DESIGNS AND ANALYSIS OF TWO-PHASE STUDIES, WITH APPLICATIONS TO GENETIC ASSOCIATION STUDIES

Ran Tao

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
2016

Approved by:

Danyu Lin

Donglin Zeng

Kari E. North

Yun Li

Quefeng Li

ABSTRACT

Ran Tao: Designs and Analysis of Two-Phase Studies, with Applications to Genetic Association Studies
(Under the direction of Danyu Lin and Donglin Zeng)

The two-phase design is a cost-effective sampling strategy when investigators are interested in evaluating the effects of covariates on an outcome but certain covariates are too expensive to be measured on all study subjects. Under such a design, the outcome of interest and the covariates that are inexpensive to measure are observed for all subjects during the first phase, and the first-phase information is used to select subjects for measurements of "expensive covariates" during the second phase. This design greatly reduces the cost associated with the collection of expensive covariate data and thus has been widely used in large epidemiological studies.

In two-phase studies, if the second-phase selection depends on multiple outcomes, then one should consider all of them simultaneously in a multivariate regression model in order to obtain valid inference. We develop an efficient likelihood-based approach to making inference under multivariate outcome-dependent sampling. We implement a computationally efficient expectation-maximization algorithm and establish the theoretical properties of the resulting maximum likelihood estimators. We demonstrate the superiority of the proposed methods over standard linear regression through extensive simulation studies. We provide applications to two large-scale sequencing studies.

In two-phase studies, the "inexpensive covariates" can be used to improve the design efficiency of second-phase sampling and control for confounding. However, accommodating continuous inexpensive covariates that are correlated with expensive covariates is very

challenging because the likelihood function involves the conditional density functions of expensive covariates given continuous inexpensive covariates. We develop a semiparametric approach to regression analysis by approximating the conditional density functions with B-spline sieves. We establish the theoretical properties of the resulting estimators. We demonstrate the superiority of the proposed methods over existing ones through extensive simulation studies. We provide applications to a large-scale whole-exome sequencing study.

Previous research on two-phase studies has largely focused on the inference procedures rather than the design aspects of two-phase studies. An important topic of investigation is the optimal study design when the primary interest is to estimate the regression coefficients of the expensive covariates. We derive optimal two-phase designs, which can be substantially more efficient than the current designs.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisers Drs. Danyu Lin and Donglin Zeng. They gave me great guidance, support, and encouragement throughout my graduate studies. I have learned a great deal from their unique perspective on research, their sharp insight into statistical problems, and their passions. All of these will inspire me to strive for excellence in my future career.

I would like to give sincere thanks to my committee members: Drs. Kari E. North, Yun Li, and Quefeng Li. I appreciate them for reading the manuscript and offering insightful comments and suggestions, which have led to significant improvements of the dissertation.

I am deeply grateful to my graduate research assistant supervisors Drs. Danyu Lin and Kari E. North. They have offered me great opportunities to participate in several large-scale genetic studies, through which I developed my collaborative and interdisciplinary research skills. In addition, their generous financial supports have helped me through my entire graduate study.

Last but not the least, I would like to extend my gratitude to my parents, my wife, and all my friends for their unconditional love, support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In epidemiological studies, the outcomes of interest (e.g, anthropometry measurements, lipids levels, or disease status) and demographical and environmental variables (e.g., age, gender, and smoking status) are typically available for all subjects. However, the covariates of main interest often involve genotyping, biomarker assay, or medical imaging and thus are prohibitively expensive to measure for all subjects, especially in a large study. If disease status or another discrete outcome is of primary interest, then the case-control design with an equal number of cases and controls is the most efficient one (Scott and Wild 1997). If a continuous outcome such as height is of primary interest, then a cost-effective strategy is the "extreme-tail" sampling design, whereby one selectively measures the "expensive covariates" only for subjects with extreme values of the primary outcome measure (Lin et al. 2013). In either case, the efficiency of the design can be improved by stratifying on the "inexpensive covariates".

The case-control and extreme-tail sampling designs can be viewed as special cases of the two-phase, outcome-dependent sampling design, which was first introduced by White (1982). In the first phase of such a design, the outcomes of interest and inexpensive covariates are observed for all study subjects; the information collected during the first phase is then used to determine which subjects to include for measurements on expensive covariates during the second phase. This design greatly reduces the cost and other practical burdens associated with the collection of expensive covariate data and thus has been widely used in large epidemiological studies.

One recent example of the two-phase design is the National Heart, Lung, and Blood

Institute (NHLBI) Exome Sequencing Project (ESP), where 4494 subjects from seven cohorts were selected for whole-exome sequencing (Lin et al. 2013). Among these subjects, 659, 806, and 657 were selected because of extremely high or low values of body mass index (BMI), blood pressure (BP) adjusted for age, gender, race, BMI, and anti-hypertensive medication, and low-density lipoprotein (LDL) adjusted for age, gender, race, and lipid medication, respectively.

In this dissertation, we develop novel statistical methods to solve problems arising in the design and analysis of two-phase studies, and provide applications to genetic association studies. The outline is as follows. In Chapter 2, we conduct a comprehensive literature review on existing methods for the designs and analysis of two-phase studies. In Chapter 3, we develop an efficient likelihood-based approach to making inference under multivariate outcome-dependent sampling. In Chapter 4, we develop efficient semiparametric inference procedures for general two-phase studies. In Chapter 5, we study optimal two-phase designs and point out some future directions.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Let Y denote the outcome of interest, $\boldsymbol{X}$ denote the vector of expensive covariates that is measured on a fraction of subjects in the study, $\boldsymbol{Z}$ denote the vector of inexpensive covariates that is potentially correlated with $\boldsymbol{X}$, and $\boldsymbol{W}$ denote the vector of inexpensive covariates that is known to be independent of $\boldsymbol{X}$ given $\boldsymbol{Z}$. The data $(Y, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$ are assumed to be generated from the joint distribution $P_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})P(\boldsymbol{X}|\boldsymbol{Z})P(\boldsymbol{Z}, \boldsymbol{W})$, where $P_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$ is a parametric regression model indexed by parameter $\boldsymbol{\theta}$, $P(\boldsymbol{X}|\boldsymbol{Z})$ is the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Z}$, and $P(\boldsymbol{Z}, \boldsymbol{W})$ is the joint distribution of $\boldsymbol{Z}$ and $\boldsymbol{W}$.

Under the two-phase design, $(Y, \boldsymbol{Z}, \boldsymbol{W})$ is measured for all n subjects in the first phase, and $\boldsymbol{X}$ is measured for a sub-sample of size n_2 in the second phase. Let R indicate, by the values 1 versus 0, whether the subject is selected for the measurement of $\boldsymbol{X}$ in the second phase or not. The key assumption for any two-phase design is that the distribution of R depends on $(Y, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$ only through the first-phase data $(Y, \boldsymbol{Z}, \boldsymbol{W})$. Under this assumption, the data on $\boldsymbol{X}$ are missing at random, such that the sampling indicators $(R_1, \ldots, R_n)$ can be omitted from the likelihood function when estimating $\boldsymbol{\theta}$. Thus, the observed-data likelihood takes the form

$$\prod_{i=1}^{n} \left\{ P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)P(\boldsymbol{X}_i|\boldsymbol{Z}_i) \right\}^{R_i} \left\{ \log \int P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}, \boldsymbol{Z}_i, \boldsymbol{W}_i)P(\boldsymbol{x}|\boldsymbol{Z}_i)d\boldsymbol{x} \right\}^{1-R_i}. \qquad (2.1)$$

In this chapter, we first review existing methods developed for regression analysis of two-phase studies with a single outcome. These methods are classified into two categories

depending on whether they used the first-phase information for subjects not selected during the second-phase or not. Then, we review existing methods for multiple outcomes. Finally, we review existing literature on design efficiencies of two-phase studies.

2.2 Methods for Analyzing Two-Phase Studies with a Single Outcome

2.2.1 Methods Using Second-Phase Subjects Only

If the first-phase information is not available for subjects not selected during the second phase, then the resulting likelihood is

$$\begin{aligned} &\prod_{i:R_i=1} P(Y_i, \boldsymbol{X}_i | \boldsymbol{Z}_i, \boldsymbol{W}_i, R_i = 1) \\ &= \prod_{i:R_i=1} \frac{P_{\boldsymbol{\theta}}(Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) P(\boldsymbol{X}_i | \boldsymbol{Z}_i) P(R_i = 1 | Y_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)}{P(R_i = 1 | \boldsymbol{Z}_i, \boldsymbol{W}_i)}, \end{aligned} \tag{2.2}$$

where $P(R = 1|\boldsymbol{Z}, \boldsymbol{W}) = \int P_{\boldsymbol{\theta}}(y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) P(\boldsymbol{x}|\boldsymbol{Z}) P(R = 1|y, \boldsymbol{Z}, \boldsymbol{W}) d\boldsymbol{x} dy$. We can also write down a conditional likelihood that does not involve $P(\boldsymbol{x}|\boldsymbol{Z})$:

$$\prod_{i:R_i=1} P(Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i, R_i = 1) = \prod_{i:R_i=1} \frac{P_{\boldsymbol{\theta}}(Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) P(R_i = 1 | Y_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)}{P(R_i = 1 | \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)}, \tag{2.3}$$

where $P(R = 1|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = \int P_{\boldsymbol{\theta}}(y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) P(R = 1|y, \boldsymbol{Z}, \boldsymbol{W}) dy$. Note that when both the outcome and inexpensive covariates are discrete, one can show that expression (2.3) is the semiparametric profile likelihood of $\boldsymbol{\theta}$ obtained from expression (2.2) by using the maximization process employed in Wild (1991) and Scott and Wild (1997).

Estimators Based on the Prospective Likelihood

If the second-phase sampling is completely random or depends on the inexpensive covariates $(\boldsymbol{Z}, \boldsymbol{W})$ only, then $P(R_i = 1|Y_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) = P(R_i = 1|\boldsymbol{Z}_i, \boldsymbol{W}_i)$. Therefore, $P(R_i =$

$1|Y_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)$ and $P(R_i = 1|\boldsymbol{Z}_i, \boldsymbol{W}_i)$ cancel out in the numerator and denominator of expression (2.3), and it is efficient to base inferences about $\boldsymbol{\theta}$ on the prospective likelihood

$$\prod_{i:R_i=1} P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i). \tag{2.4}$$

If there are no inexpensive covariates $(\boldsymbol{Z}, \boldsymbol{W})$ and the second-phase sampling depends on a binary outcome, Anderson (1972) and Prentice and Pyke (1979) showed that standard logistic regression based on the prospective likelihood (2.4) gives valid inferences for all regression coefficients except for the intercept. In fact, Prentice and Pyke (1979) showed that the prospective likelihood (2.4) is the profile likelihood of $\boldsymbol{\theta}$ based on the conditional likelihood (2.2) with the marginal distribution of $\boldsymbol{X}$ maximized out nonparametrically. Unfortunately, this feature does not carry over to arbitrary regression models in general two-phase, outcome-dependent sampling studies. If the second-phase sampling depends on the outcome of interest, then estimators based on expressions (2.4) are generally biased.

Maximum Semiparametric Empirical Likelihood Estimator

When the outcome is continuous but the second-phase selection depends on a small number of strata, Zhou et al. (2002) proposed a maximum semiparametric empirical likelihood estimator based on maximizing expression (2.2). Suppose that the domain of Y can be partitioned into K mutually exclusive and exhaustive strata by the known constants $-\infty = a_0 < a_1 < \cdot < a_{K-1} < a_K = \infty$. A simple random sample of size n_k is drawn from the kth stratum $(k = 1, \ldots, K)$ during the second phase. Assuming further that there are no inexpensive covariates, the conditional likelihood (2.2) can be rewritten as

$$\begin{aligned}
&\prod_{k=1}^{K}\prod_{j=1}^{n_k} \frac{P_{\boldsymbol{\theta}}(Y_{kj}|\boldsymbol{X}_{kj})P(\boldsymbol{X}_{kj})}{F(a_k) - F(a_{k-1})} \\
=&\prod_{k=1}^{K}\prod_{j=1}^{n_k} \frac{P_{\boldsymbol{\theta}}(Y_{kj}|\boldsymbol{X}_{kj})}{F(a_k|X_{kj}) - F(a_{k-1}|X_{kj})} \prod_{k=1}^{K}\prod_{j=1}^{n_k} P(\boldsymbol{X}_{kj}) \frac{F(a_k|X_{kj}) - F(a_{k-1}|X_{kj})}{F(a_k) - F(a_{k-1})},
\end{aligned} \tag{2.5}$$

where $(Y_{kj}, \boldsymbol{X}_{kj})$ is the data for the jth subject in the kth stratum ($k = 1, \ldots, K$, $j = 1, \ldots, n_k$), $F(u) = P(Y \leq u)$, and $F(u|X) = P(Y \leq u|X)$. To estimate $\boldsymbol{\theta}$, Zhou et al. (2002) first profiled the likelihood function (2.5) by fixing $\boldsymbol{\theta}$ and obtaining the empirical likelihood function of $P(\boldsymbol{X})$ over all distributions whose support contain the observed X values. They then maximized the resulting profile likelihood function with respect to $\boldsymbol{\theta}$.

