Observational studies: why are they so important?', *Sao Paulo Medical Journal* **132**(1), 01–02.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models*, Vol. 37, CRC press.

Mukherjee, B. and Chatterjee, N. (2008), 'Exploiting gene-environment independence for analysis of case–control studies: An empirical bayes-type shrinkage estimator to trade-off between bias and efficiency', *Biometrics* **64**(3), 685–694.

Murphy, S. A. and Van der Vaart, A. W. (2000), 'On profile likelihood', *Journal of the American Statistical Association* **95**(450), 449–465.

Niswander, K. and Gordon, M. (1972), 'The women and their pregnancies. washington:(dc): Us department of health, education, and welfare', *Public Health Service, National Institutes of Health* .

Owen, A. B. (1988), 'Empirical likelihood ratio confidence intervals for a single functional', *Biometrika* **75**(2), 237–249.

Owen, A. B. (2001), *Empirical likelihood*, CRC press.

Pepe, M. S. and Fleming, T. R. (1991), 'A nonparametric method for dealing with mismeasured covariate data', *Journal of the American Statistical Association* **86**(413), 108–113.

Prentice, R. (1982), 'Covariate measurement errors and parameter estimation in a failure time regression model', *Biometrika* **69**(2), 331–342.

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

Prentice, R. L. and Pyke, R. (1979), 'Logistic disease incidence models and case-control studies', *Biometrika* **66**(3), 403–411.

Qin, G. and Zhou, H. (2010), 'Partial linear inference for a 2-stage outcome-dependent sampling design with a continuous outcome', *Biostatistics* p. kxq070.

Qin, J. and Lawless, J. (1994), 'Empirical likelihood and general estimating equations', *The Annals of Statistics* pp. 300–325.

Qin, J., Zhang, B. and Leung, D. H. (2012), 'Empirical likelihood in missing data problems', *Journal of the American Statistical Association* .

Reilly, M. and Pepe, M. S. (1995), 'A mean score method for missing and auxiliary covariate data in regression models', *Biometrika* **82**(2), 299–314.

Robins, J. M., 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)* pp. 409–424.

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.

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996), 'A semiparametric mixture approach to case-control studies with errors in covariables', *Journal of the American Statistical Association* **91**(434), 722–732.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

Scott, A. J., Lee, A. J. and Wild, C. (2007), 'On the breslow–holubkov estimator', *Lifetime data analysis* **13**(4), 545–563.

Scott, A. J. and Wild, C. (1986), 'Fitting logistic models under case-control or choice based sampling', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 170–182.

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

Scott, A. J. and Wild, C. J. (2011), 'Fitting regression models with response-biased samples', *Canadian Journal of Statistics* **39**(3), 519–536.

Scott, A. and Wild, C. (1991), 'Fitting logistic regression models in stratified case-control studies', *Biometrics* pp. 497–510.

Shao, J. (2003), *Mathematical Statistics*, Springer Science Business Media, Springer.
URL: *https://books.google.com/books?id=cyqTPotl7QcC*

Song, J. W. and Chung, K. C. (2010), 'Observational studies: cohort and case-control studies', *Plastic and reconstructive surgery* **126**(6), 2234.

Song, R., Zhou, H. and Kosorok, M. R. (2009), 'A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome', *Biometrika* p. asn073.

Suissa, S. (1991), 'Binary methods for continuous outcomes: a parametric alternative', *Journal of clinical epidemiology* **44**(3), 241–248.

Tsiatis, A. A. and Ma, Y. (2004), 'Locally efficient semiparametric estimators for functional measurement error models', *Biometrika* **91**(4), 835–848.

Wang, W., Scharfstein, D., Tan, Z. and MacKenzie, E. J. (2009), 'Causal inference in outcome-dependent two-phase sampling designs', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(5), 947–969.

Wang, X., Wu, Y. and Zhou, H. (2009), 'Outcome-and auxiliary-dependent subsampling and its statistical inference', *Journal of biopharmaceutical statistics* **19**(6), 1132–1150.

Wang, X. and Zhou, H. (2006), 'A semiparametric empirical likelihood method for biased sampling schemes with auxiliary covariates', *Biometrics* **62**(4), 1149–1160.

Wang, X. and Zhou, H. (2010), 'Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling', *Biometrics* **66**(2), 502–511.

Wasserman, L. (2006), *All of nonparametric statistics*, Springer Science & Business Media.

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.

White, J. E. (1982), '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.

Wild, C. (1991), 'Fitting prospective regression models to case-control data', *Biometrika* **78**(4), 705–717.

Yu, Y. and Ruppert, D. (2002), 'Penalized spline estimation for partially linear single-index models', *Journal of the American Statistical Association* **97**(460), 1042–1054.

Zeng, D. and Lin, D. (2006), 'Efficient estimation of semiparametric transformation models for counting processes', *Biometrika* **93**(3), 627–640.

Zhao, L. and Lipsitz, S. (1992), 'Designs and analysis of two-stage studies', *Statistics in medicine* **11**(6), 769–782.

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 (Cambridge, Mass.)* **18**(4), 461.

Zhou, H. and Pepe, M. S. (1995), 'Auxiliary covariate data in failure time regression', *Biometrika* **82**(1), 139–149.

Zhou, H., Song, R., Wu, Y. and Qin, J. (2011), 'Statistical inference for a two-stage outcome-dependent sampling design with a continuous outcome', *Biometrics* **67**(1), 194–202.

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

Zhou, H., Wu, Y., Liu, Y. and Cai, J. (2011), 'Semiparametric inference for a 2-stage outcome-auxiliary-dependent sampling design with continuous outcome', *Biostatistics* **12**(3), 521–534.

Zhou, H., Xu, W., Zeng, D. and Cai, J. (2014), 'Semiparametric inference for data with a continuous outcome from a two-phase probability-dependent sampling scheme', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1), 197–215.

Zhou, H., You, J., Qin, G. and Longnecker, M. P. (2011), 'A partially linear regression model for data from an outcome-dependent sampling design', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60**(4), 559–574.