Weighted Estimators

If every study subject have a positive probability of being selected during the second phase, then the Horvitz-Thompson approach (Horvitz and Thompson 1952) commonly used in survey sampling can be adopted (Hsieh et al. 1985, Scott and Wild 1986, Kalbfleisch and Lawless 1988, Zhao and Lipsitz 1992, Whittemore 1997). If all variables had been fully observed for all n subjects, then the log-likelihood function would be $\sum_{i=1}^{n} \log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)$. An estimate of this quantity is obtained if we use the completely observed units only and weight their contributions inversely according to their probability of selection, i.e.,

$$\sum_{i=1}^{n} \frac{R_i}{\pi_i} \log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i), \tag{2.6}$$

where $\pi = P(R = 1|Y, \boldsymbol{Z}, \boldsymbol{W})$. The Horvitz-Thompson estimator of $\boldsymbol{\theta}$ is based on maximizing expression (2.6). It provides unbiased estimation of the overall association in all study subjects even when the regression model is misspecified. However, the Horvitz-Thompson estimator is inefficient, especially when the inclusion probabilities are highly variable, which is usually the case for an efficient two-phase design.

Efficiency can be improved by modifying the sampling weights. When the regression model is linear, Magee (1998) proposed to weight each subject selected during the second phase by $\{\pi a_{\boldsymbol{\alpha}}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})\}^{-1}$ instead of π^{-1}, where $a_{\boldsymbol{\alpha}}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$ belongs to a parameterized family of functions indexed by the vector parameter $\boldsymbol{\alpha}$. They showed that under certain moment assumptions, any estimator with positive weights $\{\pi a_{\boldsymbol{\alpha}}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})\}^{-1}$ is consistent

for $\boldsymbol{\theta}$. Therefore, one can chose the optimal $\boldsymbol{\alpha}$ that minimizes a scalar variance criterion such as the determinant or the trace of the asymptotic variance estimator. The choice of the function $a_{\boldsymbol{\alpha}}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$ is up to the analyst but the obvious idea is to choose a function that is believed to be approximately inversely proportional to the residual variance under the sample model.

Pfeffermann and Sverchkov (1999) proposed another modification. They showed that

$$\mathrm{E}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = \frac{\mathrm{E}(\pi^{-1}Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}, R=1)}{\mathrm{E}(\pi^{-1}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}, R=1)},$$

and proposed to use the weight $\pi^{-1}/\mathrm{E}(\pi^{-1}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}, R=1)$. This weight accounts only for the aspect of the second-phase selection process that is not determined by the covariates in the regression model. Because of the reduced variation of the weights, the resulting estimator tends to be more powerful than the Horvitz-Thompson estimator.

2.2.2 Methods Using All Study Subjects

Pseudo-Likelihood Estimators

If the first-phase information is available for all study subjects, then it can be utilized to improve efficiency. When the second-phase sampling is completely random or depends on the inexpensive covariates $(\boldsymbol{Z}, \boldsymbol{W})$ only, Pepe and Fleming (1991) and Carroll and Wand (1991) proposed to estimate θ by maximizing the likelihood

$$\prod_{i=1}^{n} \left\{ P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) \right\}^{R_i} \left\{ \int P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}, \boldsymbol{Z}_i, \boldsymbol{W}_i) P(\boldsymbol{x}|\boldsymbol{Z}_i) d\boldsymbol{x} \right\}^{1-R_i}, \tag{2.7}$$

where $P(\boldsymbol{x}|\boldsymbol{Z})$ is estimated nonparametrically using the second-phase sample alone. If $\boldsymbol{Z}$ is discrete, then Pepe and Fleming (1991) estimated $P(\boldsymbol{x}|\boldsymbol{Z})$ by

$$\widehat{P}(\boldsymbol{x}|\boldsymbol{Z}) = \widehat{P}(\boldsymbol{x}|\boldsymbol{Z}, R=1) = \left\{ \sum_{i=1}^{n} R_i I(\boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{Z}_i = \boldsymbol{Z}) \right\} \Big/ \left\{ \sum_{i=1}^{n} R_i I(\boldsymbol{Z}_i = \boldsymbol{Z}) \right\}. \tag{2.8}$$

If $\boldsymbol{Z}$ contains continuous components, then Carroll and Wand (1991) estimated $P(\boldsymbol{x}|\boldsymbol{Z})$ with kernel smoothing techniques. That is,

$$\widehat{P}(\boldsymbol{x}|\boldsymbol{Z}) = \widehat{P}(\boldsymbol{x}|\boldsymbol{Z}, R=1) = \frac{\sum_{i=1}^{n} R_i I(\boldsymbol{X}_i = \boldsymbol{x}) K(\|\boldsymbol{Z} - \boldsymbol{Z}_i\|/h)}{\sum_{i=1}^{n} R_i K(\|\boldsymbol{Z} - \boldsymbol{Z}_i\|/h)}, \tag{2.9}$$

where $K(\cdot)$ is a symmetric density function and h is the bandwidth. In addition, for scaler $\boldsymbol{Z}$, they obtained a representation for an optimal bandwidth through a detailed analysis of the mean-squared error of the parameter estimate.

Mean Score Estimator

When both the outcome and inexpensive covariates are discrete, Reilly and Pepe (1995) proposed a mean score estimator (MSE). It is based on solving the estimating equation

$$\sum_{i=1}^{n} R_i l_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) + \sum_{i=1}^{n} (1 - R_i) \mathrm{E}\{l_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}, \boldsymbol{Z}_i, \boldsymbol{W}_i)|Y_i, \boldsymbol{Z}_i, \boldsymbol{W}_i\} = 0, \tag{2.10}$$

where $l_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = \partial \log P_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})/\partial \boldsymbol{\theta}$. Reilly and Pepe (1995) proposed estimating $\mathrm{E}\{l_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})|Y, \boldsymbol{Z}, \boldsymbol{W}\}$ for a subject not selected during the second phase by $\int l_{\boldsymbol{\theta}}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) d\widehat{P}(\boldsymbol{x}|Y, \boldsymbol{Z})$, where $\widehat{P}(\boldsymbol{x}|Y, \boldsymbol{Z})$ is the empirical distribution of $\boldsymbol{X}$ given $(Y, \boldsymbol{Z})$ in the second-phase sample. This purely empirical mean score approach is valid because $P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z}) = P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z}, R=1)$.

Maximum Likelihood Estimators Assuming Discrete First-Phase Information

When both the outcome and inexpensive covariates are discrete, Scott and Wild (1997) proposed estimating $\boldsymbol{\theta}$ by maximizing the full likelihood (2.1). This maximum likelihood estimator (MLE) is the most efficient among all valid estimators. Breslow and Holubkov (1997) considered the special case of logistic regression.

For continuous first-phase data, Lawless et al. (1999) suggested to discretize them into

a small number of strata and then use the stratum membership to select subjects in the second phase. Specifically, suppose that the range of $(Y, \boldsymbol{Z}, \boldsymbol{W})$ is partitioned into K strata $\mathcal{S}_1, \ldots, \mathcal{S}_K$. The observed-data likelihood is

$$\prod_{j=1}^{K} \left\{ \prod_{i \in D_j} P_{\boldsymbol{\theta}}(Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) g(\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) \right\} Q_j(\boldsymbol{\theta}, G)^{n_j - n_{2j}}, \tag{2.11}$$

where $Q_j(\boldsymbol{\theta}, G) = \Pr\{(Y, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) \in \mathcal{S}_j\}$, $D_j = \{i\colon (Y, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) \in \mathcal{S}_j, R_i = 1\}$, $n_{2j} = |D_j|$, n_j is the total number of subjects in stratum $\mathcal{S}_j$, $j = 1, \ldots, K$, and $G(\cdot)$ and $g(\cdot)$ are the distribution and density functions corresponding to $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$, respectively. From expression (2.11), we see that for subjects not selected in the second phase, only the stratum membership is used in the inference. Breslow et al. (2003) established the asymptotic properties of the corresponding MLE. Note that the discretization of first-phase data for subjects not selected during the second-phase entails a substantial loss of information and may even bias parameter estimation.

Pseudo-Score Estimators

To improve efficiency, Chatterjee et al. (2003) proposed a pseudo-score estimator (PSE). It allows the outcome of interest to be continuous but require the inexpensive covariates to be discrete. This estimator of $\boldsymbol{\theta}$ is based on solving the estimating equation

$$\begin{aligned} &\sum_{i=1}^{n} R_i l_{\boldsymbol{\theta}}(Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) \\ &+ \sum_{i=1}^{n} (1 - R_i) \frac{\int l_{\boldsymbol{\theta}}(Y_i | \boldsymbol{x}, \boldsymbol{Z}_i, \boldsymbol{W}_i) P_{\boldsymbol{\theta}}(Y_i | \boldsymbol{x}, \boldsymbol{Z}_i, \boldsymbol{W}_i) P(\boldsymbol{x} | \boldsymbol{Z}_i) d\boldsymbol{x}}{\int P_{\boldsymbol{\theta}}(Y_i | \boldsymbol{x}, \boldsymbol{Z}_i, \boldsymbol{W}_i) P(\boldsymbol{x} | \boldsymbol{Z}_i) d\boldsymbol{x}} = 0, \end{aligned} \tag{2.12}$$

where the left-hand side is obtained by first taking log of expression (2.7) and then differentiating with respect to $\boldsymbol{\theta}$. Next, one needs to find a valid estimator of the conditional probability $P(\boldsymbol{x}|\boldsymbol{Z})$. If the second-phase sampling depends on the outcome, then

$P(\boldsymbol{x}|\boldsymbol{Z}) \neq P(\boldsymbol{x}|\boldsymbol{Z}, R = 1)$, and one cannot use expression (2.8), as in Pepe and Fleming (1991), to estimate $P(\boldsymbol{x}|\boldsymbol{Z})$ anymore. From Bayes's theorem, if $P(\boldsymbol{x}|\boldsymbol{Z}, R = 1) > 0$ almost surely, then

$$P(\boldsymbol{x}|\boldsymbol{Z}) = \frac{P(\boldsymbol{x}|\boldsymbol{Z}, R = 1)P(R = 1|\boldsymbol{Z}, \boldsymbol{W})}{P(R = 1|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W})}, \tag{2.13}$$

where $P(R = 1|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) = \int P(R = 1|y, \boldsymbol{Z}, \boldsymbol{W})P_{\boldsymbol{\theta}}(y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})dy$. Chatterjee et al. (2003) proposed to estimate $P(\boldsymbol{x}|\boldsymbol{Z})$ by using expression (2.13), where $P(\boldsymbol{x}|\boldsymbol{Z}, R = 1)$ is estimated by expression (2.8) and $P(R = 1|\boldsymbol{Z}, \boldsymbol{W})$ is ignored because it cancels out in the numerator and denominator of the second term in the left-hand side of expression (2.12).

In order to accommodate continuous covariates, Chatterjee and Chen (2007) considered the kernel smoothing approach similar to that considered by Carroll and Wand (1991). There are, however, some complications if the second-phase sampling depends on $\boldsymbol{Z}$. Specifically, if $\boldsymbol{Z}$ is partitioned into a fixed number of strata, such that subjects are sampled with different selection probabilities across different strata during the second phase, then the discontinuity of the selection probabilities would cause the conditional expectation function $\mathrm{E}(U|\boldsymbol{Z}, R = 1)$ for a random variable U to have jumps between strata, even though $\mathrm{E}(U|\boldsymbol{Z})$ could be continuous and smooth in the whole range of $\boldsymbol{Z}$. To account for these discontinuities, Chatterjee and Chen (2007) proposed to apply kernel smoothing within each partition of $\boldsymbol{Z}$ separately. However, if the second-phase sampling depends on the partitions of the residuals calculated from the regression model relating Y to $\boldsymbol{Z}$ and $\boldsymbol{W}$, then the corresponding partition of $\boldsymbol{Z}$ is hard to determine. In addition, it is difficult to calculate $P(R = 1|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W})$ in this case. Consequently, the PSE method of Chatterjee and Chen (2007) is only applicable when the second-phase sampling depends on only discrete covariates.

Maximum Estimated Likelihood Estimator

Weaver and Zhou (2005) proposed a maximum estimated likelihood estimator (MELE).

Similar to the PSE method of Chatterjee et al. (2003), it allows the outcome of interest to be continuous but requires the inexpensive covariates to be discrete and the second-phase selection to depend on a small number of strata. The MELE of $\boldsymbol{\theta}$ is based on maximizing expression (2.7), where $P(\boldsymbol{x}|\boldsymbol{Z})$ is estimated by

$$\widehat{P}(\boldsymbol{x}|\boldsymbol{Z}=\boldsymbol{z}_j)=\sum_{k=1}^{K}\widehat{P}_k(\boldsymbol{x}|\boldsymbol{Z}=\boldsymbol{z}_j)\frac{N_k(\boldsymbol{z}_j)}{N(\boldsymbol{z}_j)}. \tag{2.14}$$

Here

$$\widehat{P}_k(\boldsymbol{x}|\boldsymbol{Z}=\boldsymbol{z}_j)=\frac{\sum_{i\in S_k}R_iI(\boldsymbol{X}_i=\boldsymbol{x}_i,\boldsymbol{Z}_i=\boldsymbol{z}_j)}{\sum_{i\in S_k}R_iI(\boldsymbol{Z}_i=\boldsymbol{z}_j)},\quad k=1,\ldots,K, \tag{2.15}$$

and $N_k(\boldsymbol{z}_j)$ and $N(\boldsymbol{z}_j)$ are the numbers of observations in the population and in the kth stratum, respectively, that satisfy $\boldsymbol{Z}=\boldsymbol{z}_j$. Simulation studies in Weaver and Zhou (2005) showed that the MELE is consistently slightly less efficient than the PSE. However, they claimed that the MELE has computational advantages over the PSE.

Maximum Likelihood Estimator Assuming No Inexpensive Covariates

Both the PSE and MELE methods are statistically inefficient. Song et al. (2009) and Lin et al. (2013) considered efficient estimation for two-phase studies without inexpensive covariates. In this case, the observed-data likelihood (2.1) reduces to

$$\prod_{i=1}^{n}\left\{P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i)P(\boldsymbol{X}_i)\right\}^{R_i}\left\{\log\int P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x})P(\boldsymbol{x})d\boldsymbol{x}\right\}^{1-R_i}. \tag{2.16}$$

Song et al. (2009) and Lin et al. (2013) maximized expression (2.16) using the nonparametric maximum likelihood estimation, where $P(\boldsymbol{X})$ is estimated by the discrete probabilities at the observed values of $\boldsymbol{X}$. We denote this MLE approach by MLE_0. If the inexpensive covariates are available for all subjects but the second-phase selection does not depend on either $\boldsymbol{Z}$ or

$\boldsymbol{W}$, then the MLE_0 method can be adopted by redefining the "expensive covariates" as $(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}}, \boldsymbol{W}^{\mathrm{T}})^{\mathrm{T}}$ and disregarding $\boldsymbol{Z}$ and $\boldsymbol{W}$ for subjects not selected in the second phase. This data reduction approach may entail a substantial loss of information. If the second-phase selection does depend on $\boldsymbol{Z}$ and $\boldsymbol{W}$, then expression (2.16) no longer correctly reflects the sampling mechanism, and the MLE_0 method is generally biased.

Semiparametric Efficient Estimator

When every study subject have a positive probability of being selected during the second phase, Robins et al. (1995) derived the efficient score function S_{eff} for general two-phase studies with inexpensive covariates. Define $\boldsymbol{O} \equiv (Y, \boldsymbol{Z}, \boldsymbol{W})$, which is the first-phase information. They showed that $S_{\text{eff}} = U(\phi_{\text{op}})$, where, for any function $\phi(\boldsymbol{o})$ taking values in R^d, d is the dimension of $\boldsymbol{\theta}$,

$$
\begin{aligned}
U(\phi) &= U^{(1)} + U^{(2)}(\phi), \\
U^{(1)} &= R l_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) - R\mathrm{E}\left\{l_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})|\boldsymbol{X}, \boldsymbol{Z}, R=1\right\}, \\
U^{(2)}(\phi) &= -\pi^{-1} R\mathrm{E}\left\{(1-\pi)\phi(Y, \boldsymbol{Z}, \boldsymbol{W})|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}\right\} + (1-R)\phi(\boldsymbol{O}),
\end{aligned}
$$

and $\phi_{\text{op}}(\boldsymbol{O})$ is the unique solution to the functional equation

$$
\begin{aligned}
\phi(\boldsymbol{O}) =& \mathrm{E}\left\{l_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})|\boldsymbol{O}\right\} - \mathrm{E}\left[\frac{\mathrm{E}\left\{\pi(\boldsymbol{O}) l_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}\right\}}{\mathrm{E}\left\{\pi(\boldsymbol{O})|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}\right\}}\middle|\boldsymbol{O}\right] \\
& - \mathrm{E}\left[\frac{\mathrm{E}\left\{(1-\pi(\boldsymbol{O}))\phi(\boldsymbol{O})|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}\right\}}{\mathrm{E}\left\{\pi(\boldsymbol{O})|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}\right\}}\middle|\boldsymbol{O}\right].
\end{aligned}
$$

Robins et al. (1995) proposed a class of estimators based on the efficient score function. Specifically, given a correctly specified model $\pi(\boldsymbol{O}; \boldsymbol{\alpha})$ for $\pi(\boldsymbol{O})$, they considered estimators

$\widehat{\boldsymbol{\theta}}(\phi, \widehat{\boldsymbol{\alpha}})$ solving

$$0 = n^{1/2}\bar{U}(\boldsymbol{\theta}, \phi, \widehat{\boldsymbol{\alpha}}) = n^{-1/2}\sum_{i=1}^{n} U_i(\boldsymbol{\theta}, \phi, \widehat{\boldsymbol{\alpha}}),$$

where $\widehat{\boldsymbol{\alpha}}$ satisfies

$$\sum_{i=1}^{n} S_{\boldsymbol{\alpha},i}(\widehat{\boldsymbol{\alpha}}) = 0,$$

$$S_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \partial\left[R\log\pi(\boldsymbol{O};\boldsymbol{\alpha}) + (1-R)\log\left\{1-\pi(\boldsymbol{O};\boldsymbol{\alpha})\right\}\right]/\partial\boldsymbol{\alpha}.$$

They showed that under certain regularity conditions, the estimator $\widehat{\boldsymbol{\theta}}(\phi, \widehat{\boldsymbol{\alpha}})$ is consistent and asymptoticly normal. In addition, if $\phi = \phi_{\text{op}}$, then $\widehat{\boldsymbol{\theta}}(\phi_{\text{op}}, \widehat{\boldsymbol{\alpha}})$ is asymptoticly efficient. However, because $\phi_{\text{op}}(\boldsymbol{O})$ depends on the unknown joint distribution of the data, $\widehat{\boldsymbol{\theta}}(\phi_{\text{op}}, \widehat{\boldsymbol{\alpha}})$ cannot be used as a "estimator". When both the outcome and the inexpensive covariates are discrete, Robins et al. (1995) proposed an adaptive semiparametric efficient estimator $\widehat{\boldsymbol{\theta}}(\widehat{\phi}_{\text{op}}, \widehat{\boldsymbol{\alpha}})$ by replacing the unknown $\phi_{\text{op}}(\boldsymbol{O})$ with a consistent estimator $\widehat{\phi}_{\text{op}}(\boldsymbol{O})$. When the outcome of interest is continuous, however, this estimator is difficult to implement because it involves numerical solution of an infinite dimensional integral equation.

2.3 Methods for Analyzing Multivariate Outcome-Dependent Sampling Studies

Weighted Estimator

Similar as in the single-outcome case, if every study subject have a positive probability of being selected during the second phase, then the Horvitz-Thompson approach can be adopted. This estimator avoids the joint modeling of the traits and thus can handle quantitative, binary, and censored time-to-event traits simultaneously. It yields unbiased effect estimation and correct type I error. Such weighting methods, however, are substantially less efficient than standard linear regression ignoring the sample design (T. Lumley, personal

communication, April 19, 2012). Efficiency is a major concern in genetic association studies since many genetic effects are small and the correction for multiple comparisons is extremely severe for tens of thousands of variants. In addition, the Horvitz-Thompson approach is not applicable to the design where not every subject has positive probability of being selected during the second phase.

Univariate Analysis Plus Meta-Analysis

Analysis methods for two-phase designs with a single outcome, such as that of Lin et al. (2013), may be applied to the multivariate outcome-dependent sampling design. As mentioned in Section 2.2.2, Lin et al. (2013) proposed a likelihood-based approach for the univariate outcome-dependent sampling design. They derived efficient estimators for both the primary trait, which is used for sampling, and the secondary trait, which is not used for sampling. Suppose that we wish to make inference on the first trait under a multivariate TDS design with K traits. We can analyze the first trait as the primary trait by treating the individuals with extreme values of the first trait as sequenced individuals and all others as nonsequenced individuals. We can also analyze the first trait as a secondary trait with each of the remaining $(K-1)$ traits as the primary trait. We can then combine the summary statistics of the K analyses. This meta-analysis is not valid because it does not account for the correlations of the K statistics caused by overlapping individuals. To avoid overlaps of sequenced individuals, we let each individual be considered "sequenced" in only one analysis. This strategy, however, will introduce bias into the univariate analysis because the "selection" for one trait depends on other traits.

2.4 Design Efficiency of Two-Phase Studies

When the outcome is continuous and there is no inexpensive covariate, Lin et al. (2013) showed that the design is more efficient if it selects subjects with more extreme values of Y.

Specifically, suppose that the regression model is $Y = \alpha + \beta X + \epsilon$, where $\epsilon \sim \mathrm{N}(0, \sigma^2)$. It can be shown that the conditional MLE is consistent and asymptotically equivalent to the full MLE. The information for the conditional MLE is approximately the conditional variance of the score function

$$\frac{X(Y - \alpha - \beta G)}{\sigma^2} - \mathrm{E}\left\{ \frac{X(Y - \alpha - \beta G)}{\sigma^2} \middle| Y \in \mathcal{C} \right\} \tag{2.17}$$

given $Y \in \mathcal{C}$, where $\mathcal{C}$ is the sampling set. After tedious calculation, this information can be expressed as

$$\mathrm{Var}(Y|Y \in \mathcal{C})\mathrm{Var}(G|Y \in \mathcal{C})/\sigma^4 + O(\beta). \tag{2.18}$$

This implies that the design is more efficient if it selects subjects with more extreme values of Y.

CHAPTER 3: ANALYSIS OF SEQUENCE DATA UNDER MULTIVARIATE TRAIT-DEPENDENT SAMPLING

3.1 Introduction

The past few years have seen progressive advances in high-throughput sequencing technologies that allow the sequencing of genomic regions for association studies. However, the cost of performing high-throughput sequencing on a large number of individuals is still high and will likely remain so in the near future. If a quantitative trait is of primary interest, then a cost-effective strategy is to sequence individuals with the extreme trait values. This trait-dependent sampling (TDS) strategy can substantially increase statistical power when compared to a random sample of the same size (Allison 1997, Page and Amos 1999, Slatkin 1999, Chen et al. 2005, Huang and Lin 2007, Lin et al. 2013).

Many sequencing studies are derived from large, population-based cohorts, such as the Atherosclerosis Risk in Communities (ARIC) study (The ARIC Investigators 1989), Cardiovascular Health Study (CHS) (Fried et al. 1991), and Framingham Heart Study (FHS) (Dawber et al. 1951). In these cohorts, hundreds of traits are measured at baseline and follow-up visits. Investigators are often interested in multiple (potentially correlated) quantitative traits. One may select an equal number of individuals from the upper and lower tails of each trait distribution or select individuals from one tail of each trait distribution and use a random sample as a common comparison group. The former design was adopted by the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) Lin et al. (2013). The latter design was recently used in the Cohorts for Heart and Aging Research in Genomic Epidemiology Targeted Sequencing Study (CHARGE-TSS) (Lin et al.

2014).

The NHLBI ESP European American (EA) sample consists of 2538 individuals who were selected for sequencing from six cohorts: ARIC, CHS, FHS, Coronary Artery Risk Development in Young Adults (CARDIA) study (Friedman et al. 1988), Multi-Ethnic Study of Atherosclerosis (MESA) (Bild et al. 2002), and Women's Health Initiative (WHI) (The Women's Health Initiative Study Group 1998). The project contains several studies, each of which was focused on a particular trait and some of which selected individuals with extreme values of quantitative traits, including low-density lipoprotein (LDL) and blood pressure (BP). The CHARGE-TSS involves three cohorts, ARIC, CHS and FHS, in which $\sim$200 individuals with extreme values from each of 14 traits, as well as a random sample of $\sim$2000 individuals, were selected for sequencing at a total of 77 genomic loci that had been identified by genome-wide association studies (GWAS) to be associated with one or more traits (Lin et al. 2013).

Standard linear regression analysis based on least squares (LS) estimation only uses the sequenced individuals and treats them as if they were randomly selected from the whole cohorts. Thus, the multivariate TDS design is ignored with this approach. If the genetic variant of interest is independent of all the traits used in the sampling, then the LS method has correct type I error. If the genetic variant affects certain traits used in the sampling, however, then the LS method yields biased estimates of the genetic effects. The type I error for testing the genetic effect on one trait may also be inflated if other traits that are used in sampling are affected by the genetic variant.

Analysis methods for the univariate TDS design, such as that of Lin et al. (2013), may be applied to the multivariate TDS design. Lin et al. (2013) analyzed the LDL data in the NHLBI ESP by performing separate analysis in each study and combining the summary statistics. This approach may not preserve the type I error because it cannot properly handle sequenced individuals with extreme values in multiple traits, as elaborated in Section 3.2.

In the CHARGE-TSS, the selection of individuals with the extreme values of the pulmonary function was based on both the forced expiratory volume in the first second (FEV_1) and the ratio of FEV_1 to forced vital capacity (FEV_1/FVC) (Lin et al. 2013). The univariate approach is not applicable to this case because it does not allow the selection of an individual to depend on multiple traits. Another limitation of the univariate approach is that it cannot perform simultaneous inference on multiple traits.

In this chapter, we develop a valid and efficient likelihood-based approach to making inferences about genetic effects under multivariate TDS. In our formulation, the sampling can depend on multiple quantitative traits in any manner. Quantitative traits are related to genetic variants and covariates through a multivariate linear regression model while the distributions of genetic variants and covariates are unspecified. We derive the likelihood that accounts for the TDS and utilizes all available data. The computation is challenging due to the presence of missing trait values with arbitrary patterns, the multivariate nature of the model, and a potentially infinite-dimensional covariate distribution. We develop a novel expectation-maximization (EM) algorithm Dempster et al. (1977) to maximize the likelihood. We establish the consistency, asymptotic normality, and asymptotic efficiency of the resulting estimators by using novel arguments to deal with the challenging issue of partially missing trait values. We construct single-variant and gene-level association tests (Li and Leal 2008, Madsen and Browning 2009, Price et al. 2010, Lin and Tang 2011, Wu et al. 2011) for assessing the marginal genetic effects on each trait or the joint effects on any subset of traits. We demonstrate the superiority of the proposed methods over the univariate approach and standard linear regression through extensive simulation studies. Finally, we provide applications to the CHARGE-TSS and NHLBI ESP data.

3.2 Methods

Let $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_K)^{\mathrm{T}}$ be a $K \times 1$ vector of quantitative traits, $\boldsymbol{G}$ be a $d \times 1$ vector of genetic variables, and $\boldsymbol{Z}$ be a $p \times 1$ vector of covariates (including the unit component). We relate $\boldsymbol{Y}$ to $\boldsymbol{G}$ and $\boldsymbol{Z}$ through the multivariate linear model:

$$\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{G} + \boldsymbol{\gamma}\boldsymbol{Z} + \boldsymbol{\epsilon}, \tag{3.19}$$

where $\boldsymbol{\beta}$ is a $K \times d$ matrix of regression parameters for the genetic effects, $\boldsymbol{\gamma}$ is a $K \times p$ matrix of regression parameters for the covariate effects, and $\boldsymbol{\epsilon}$ is a K-variate normal random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. In single-variant analysis, $d = 1$, and G is a scalar that codes the number of minor alleles the individual carries at the variant site under the additive model or indicates whether the individual carries any minor allele (or two minor alleles) at that site under the dominant (or recessive) model. In gene-level analysis for rare variants, $\boldsymbol{G}$ is a (weighted) sum of the numbers of mutations across multiple variant sites within a gene or the vector of genotypes for individual variants.

Under the multivariate TDS design, $\boldsymbol{Y}$ is measured on all the N individuals in the cohort (with potential missing values), and $\boldsymbol{G}$ is only collected for a sub-sample of size n. The selection may depend on observed $\boldsymbol{Y}$ in an arbitrary manner. Under the "one-tail" design used in the CHARGE-TSS, the sequenced individuals include those with extreme values of each quantitative trait of interest plus a random sample. Under the "two-tail" design used in the NHLBI ESP, the sequenced individuals have the largest or smallest trait values. If $\boldsymbol{Z}$ contains demographic/environmental variables and ancestry information, such as the percentage of African ancestry or the principal components (PCs) for ancestry, which is estimated from the GWAS marker data, then $\boldsymbol{Z}$ may potentially be available for all N individuals. If the ancestry information is obtained from the sequence data, then $\boldsymbol{Z}$ is available only for the n sequenced individuals. Because it is often difficult to retrieve covariate information for

nonsequenced individuals, especially when multiple cohorts are involved, we require $\boldsymbol{Z}$ to be available only for the n sequenced individuals.

We arrange the records such that the first n individuals are the sequenced ones and the remaining $(N-n)$ are the nonsequenced ones. Then the data consist of $(\boldsymbol{Y}_i^{obs}, \boldsymbol{Z}_i, \boldsymbol{G}_i)$ for $i = 1, \ldots, n$ and $\boldsymbol{Y}_i^{obs}$ for $i = n+1, \ldots, N$, where $\boldsymbol{Y}_i^{obs}$ is the observed part of $\boldsymbol{Y}_i$. We include all the individuals with at least one nonmissing trait — the largest possible sample — in the analysis. We assume that the observations on $\boldsymbol{Y}$ are missing at random. We require $\boldsymbol{Z}$ to be completely observed for all sequenced individuals, which is the case in both the CHARGE-TSS and NHLBI ESP.

We represent $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\Sigma}$ by $\boldsymbol{\theta}$. We show in Section 3.6 that the observed-data likelihood takes the form

$$\prod_{i=1}^{n} \left[f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs} | \boldsymbol{Z}_i, \boldsymbol{G}_i) f(\boldsymbol{Z}_i, \boldsymbol{G}_i) \right] \prod_{i=n+1}^{N} \int_{\boldsymbol{z},\boldsymbol{g}} f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs} | \boldsymbol{z}, \boldsymbol{g}) dF(\boldsymbol{z}, \boldsymbol{g}), \tag{3.20}$$

where $f_{\boldsymbol{\theta}}(\cdot|\boldsymbol{z}, \boldsymbol{g})$ is the joint density of $\boldsymbol{Y}^{obs}$ conditional on $(\boldsymbol{Z}, \boldsymbol{G}) = (\boldsymbol{z}, \boldsymbol{g})$, $f(\cdot, \cdot)$ is the joint density of $(\boldsymbol{Z}, \boldsymbol{G})$, and $F(\cdot, \cdot)$ is the distribution function of $f(\cdot, \cdot)$. Note that we do not assume a specific form for $f(\cdot, \cdot)$ in (3.20). Thus, $f(\cdot, \cdot)$ is infinite-dimensional when $\boldsymbol{Z}$ contains continuous covariates. We estimate $f(\cdot, \cdot)$ by the discrete probabilities at the observed distinct values of $(\boldsymbol{Z}_i, \boldsymbol{G}_i)$, $i = 1, \ldots, n$, denoted by $(\boldsymbol{z}_1, \boldsymbol{g}_1), \ldots, (\boldsymbol{z}_m, \boldsymbol{g}_m)$, $m \leq n$, and maximize the above function over other parameters. Denote the point mass at $(\boldsymbol{z}_j, \boldsymbol{g}_j)$ as q_j, $j = 1, \ldots, m$. The objective function to be maximized is equivalent to

$$\begin{aligned} &\sum_{i=1}^{n} \left[\log f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs} | \boldsymbol{Z}_i, \boldsymbol{G}_i) + \log \sum_{j=1}^{m} I\left\{ (\boldsymbol{Z}_i, \boldsymbol{G}_i) = (\boldsymbol{z}_j, \boldsymbol{g}_j) \right\} q_j \right] \\ &+ \sum_{i=n+1}^{N} \log \sum_{j=1}^{m} f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs} | \boldsymbol{z}_j, \boldsymbol{g}_j) q_j, \end{aligned} \tag{3.21}$$

where $I(\cdot)$ is the indicator function.

We present in Section 3.7.2 a novel EM algorithm for maximizing (3.21) that is computationally efficient and numerically stable. In addition, we prove in Section 3.7.3 that the resulting maximum likelihood estimators (MLEs) are consistent, asymptotically normal, and asymptotically efficient. Thus, the corresponding association tests have correct type I error and are the most powerful of all valid tests.

Inferences about the genetic effects on the traits of interest are flexible under our likelihood framework, as detailed in Section 3.7.4. For single-variant analysis, G is a scalar, and $\boldsymbol{\beta}$ reduces to a $K \times 1$ vector. We can use the Wald, score, or likelihood ratio statistics to test any subset of $\boldsymbol{\beta}$. The Wald tests are the most efficient computationally because we only need to fit the model once no matter how many and what kind of hypotheses we are interested in; to perform the score or likelihood ratio tests, we need to obtain the restricted MLEs under each null hypothesis. For variants with moderate minor allele frequencies (MAFs), the three types of tests give similar results.

To perform a burden test for rare variants, we define G as the total number of mutations among variants whose MAFs are below a pre-specified threshold, such as 1% or 5%, with the corresponding tests denoted by T1 and T5, respectively; alternatively, we define G as a weighted sum of the mutation counts, using weights such as those defined by Madsen and Browning (2009) to reflect each variant's MAF, with the corresponding test denoted by MB. For detecting variants with opposite effects on the traits, we extend the sequence kernel association test (SKAT) (Wu et al. 2011) to the multivariate TDS setting. We can test the null hypothesis that there is no genetic effect on a particular trait or the "global" null hypothesis that there is no genetic effect on any trait. All our gene-level tests are based on the score statistics, which are statistically more accurate and numerically more stable than the Wald statistics for rare variants (Lin and Tang 2011).

Lin et al. (2013) proposed a likelihood-based approach for the univariate TDS design. They derived efficient estimators for both the primary trait, which is used for sampling, and

the secondary trait, which is not used for sampling. Suppose that we wish to make inference on the first trait under a multivariate TDS design with K traits. We can analyze the first trait as the primary trait by treating the individuals with extreme values of the first trait as sequenced individuals and all others as nonsequenced individuals. We can also analyze the first trait as a secondary trait with each of the remaining $(K-1)$ traits as the primary trait. We can then combine the summary statistics of the K analyses. This meta-analysis is not valid because it does not account for the correlations of the K statistics caused by overlapping individuals. To avoid overlaps of sequenced individuals, we let each individual be considered "sequenced" in only one analysis. This strategy, however, will introduce bias into the univariate analysis because the "selection" for one trait depends on other traits. We label these two methods as (a) and (b), respectively.

For the design that contains a random sample, such as the one-tail design adopted by the CHARGE-TSS, each individual in the cohort has a positive probability of being selected. Then the inverse probability weighting (IPW) method commonly used in survey sampling can be adopted. The IPW method avoids the joint modeling of the traits and thus can handle quantitative, binary, and censored traits simultaneously. It yields unbiased effect estimation and correct type I error. Such weighting methods, however, are substantially less efficient than the LS method (T. Lumley, personal communication, April 19, 2012). Efficiency is a major concern in association studies since many genetic effects are small and the correction for multiple comparisons is extremely severe for tens of thousands of variants. In addition, IPW is not applicable to the design that does not contain a random sample.

3.3 Simulation Studies

We evaluated the performance of the MLE and LS methods in extensive simulation studies. The ARIC data in the CHARGE-TSS are more complex than the NHLBI ESP data because the former contain more sampling traits and more sequenced individuals with

extreme trait values than the latter. Thus, we designed our simulation studies to mimic the ARIC data in the CHARGE-TSS.

We generated 11 traits from the multivariate linear model given in (3.19) in which G is the number of minor alleles for a SNP with MAF of 0.1, Z is a normally distributed confounder (representing a PC for ancestry or some other genetically related variable) with mean G and unit variance, and the error terms are multivariate normal with mean 0, variances 1, and correlations r under compound symmetry. (The Pearson correlation between G and Z is $\sim$0.17.) We generated a cohort of 9000 individuals and selected individuals for sequencing as follows: we first selected a random sample of 1000 individuals; we then selected 100 individuals with the largest values of Y_1 from the remaining 8000 individuals; and we continued to select 100 individuals with the largest values of Y_2 from the remaining 7900 individuals, and so on, until we reached a "sequenced" sample of 2100 individuals. We set $\beta_1 = 0$ and considered two cases of non-zero effects for the other 10 traits: Case 1. five traits with the same effect, i.e., $\beta_2 = \cdots = \beta_6 = 0.2$, $\beta_7 = \cdots = \beta_{11} = 0$; and Case 2. six traits with opposite effects, i.e., $\beta_2 = \beta_4 = \beta_6 = 0.2$, $\beta_3 = \beta_5 = \beta_7 = -0.2$, $\beta_8 = \cdots = \beta_{11} = 0$. The value of 0.2 for β corresponds to R^2 of 0.7% and 4.0% under $\gamma = 0$ and 0.3, respectively; the value of -0.2 corresponds to R^2 of 0.7% and 0.2% under $\gamma = 0$ and 0.3, respectively. We assessed the bias, type I error, and power of the MLE and LS methods. The nominal significance level α was set to 0.001. All results are based on 100,000 replicates.

Table 3.1 shows the results for trait 1 (null effect) and trait 2 (positive effect) in Case 1. The MLE method provides unbiased estimation of genetic effects and correct type I error. The LS method is approximately unbiased for β_1 when the confounder has no effect and the traits are strongly correlated, and it has a negative bias for β_1 when there is confounding or the traits are weakly correlated or independent. When the confounder has no effect, the LS method substantially overestimates β_2. The bias is larger when the correlations are lower. When there is confounding, the bias decreases as the correlations increase. When the traits

are weakly correlated or independent, the LS method yields highly inflated type I error, whether or not the confounder has an effect. The type I error is also inflated when the traits are strongly correlated and the confounder has an effect. The MLE method is more powerful than the LS method because its standardized test statistic tends to be larger. The largest power difference is 0.188 under $\gamma = 0.3$ and $r = 0.5$. The MLE method always yields smaller root mean squared error (RMSE) than the LS method (see Table 3.2).

Table 3.1: Simulation Results for Estimating the Genetic Effects on Trait 1 (Null Effect) and Trait 2 (Positive Effect) in Case 1, Five Traits with the Same Effect

			MLE				LS			
Trait	γ	r	Bias	SE	SEE	Power	Bias	SE	SEE	Power
1	0.0	0.00	0.000	0.048	0.048	0.0010	-0.018	0.059	0.060	0.0014
		0.05	0.000	0.049	0.049	0.0010	-0.015	0.059	0.059	0.0012
		0.10	0.000	0.050	0.049	0.0011	-0.010	0.058	0.059	0.0010
		0.20	0.000	0.050	0.050	0.0010	-0.007	0.058	0.059	0.0010
		0.50	0.001	0.050	0.050	0.0010	0.002	0.058	0.059	0.0008
	0.3	0.00	0.000	0.044	0.044	0.0010	-0.026	0.053	0.053	0.0023
		0.05	0.000	0.044	0.044	0.0008	-0.028	0.053	0.053	0.0026
		0.10	0.000	0.045	0.045	0.0010	-0.032	0.052	0.053	0.0032
		0.20	0.000	0.046	0.046	0.0010	-0.032	0.052	0.053	0.0031
		0.50	0.000	0.048	0.048	0.0009	-0.034	0.051	0.053	0.0030
2	0.0	0.00	0.000	0.048	0.048	0.817	0.033	0.060	0.059	0.732
		0.05	0.000	0.048	0.048	0.805	0.033	0.059	0.059	0.743
		0.10	0.000	0.049	0.049	0.793	0.033	0.059	0.059	0.749
		0.20	0.001	0.049	0.049	0.775	0.031	0.058	0.058	0.753
		0.50	0.001	0.051	0.051	0.744	0.024	0.057	0.058	0.722
	0.3	0.00	0.000	0.044	0.043	0.902	0.018	0.053	0.053	0.799
		0.05	0.000	0.044	0.044	0.888	0.013	0.053	0.052	0.780
		0.10	0.000	0.045	0.045	0.876	0.009	0.052	0.052	0.761
		0.20	0.000	0.046	0.046	0.854	0.002	0.052	0.052	0.723
		0.50	0.000	0.049	0.049	0.799	-0.013	0.051	0.052	0.611

NOTE: SE and SEE stand for standard error and standard error estimate, respectively.

Table 3.3 shows the results for trait 1 (null effect), trait 2 (positive effect), and trait 3 (negative effect) in Case 2. The MLE method continues to provide unbiased estimation of

Table 3.2: Percentage of Bias and RMSE for Estimating the Genetic Effects on Trait 2 (Positive Effect) in Case 1, and Traits 2 (Positive Effect) and 3 (Negative Effect) in Case 2 Under the One-Tail Design

		Case 1: trait 2				Case 2: trait 2				Case 2: trait 3			
γ	r	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
0.0	0.00	0.2%	0.067	16.5%	0.091	0.2%	0.069	26.0%	0.102	0.2%	0.071	22.2%	0.097
	0.05	0.2%	0.068	16.5%	0.090	0.2%	0.069	24.2%	0.099	0.1%	0.072	20.5%	0.095
	0.10	0.2%	0.069	16.3%	0.089	0.2%	0.070	22.4%	0.097	0.1%	0.072	18.8%	0.093
	0.20	0.3%	0.070	15.5%	0.088	0.3%	0.071	19.0%	0.093	0.0%	0.072	15.5%	0.089
	0.50	0.3%	0.072	12.1%	0.085	0.3%	0.071	10.2%	0.086	0.0%	0.071	7.3%	0.085
0.3	0.00	0.1%	0.061	8.9%	0.077	0.1%	0.063	18.5%	0.085	0.1%	0.064	23.3%	0.089
	0.05	0.1%	0.063	6.7%	0.076	0.1%	0.064	15.4%	0.082	0.1%	0.065	23.3%	0.089
	0.10	0.1%	0.063	4.7%	0.075	0.2%	0.064	12.6%	0.080	0.1%	0.065	23.3%	0.089
	0.20	0.2%	0.065	1.0%	0.073	0.2%	0.066	7.4%	0.077	0.1%	0.067	22.8%	0.088
	0.50	0.2%	0.069	6.6%	0.074	0.2%	0.069	4.1%	0.075	0.0%	0.068	19.7%	0.084

NOTE: RMSE stands for root mean squared error.

genetic effects and correct type I error. The LS method tends to overestimate the effect on trait 2 and underestimate the effect on trait 3, and the bias can be as high as 26%, which is higher than in Case 1. The LS method also has inflated type I error (as high as 80%) when there is confounding. When the confounder has no effect, the LS method generally has correct type I error, although it is not as powerful as the MLE method; the power differences are larger when the correlations are higher, which is opposite to what we find in Case 1. The MLE method always yields smaller root mean squared error (RMSE) than the LS method (see Table 3.2). For both Case 1 and Case 2, we conducted other simulations with larger genetic effects and lower MAFs or with 10% random missingness in all traits. The results are similar to those of Tables 3.1 and 3.3 and thus not shown.

Due to the presence of a random sample, it was possible to evaluate the IPW method. We set the weights for individuals with extreme trait values at 1 and set the weights for individuals in the random sample at 9. These weights are not exactly equal to the inverse selection probabilities, which are difficult to calculate under the sequential selection mechanism, but the approximations are good enough for our illustration. The results for Case 1 and Case 2 are summarized in Table 3.4. Comparing Table 3.4 with Tables 3.1 and 3.3,

Table 3.3: Simulation Results for Estimating the Genetic Effects on Trait 1 (Null Effect), Trait 2 (Positive Effect), and Trait 3 (Negative Effect) in Case 2, Six Traits with Opposite Effects

			MLE				LS			
Trait	γ	r	Bias	SE	SEE	Power	Bias	SE	SEE	Power
1	0.0	0.00	0.000	0.050	0.050	0.0011	-0.003	0.061	0.061	0.0010
		0.05	0.000	0.050	0.050	0.0011	-0.003	0.061	0.061	0.0011
		0.10	0.000	0.050	0.050	0.0010	-0.003	0.061	0.061	0.0010
		0.20	0.000	0.051	0.051	0.0010	-0.002	0.060	0.060	0.0009
		0.50	0.000	0.050	0.050	0.0010	-0.001	0.060	0.060	0.0009
	0.3	0.00	0.000	0.045	0.045	0.0011	-0.008	0.054	0.054	0.0012
		0.05	0.000	0.045	0.045	0.0010	-0.011	0.054	0.054	0.0011
		0.10	0.000	0.046	0.046	0.0011	-0.014	0.053	0.054	0.0013
		0.20	0.000	0.047	0.046	0.0010	-0.020	0.054	0.054	0.0018
		0.50	0.000	0.049	0.049	0.0010	-0.025	0.052	0.053	0.0018
2	0.0	0.00	0.000	0.049	0.049	0.792	0.052	0.062	0.061	0.787
		0.05	0.000	0.049	0.049	0.782	0.048	0.062	0.061	0.780
		0.10	0.001	0.050	0.049	0.773	0.045	0.061	0.060	0.772
		0.20	0.001	0.050	0.050	0.762	0.038	0.060	0.060	0.751
		0.50	0.001	0.051	0.050	0.753	0.020	0.059	0.059	0.668
	0.3	0.00	0.000	0.044	0.044	0.888	0.037	0.055	0.054	0.859
		0.05	0.000	0.045	0.045	0.875	0.031	0.054	0.054	0.840
		0.10	0.000	0.046	0.046	0.862	0.025	0.054	0.054	0.818
		0.20	0.000	0.047	0.047	0.842	0.015	0.053	0.053	0.771
		0.50	0.000	0.049	0.049	0.798	-0.008	0.052	0.053	0.622
3	0	0.00	0.000	0.050	0.050	0.754	-0.044	0.060	0.061	0.759
		0.05	0.000	0.051	0.051	0.746	-0.041	0.060	0.061	0.750
		0.10	0.000	0.051	0.051	0.742	-0.038	0.059	0.061	0.741
		0.20	0.000	0.051	0.051	0.734	-0.031	0.059	0.060	0.718
		0.50	0.000	0.050	0.050	0.747	-0.015	0.058	0.059	0.631
	0.3	0.00	0.000	0.045	0.045	0.873	-0.047	0.053	0.054	0.895
		0.05	0.000	0.046	0.046	0.861	-0.047	0.053	0.054	0.900
		0.10	0.000	0.046	0.046	0.850	-0.047	0.053	0.054	0.904
		0.20	0.000	0.047	0.047	0.831	-0.046	0.052	0.054	0.907
		0.50	0.000	0.048	0.048	0.798	-0.039	0.051	0.054	0.888

NOTE: SE and SEE stand for standard error and standard error estimate, respectively.

we observe that although the IPW method preserves the type I error, it is substantially less powerful than the MLE and LS methods.

Table 3.4: Simulation Results for the IPW Method Under the One-Tail Design

		Case 1				Case 2					
		Trait 1		Trait 2		Trait 1		Trait 2		Trait 3	
γ	r	Bias	Power	Bias	Power	Bias	Power	Bias	Power	Bias	Power
0.0	0.00	-0.001	0.0011	0.010	0.366	0.000	0.0011	0.012	0.360	-0.009	0.345
	0.05	0.000	0.0010	0.011	0.370	0.000	0.0010	0.011	0.356	-0.008	0.341
	0.10	0.002	0.0011	0.012	0.373	0.000	0.0011	0.011	0.353	-0.007	0.336
	0.20	0.004	0.0011	0.013	0.376	0.001	0.0011	0.009	0.347	-0.005	0.326
	0.50	0.009	0.0010	0.013	0.369	0.002	0.0011	0.006	0.325	-0.001	0.302
0.3	0.00	-0.003	0.0011	0.008	0.411	-0.001	0.0010	0.011	0.412	-0.010	0.406
	0.05	-0.002	0.0010	0.009	0.411	-0.001	0.0010	0.010	0.406	-0.009	0.401
	0.10	-0.001	0.0011	0.009	0.409	-0.001	0.0011	0.009	0.398	-0.009	0.396
	0.20	0.001	0.0011	0.009	0.404	-0.001	0.0011	0.008	0.385	-0.007	0.387
	0.50	0.003	0.0010	0.008	0.384	-0.001	0.0011	0.003	0.349	-0.004	0.357

We also conducted simulation studies under the two-tail design. Specifically, we generated the cohort in the same manner as in the previous simulation studies but sequentially selected 95 individuals from the upper and lower tails of each trait distribution to reach a "sequenced" sample of 2090 individuals. The results that are analogous to those shown in Tables 3.1 and 3.3 are summarized in Tables 3.5 and 3.6. The MLE method continues to perform well. Because the two-tail sampling is more extreme than the one-tail sampling used in the previous simulation studies, the LS method tends to yield more bias. The loss of power by the LS method compared to the MLE method tends to be more severe under the two-tail design than under the one-tail design (with maximal differences of 0.583 vs. 0.188). In addition, the MLE method is generally more powerful under the two-tail design than under the one-tail design (with the power difference being as high as 0.184).

We conducted additional simulation studies under simple random sampling. We generated the cohort in the same manner as before but selected a simple random sample of 2100 individuals. The LS method is valid in this setting. The power is approximately 0.61 for all traits with non-zero effects (positive or negative) in both Case 1 and Case 2 with any combination of γ and r. When comparing with the power estimates for trait 2 in Tables 3.1 and 3.5 and traits 2 and 3 in Tables 3.3 and 3.6, we see that the two multivariate TDS

Table 3.5: Simulation Results for Estimating the Genetic Effects on Trait 1 (Null Effect) and Trait 2 (Positive Effect) in Case 1 Under the Two-Tail Design

			MLE				LS			
Trait	γ	r	Bias	SE	SEE	Power	Bias	SE	SEE	Power
1	0.0	0.00	0.000	0.046	0.045	0.0010	0.000	0.067	0.067	0.0009
		0.05	0.000	0.046	0.046	0.0011	0.018	0.066	0.066	0.0014
		0.10	0.000	0.046	0.046	0.0011	0.033	0.066	0.067	0.0026
		0.20	0.000	0.045	0.045	0.0010	0.058	0.069	0.070	0.0066
		0.50	0.000	0.042	0.041	0.0010	0.098	0.078	0.078	0.0198
	0.3	0.00	0.000	0.044	0.044	0.0010	-0.027	0.063	0.064	0.0017
		0.05	0.000	0.046	0.046	0.0010	-0.017	0.061	0.062	0.0013
		0.10	0.000	0.046	0.046	0.0011	-0.011	0.061	0.062	0.0011
		0.20	0.000	0.046	0.046	0.0010	0.000	0.062	0.063	0.0008
		0.50	0.000	0.043	0.043	0.0010	0.017	0.069	0.071	0.0010
2	0.0	0.00	0.001	0.046	0.046	0.856	0.085	0.066	0.066	0.844
		0.05	0.001	0.046	0.046	0.856	0.098	0.066	0.066	0.886
		0.10	0.001	0.046	0.046	0.860	0.111	0.066	0.067	0.911
		0.20	0.001	0.046	0.046	0.864	0.125	0.068	0.068	0.928
		0.50	0.000	0.043	0.043	0.909	0.135	0.076	0.076	0.865
	0.3	0.00	0.000	0.045	0.045	0.874	0.047	0.061	0.062	0.754
		0.05	0.000	0.046	0.046	0.864	0.049	0.061	0.062	0.767
		0.10	0.000	0.046	0.046	0.857	0.050	0.061	0.062	0.775
		0.20	0.001	0.046	0.046	0.850	0.053	0.062	0.063	0.776
		0.50	0.000	0.045	0.045	0.875	0.053	0.067	0.068	0.660

NOTE: SE and SEE stand for standard error and standard error estimate, respectively.

Table 3.6: Simulation Results for Estimating the Genetic Effects on Trait 1 (Null Effect), Trait 2 (Positive Effect), and Trait 3 (Negative Effect) in Case 2 Under the Two-Tail Design

			MLE				LS			
Trait	γ	r	Bias	SE	SEE	Power	Bias	SE	SEE	Power
1	0.0	0.00	0.000	0.046	0.046	0.0010	0.000	0.065	0.066	0.0009
		0.05	0.000	0.046	0.046	0.0010	0.000	0.065	0.066	0.0010
		0.10	0.000	0.046	0.046	0.0011	-0.001	0.066	0.066	0.0011
		0.20	0.000	0.044	0.044	0.0010	-0.001	0.070	0.070	0.0008
		0.50	0.000	0.043	0.043	0.0011	-0.006	0.075	0.076	0.0009
	0.3	0.00	0.000	0.045	0.045	0.0010	-0.011	0.064	0.064	0.0010
		0.05	0.000	0.045	0.045	0.0009	-0.014	0.063	0.064	0.0010
		0.10	0.000	0.045	0.045	0.0009	-0.018	0.063	0.064	0.0012
		0.20	0.000	0.045	0.045	0.0011	-0.025	0.065	0.066	0.0016
		0.50	0.000	0.042	0.042	0.0009	-0.038	0.070	0.072	0.0021
2	0.0	0.00	0.001	0.046	0.046	0.860	0.084	0.065	0.066	0.846
		0.05	0.001	0.046	0.046	0.861	0.080	0.066	0.066	0.826
		0.10	0.001	0.046	0.046	0.863	0.075	0.066	0.067	0.800
		0.20	0.001	0.045	0.045	0.874	0.064	0.067	0.068	0.725
		0.50	0.000	0.042	0.042	0.924	0.028	0.075	0.076	0.386
	0.3	0.00	0.000	0.045	0.045	0.869	0.064	0.062	0.063	0.824
		0.05	0.000	0.046	0.046	0.860	0.056	0.062	0.063	0.791
		0.10	0.001	0.046	0.046	0.856	0.048	0.062	0.063	0.752
		0.20	0.001	0.046	0.046	0.853	0.033	0.062	0.063	0.656
		0.50	0.000	0.044	0.044	0.890	-0.007	0.067	0.069	0.307
3	0.0	0.00	-0.001	0.046	0.046	0.861	-0.085	0.066	0.066	0.846
		0.05	-0.001	0.046	0.046	0.863	-0.081	0.066	0.066	0.830
		0.10	-0.001	0.046	0.045	0.868	-0.077	0.066	0.067	0.806
		0.20	-0.001	0.045	0.045	0.878	-0.068	0.068	0.068	0.739
		0.50	0.000	0.042	0.042	0.931	-0.043	0.075	0.076	0.457
	0.3	0.00	0.000	0.045	0.045	0.870	-0.082	0.062	0.063	0.882
		0.05	0.000	0.046	0.046	0.864	-0.083	0.062	0.063	0.884
		0.10	0.000	0.046	0.046	0.859	-0.082	0.062	0.063	0.882
		0.20	0.000	0.046	0.046	0.861	-0.080	0.063	0.064	0.864
		0.50	0.000	0.044	0.044	0.900	-0.068	0.068	0.070	0.713

NOTE: SE and SEE stand for standard error and standard error estimate, respectively.

Table 3.7: Simulation Results for Estimating the Genetic Effects on Trait 1 (Null Effect) and Trait 2 (Positive Effect) in Case 1 When the Traits Follow Multivariate T Distributions

	Trait 1						Trait 2					
	MLE		LS		MLE-INV		MLE		LS		MLE-INV	
ν	Bias	Power	Bias	Power	Bias	Power	Bias	Power	Bias	Power	Bias	Power
5	0.021	0.0023	-0.019	0.0012	0.013	0.0017	-0.052	0.319	0.010	0.158	-0.078	0.358
10	0.015	0.0015	-0.025	0.0015	0.009	0.0012	-0.023	0.648	0.012	0.431	-0.049	0.631
15	0.010	0.0012	-0.026	0.0018	0.007	0.0010	-0.015	0.744	0.013	0.551	-0.038	0.725
20	0.008	0.0011	-0.028	0.0018	0.005	0.0010	-0.011	0.788	0.013	0.612	-0.032	0.771
30	0.005	0.0010	-0.028	0.0023	0.003	0.0009	-0.007	0.826	0.013	0.672	-0.026	0.812

designs are much more efficient than simple random sampling.

To assess the robustness to the normality assumption, we simulated data in the setup of Case 1 under the one-tail design but let $\boldsymbol{\epsilon}$ follow a multivariate t distribution $t_\nu(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the scale matrix, and ν is the degrees of freedom. We set $\gamma = 0.3$ and $r = 0.05$. We added a variation of the MLE method that applies the inverse normal transformation to the trait values, which is referred to as MLE-INV. The results are summarized in Table 3.7. The MLE method has appreciable bias and inflated type I error for trait 1 (null effect) when ν is small but performs reasonably well when ν is moderate or large. The MLE-INV method has better control of the type I error than the MLE method when ν is small. The LS method is biased and its performance worsens as ν increases.

To compare our multivariate approach with the univariate approach of Lin et al. (2013), we simulated a cohort of 10,000 individuals with two traits. We set the genetic variable to be the number of minor alleles for a SNP with MAF of 0.1, the effect sizes at 0.2 and 0 for the two traits; we did not include any confounder in the model. We adopted the two-tail design by sequentially selecting 250 individuals from the upper and lower tails of the two trait distributions. We used score tests for both approaches. We set the nominal significance level at 0.001 and varied the correlation between the two traits and the proportion of random missingness for each trait. As shown in Table 3.8, the univariate approach has inflated type I error, which is caused by the underestimation of the variance in method (a) and the bias

Table 3.8: Simulation Results for Comparing the Multivariate and Univariate Approaches

			Multivariate				Univariate (a)				Univariate (b)			
Trait	r	% Missing	Mean	SE	SEE	Power	Mean	SE	SEE	Power	Mean	SE	SEE	Power
1	0.0	0	114.5	24.0	24.1	0.935	120.0	25.7	25.0	0.936	118.9	25.0	25.5	0.931
		20	103.2	22.8	22.9	0.892	107.8	24.3	23.7	0.898	107.0	23.7	24.1	0.885
	0.1	0	114.5	24.0	24.1	0.934	120.3	25.9	25.1	0.936	119.0	25.1	25.5	0.926
		20	103.3	22.8	22.9	0.895	108.1	24.3	23.7	0.899	107.1	23.7	24.2	0.888
	0.3	0	114.5	23.9	24.1	0.934	122.7	26.6	25.3	0.936	120.1	25.3	25.9	0.925
		20	103.6	22.9	22.9	0.894	109.7	25.0	23.9	0.900	107.7	24.1	24.5	0.879
	0.5	0	114.9	24.1	24.1	0.934	129.6	29.2	26.2	0.937	122.4	26.7	27.1	0.901
		20	104.6	23.2	23.0	0.895	114.5	26.8	24.5	0.904	109.3	25.0	25.3	0.861
	0.7	0	117.0	24.4	24.2	0.941	149.6	34.6	28.7	0.949	125.4	30.0	30.6	0.802
		20	108.0	23.5	23.3	0.910	127.7	30.5	26.3	0.922	112.1	27.2	27.5	0.792
2	0.0	0	0.0	24.2	24.3	0.0009	0.0	25.8	24.8	0.0017	0.0	24.2	24.2	0.0009
		20	0.0	23.1	23.0	0.0009	0.0	24.5	23.5	0.0015	0.0	23.0	23.0	0.0009
	0.1	0	0.0	24.3	24.3	0.0010	0.0	26.0	24.8	0.0017	-2.4	24.2	24.2	0.0011
		20	0.0	23.1	23.0	0.0010	0.1	24.6	23.5	0.0016	-1.8	23.0	23.0	0.0011
	0.3	0	0.0	24.5	24.5	0.0010	0.1	27.1	25.3	0.0021	-7.6	24.2	24.2	0.0017
		20	0.0	23.3	23.3	0.0009	0.1	25.5	23.9	0.0017	-5.8	23.0	23.0	0.0014
	0.5	0	0.1	25.1	25.1	0.0010	0.1	29.7	26.6	0.0035	-14.4	24.3	24.4	0.0032
		20	0.0	23.8	23.8	0.0011	0.1	27.3	24.8	0.0027	-10.7	23.1	23.1	0.0022
	0.7	0	0.1	26.1	26.2	0.0010	0.1	34.7	29.3	0.0053	-24.9	25.3	25.5	0.0096
		20	0.1	24.8	24.8	0.0011	0.2	30.9	26.8	0.0039	-17.5	23.8	23.9	0.0047

NOTE: SE and SEE stand for the standard error and standard error estimate of the score statistic.

in method (b). The inflation increases as the correlation between the two traits becomes stronger. There is power loss in (b) as compared to the multivariate approach, which is caused by the larger variances of the test statistics. The power difference is larger when the correlation is higher and is not affected much by the level of missingness.

3.4 CHARGE-TSS ARIC Data

We considered the ARIC data in the CHARGE-TSS. As described, a random sample plus individuals with extreme values for 11 traits were selected from ∼9000 ARIC whites who provided informed consent for use of their genetic data and had sufficient DNA for sequencing. The selected individuals were sequenced for 77 genomic loci that had previously been found to be associated with one or more of 14 traits. (Three traits were not used for sampling in the ARIC data.) After quality control (QC), the genotype data included 31,813

SNPs and 2003 individuals. Details for the design, sample selection criteria, genotype QC, and annotation can be found in Lin et al. (2013).

We removed individuals without PCs (calculated from GWAS data) and obtained 9103 individuals, among whom 1927 were sequenced. Table 3.9 shows the number of individuals with nonmissing trait values in the cohort, the specific sampling strategy, and the achieved number of extreme cases for sequencing, as well as that number after QC for each of the 11 traits. (Note that the numbers of extreme cases for all traits may add up to be greater than n since some individuals may have extreme values for multiple traits.) Of the 11 traits used for sampling, stroke is an age-at-onset trait that cannot be incorporated into our model. We treated the 60 individuals who were selected solely due to stroke as nonsequenced individuals. As noted before, the pulmonary function trait comprised two traits — FEV_1 and FEV_1/FVC — such that the total number of traits entering into the analysis remained at 11. C-reactive protein (CRP) and retinal venule diameter have about 20% missingness in the whole cohort, while all the other traits have less than 5% missingness.

In the CHARGE-TSS, the selections for certain traits were based on the residuals of the original values adjusted for various covariates. For those traits, we used the residuals in the analysis. Most of the traits are positively correlated, and there is no pairwise correlation less than -0.15. The correlations are 0.56 between fast insulin and body mass index (BMI), 0.49 between the two pulmonary function traits, 0.30 between BMI and CRP, and 0.22 between fast insulin and hematocrit, as well as between fast insulin and CRP. All the other positive correlations are well below 0.2, and many of them are essentially 0 (see Table 3.10). We included age, gender, study centers, and the top five PCs as covariates.

We focused on BMI. We restricted the single-variant analysis to SNPs with MAFs larger than 5% and ended up with 2971 SNPs. We chose the additive genetic model. Table 3.11 shows the top 10 SNPs for the MLE method and the corresponding LS results. The LS method consistently yields larger effect estimates for SNPs with positive effects and smaller

Table 3.9: Summary of the ARIC Data in the CHARGE-TSS

Trait	No. (%) of non-missing values	Sampling strategy	No. sequenced (No. after QC)
ECG PR interval	8996 (98.82)	high residual	94 (92)
ECG QRS interval	9053 (99.45)	high residual	90 (89)
Blood pressure	9091 (99.87)	high/low residual	93 (89)
Body mass index	9095 (99.91)	high	90 (79)
Fasting insulin	8896 (97.73)	high	94 (94)
C-reactive protein	7211 (79.22)	high residual	93 (90)
Hematocrit	9071 (99.65)	low residual	97 (85)
Retinal venule diameter	7099 (77.99)	high	156 (154)
Carotid wall thickness	8725 (95.85)	high	91 (87)
Pulmonary: FEV_1	8958 (98.41)	low	186 (185)
Pulmonary: FEV_1/FVC	8956 (98.39)		
Stroke		early onset	74 (70)
Random sample			946 (913)
Total	9103 (100.00)		2003 (1927)

NOTE: For the sampling strategy, "high" ("low") means sampling from the upper (lower) tail of the trait distribution; "residual" indicates that the sampling is based on the residuals of the original values adjusted for covariates.

effect estimates for SNPs with negative effects. This is similar to what we find in most scenarios under Case 2 in the simulation studies. As shown in Figure 3.1 of the Supplemental Material, the p-values for the MLE and LS methods are similar.

In gene-level analysis of rare variants, we considered "functional coding" variants, i.e., non-synonymous, splicing, and stop-gain variants, and ended up with a total of 2360 variants. We removed any targeted region with minor allele count (MAC) — the number of individuals with at least one mutation — less than five. For MB and SKAT tests, we only included variants with MAFs less than 5%. Table 3.12 shows the results for the top five targeted regions in each of the four types of tests based on the MLE method. We also performed gene-level tests of the global null hypothesis that there is no genetic effect on any trait. Table 3.13 shows the results for the top five targeted regions in each of the four types of tests. It would be worthwhile to follow up the regions identified in Tables 3.12 and 3.13 in larger

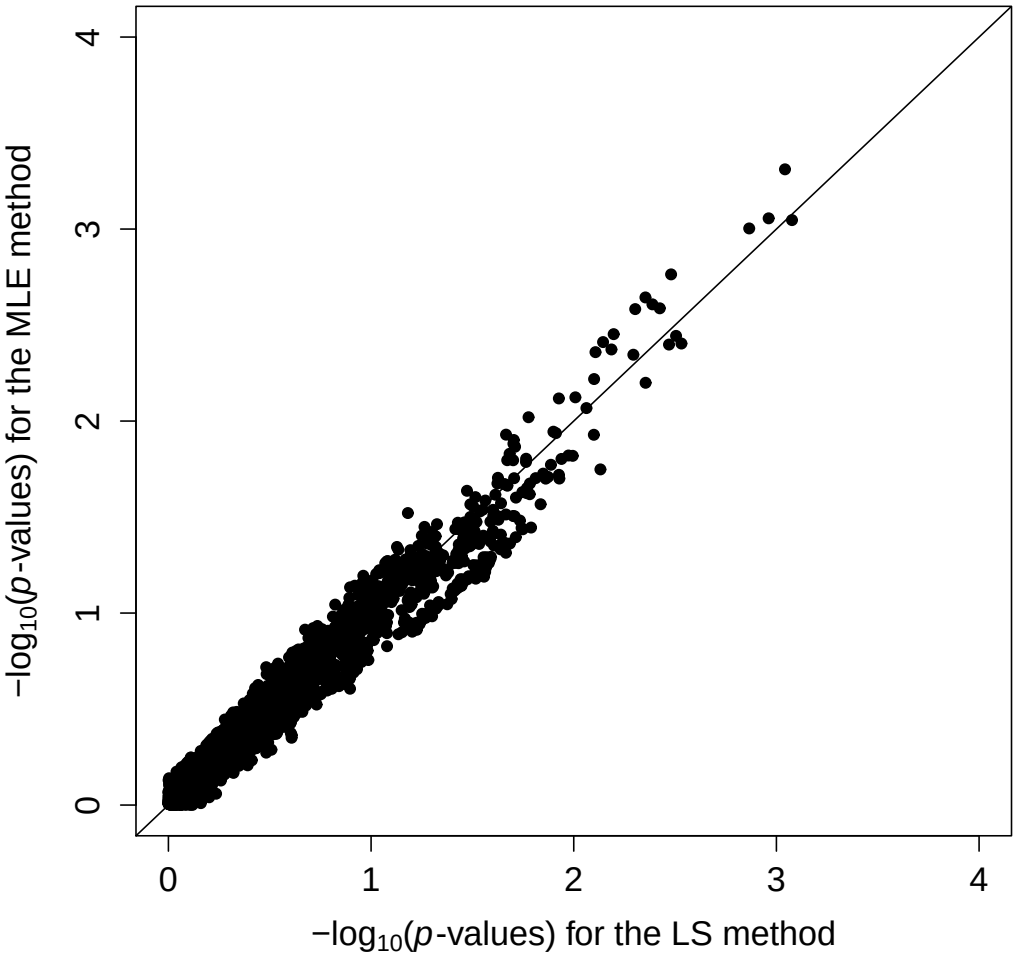

Figure 3.1: Plot of the p-values for the MLE versus LS methods in the analysis of the BMI data in the CHARGE-TSS ARIC sample. SNPs with MAFs greater than 5% are included.

Table 3.10: Pairwise Correlations of the 11 Traits Used for Sampling in the CHARGE-TSS ARIC Data

	PR	QRS	BP	BMI	FI	CRP	HEMA	EYE	IMT	FEV_1
QRS	0.04									
BP	0.02	0.04								
BMI	0.00	0.00	0.00							
FI	0.00	−0.04	0.12	0.56						
CRP	−0.01	−0.03	0.04	0.30	0.22					
HEMA	−0.02	−0.02	0.11	0.13	0.22	0.06				
EYE	0.00	−0.04	−0.05	0.03	0.04	0.08	0.18			
IMT	0.02	0.01	0.08	0.08	0.07	0.08	0.05	0.07		
FEV_1	0.01	0.03	−0.07	−0.04	−0.11	−0.14	−0.05	−0.05	−0.06	
FEV_1/FVC	0.02	0.04	−0.02	0.17	0.15	0.05	0.01	−0.02	−0.01	0.49

NOTE: PR: ECG PR interval; QRS: ECG QRS interval; BP: blood pressure; BMI: body mass index; FI: fast insulin; CRP: C-reactive protein; HEMA: hematocrit; EYE: retinal venule diameter; IMT: carotid wall thickness; FEV_1: forced expiratory volume in 1 second; FVC: forced vital capacity.

samples.

3.5 NHLBI ESP EA Data

The NHLBI ESP EA data consist of the six cohorts mentioned previously and include four types of study designs. The first study is a TDS study consisting of 872 individuals who were selected from the upper and lower tails of the LDL and BP distributions. The second study is a random sample of 721 individuals with measurements on a common set of phenotypes; this study is referred to as the deeply phenotyped reference (DPR). The third study is a case-control study of early myocardial infarction (MI) consisting of 220 cases and 390 controls. The fourth study is a case-only study of stroke consisting of 335 individuals with ischemic stroke. Exome sequencing was performed on the selected individuals at the University of Washington and the Broad Institute. We implemented the genotype QC steps described by Lin et al. (2013) and obtained 1,281,645 variants.

In the TDS study, we excluded individuals (either sequenced or nonsequenced) who were

Table 3.11: Top 10 SNPs in the Single-Variant Analysis of the BMI Data in the CHARGE-TSS ARIC Sample

			MLE			LS		
Variant ID	Gene	MAF	Est	SE	p-value	Est	SE	p-value
chr02:000649384	*TMEM18*	2.87E-01	1.12E-01	3.21E-02	4.89E-04	1.34E-01	4.04E-02	9.07E-04
chr02:000669959	*TMEM18*	2.98E-01	−1.06E-01	3.19E-02	8.79E-04	−1.34E-01	4.11E-02	1.09E-03
chr12:000547464	*NINJ2*	6.43E-02	−1.96E-01	5.92E-02	8.98E-04	−2.47E-01	7.41E-02	8.38E-04
chr01:068340029	*WLS*	4.94E-01	−9.41E-02	2.86E-02	9.93E-04	−1.17E-01	3.65E-02	1.36E-03
chr02:000648937	*TMEM18*	2.95E-01	1.01E-01	3.23E-02	1.72E-03	1.19E-01	4.07E-02	3.31E-03
chr02:000648595	*TMEM18*	3.00E-01	9.75E-02	3.19E-02	2.27E-03	1.15E-01	4.04E-02	4.43E-03
chr02:000645222	*TMEM18*	1.12E-01	−1.44E-01	4.74E-02	2.47E-03	−1.71E-01	5.96E-02	4.09E-03
chr02:000649218	*TMEM18*	2.60E-01	1.01E-01	3.36E-02	2.59E-03	1.23E-01	4.24E-02	3.76E-03
chr02:000647954	*TMEM18*	2.95E-01	9.83E-02	3.27E-02	2.61E-03	1.15E-01	4.10E-02	4.97E-03
chr02:000648157	*TMEM18*	2.99E-01	9.34E-02	3.20E-02	3.53E-03	1.10E-01	4.04E-02	6.35E-03

NOTE: Variant ID is in "chromosome:position" format, where the positions are based on the reference human genome (NCBI Genome Build 36, 2006). Est and SE stand for the genetic effect estimate and standard error, respectively.

not eligible for either the LDL or BP selection. In the FHS, which contains related individuals, we removed one individual from each pair of first- or second-degree relatives. The actual sample selections for LDL and BP were based on the residuals rather than the original values. We used the LDL residuals (log-transformed LDL values adjusted for age, age-squared, gender, and lipid medication) and BP residuals (mean of the residuals for diastolic and systolic BPs adjusted for age, gender, BMI, and anti-hypertensive medication) as the trait values in the analysis. We considered LDL as the trait of interest and removed individuals with missing LDL values in the DPR, MI, and stroke studies. Note that individuals with missing LDL or BP values (but not both) were still included in the analysis of the TDS study. Table 3.14 summarizes the sample sizes of the four studies in each cohort after QC.

In the TDS study, we used both the MLE and LS methods to analyze LDL. For case-control and case-only studies with rare diseases, standard linear regression analysis of secondary quantitative traits conditional on the disease status yields approximately correct results (Lin and Zeng 2009). Because early MI and ischemic stroke are relatively rare, we performed standard linear regression in the MI (adjusted for the MI status), stroke, and DPR studies. We included cohorts and sequencing centers/targets as covariates. We performed

Table 3.12: Top Five Targeted Regions for the T1, T5, MB, and SKAT Tests in the Analysis of the BMI Data Using the MLE Method in the CHARGE-TSS ARIC Sample

Test	Region	MAC	MLE p-value	LS p-value
T1	chr05:087819438-088215292	6	1.90E-03	2.21E-01
	chr01:168853417-168975265	6	3.52E-03	4.59E-01
	chr12:111338491-111436622	8	1.73E-02	5.37E-01
	chr05:156830995-156936446	60	1.84E-02	2.16E-01
	chr07:100054874-100079499	48	2.59E-02	2.88E-02
T5	chr05:087819438-088215292	6	1.90E-03	2.21E-01
	chr01:168853417-168975265	6	3.52E-03	4.59E-01
	chr12:111338491-111436622	8	1.73E-02	5.37E-01
	chr07:100054874-100079499	48	2.59E-02	2.88E-02
	chr10:104579177-104619322	23	2.90E-02	9.45E-01
MB	chr13:109599195-109758700	18	2.32E-02	2.52E-01
	chr06:025857845-025987550	57	2.78E-02	4.67E-01
	chr10:104579177-104619322	6	4.64E-02	1.85E-01
	chr11:046720500-046832766	6	7.68E-02	8.90E-01
	chr12:110374301-110521963	46	8.63E-02	7.69E-02
SKAT	chr05:156830995-156936446	71	3.18E-03	4.57E-01
	chr06:025857845-025987550	57	9.32E-03	3.41E-01
	chr06:135322113-135417715	58	1.34E-02	9.81E-03
	chr13:109599195-109758700	18	2.31E-02	2.52E-01
	chr10:104579177-104619322	6	4.65E-02	1.85E-01

NOTE: Region is in "chromosome:start-stop" format, where the positions are based on the reference human genome (NCBI Genome Build 36, 2006).

Table 3.13: Top Five Targeted Regions for the T1, T5, MB, and SKAT Tests of the Global Null Hypothesis in the CHARGE-TSS ARIC Sample

Test	Region	MAC	p-value
T1	chr05:156830995-156936446	60	1.03E-03
	chr11:046720500-046832766	43	2.74E-02
	chr12:101312706-101455233	7	4.51E-02
	chr12:111338491-111436622	8	4.70E-02
	chr07:115925580-115935931	5	4.85E-02
T5	chr05:156830995-156936446	104	1.05E-02
	chr11:046720500-046832766	43	2.74E-02
	chr11:016764687-016993639	155	3.30E-02
	chr11:046695000-046720000	53	4.15E-02
	chr12:101312706-101455233	7	4.51E-02
MB	chr05:156830995-156936446	104	2.15E-03
	chr11:016764687-016993639	155	1.54E-02
	chr10:070698661-070832743	41	3.68E-02
	chr07:115925580-115935931	5	4.83E-02
	chr12:111338491-111436622	8	4.85E-02
SKAT	chr06:135322113-135417715	85	3.57E-03
	chr12:111338491-111436622	8	4.21E-03
	chr13:109599195-109758700	102	2.07E-02
	chr07:115925580-115935931	5	2.52E-02
	chr11:046695000-046720000	53	3.03E-02

NOTE: Region is in "chromosome:start-stop" format, where the positions are based on the reference human genome (NCBI Genome Build 36, 2006).

Table 3.14: Sample Size Summary of the NHLBI ESP EA Data

			With nonmissing LDL			
	LDL	BP	DPR	MI	Stroke	Nonsequenced
ARIC	172	93	84	136	6	9553
CARDIA	14	66	32	0	0	1530
CHS	15	3	77	43	1	1186
FHS	12	52	34	147	15	2245
MESA	60	19	159	0	7	1310
WHI	46	8	286	156	49	5115
Total	319	241	672	482	78	20939

meta-analysis of the four studies using software MASS (Tang and Lin 2013).

We restricted the single-variant analysis to SNPs with MACs ≥ 5 and ended up with 109,607 SNPs. We chose the additive model and used score statistics to ensure numerical accuracy for SNPs with low MACs. Figure 3.2 shows the quantile-quantile plots using the MLE and LS methods in the TDS study only and in all four studies. Although the trends in the quantile-quantile plots of the TDS study appear to be similar between the MLE and LS methods, the MLE method clearly produces more significant results than the LS method in the meta-analysis. Table 3.15 lists the top 10 SNPs for the MLE method in the meta-analysis. For the MLE method, the top SNP (chr19:45397229) in the meta-analysis is also the top SNP in the TDS study, with the p-value in the meta-analysis being much more significant (2.08×10^{-10} vs. 2.64×10^{-7}). For the LS method, although the top SNP remains the same, its p-value in the meta-analysis is less significant than that in the TDS study (1.17×10^{-6} vs. 4.29×10^{-7}).

The forest plots shown in Figure 3.3 help to explain the results in Figure 3.2 and Table 3.15. The MLE estimates in the TDS study are very similar to the estimates in the DPR, MI, and stroke studies. (The estimates in the stroke study tend to have large standard errors due to its small sample size.) Thus, the MLE estimates from the meta-analysis are similar to the MLE estimates in the TDS study but with smaller standard errors. Because of its bias,

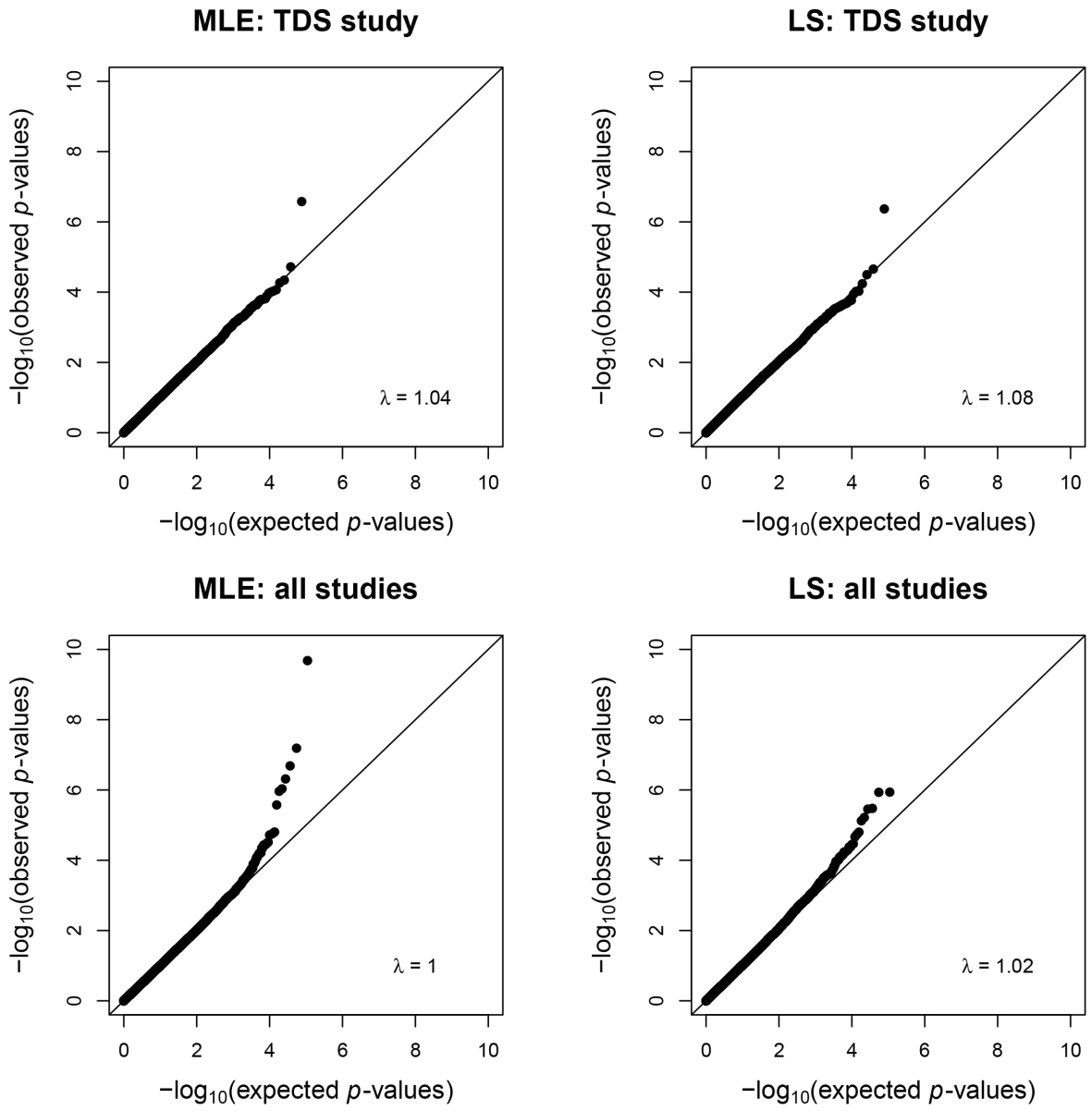

Figure 3.2: Quantile-quantile plots for the single-variant analysis of the LDL data using the MLE and LS methods in the TDS study only and in all four studies included in the NHLBI ESP EA sample. The values of the genomic control λ, defined as the ratio between the observed median of the test statistics and the median of the χ^2_1 distribution, are also shown.

the LS method yields larger effect estimates as well as (proportionately) larger standard errors than the MLE method in the TDS study, such that the two methods have similar standardized test statistics in the TDS study. Because the LS estimates in the TDS study are much larger than the LS estimates in the other three studies, meta-analysis of the LS

Table 3.15: Top 10 SNPs in the Single-Variant Analysis of the LDL Data in the NHLBI ESP EA Sample

			MLE		LS	
Variant ID	Gene	MAC	All studies	TDS study	All studies	TDS study
chr19:045397229	*TOMM40*	132	2.08E-10	2.64E-07	1.17E-06	4.29E-07
chr01:109814880	*CELSR2*	538	6.48E-08	8.57E-05	3.51E-06	9.42E-05
chr12:101685691	*UTP20*	546	2.06E-07	2.45E-04	6.08E-06	2.10E-04
chr12:101685852	*UTP20*	548	4.85E-07	5.25E-04	7.53E-06	4.61E-04
chr12:101693534	*UTP20*	614	9.28E-07	1.62E-03	3.35E-06	1.44E-03
chr12:101776996	*UTP20*	554	1.09E-06	6.76E-04	1.85E-05	6.15E-04
chr19:002039746	*MKNK2*	9	2.66E-06	1.91E-06	1.20E-02	9.17E-06
chr07:121513561	*PTPRZ1*	492	1.57E-05	3.89E-03	5.87E-05	3.84E-03
chr01:186089112	*HMCN1*	916	1.73E-05	1.08E-04	4.28E-03	1.14E-04
chr12:101705477	*UTP20*	560	1.83E-05	3.67E-03	1.05E-04	3.55E-03

NOTE: Variant ID is in "chromosome:position" format, where the positions are based on the human reference sequence (UCSC Genome Browser, hg19).

estimates from the four studies yields less significant results than the MLE meta-analysis.

We also performed single-variant analysis in the TDS study using the univariate approach of Lin et al. (2013). Figure 3.4 compares the p-values for the multivariate and univariate methods. The two methods yield similar results for most SNPs. This is because the correlation between LDL and BP among individuals in the TDS study is only 0.01. Note that the multivariate approach produces a more significant p-value for the top SNP (chr19:45397229) than the univariate approach does (2.64×10^{-7} vs. 1.24×10^{-5}).

In gene-level analysis for rare variants, we considered variants that are nonsynonymous, stop-gain, stop-loss, or splicing mutations. Other steps were the same as in the analysis of the CHARGE-TSS ARIC data. The results are displayed in Figures 3.5–3.8 and in Tables 3.16–3.19. The conclusions regarding the performance of the MLE and LS methods are similar to those of the single-variant analysis. Again, the MLE method yields more significant results than the LS method. We also performed gene-level tests of the global null hypothesis. The results are displayed in Figure 3.9 and in Tables 3.20–3.22. The strongest signals appear in the T1 tests.

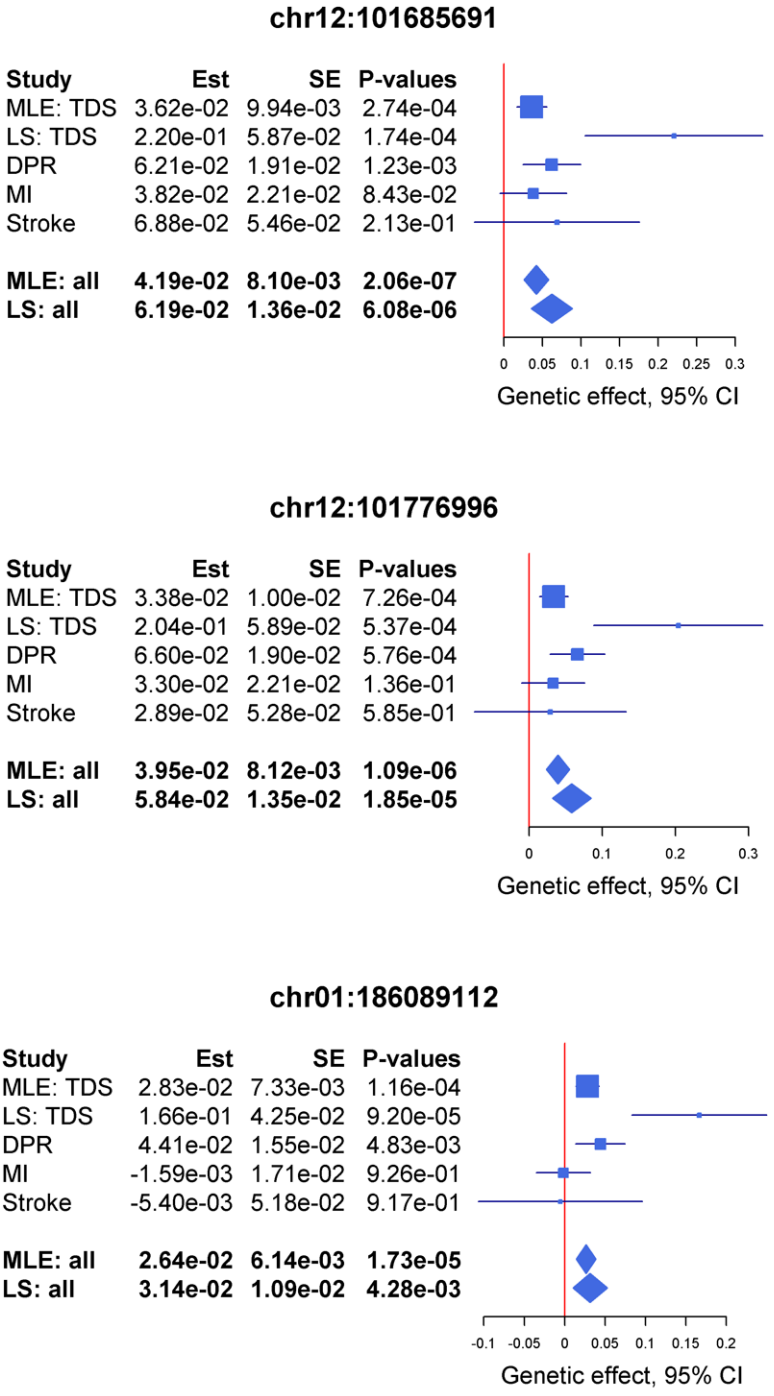

Figure 3.3: Forest plots based on the MLE and LS methods for the third, sixth, and ninth most significant SNPs in the analysis of the LDL data in the NHLBI ESP EA sample. Est, SE, and CI stand for the genetic effect estimate, standard error, and confidence interval, respectively.

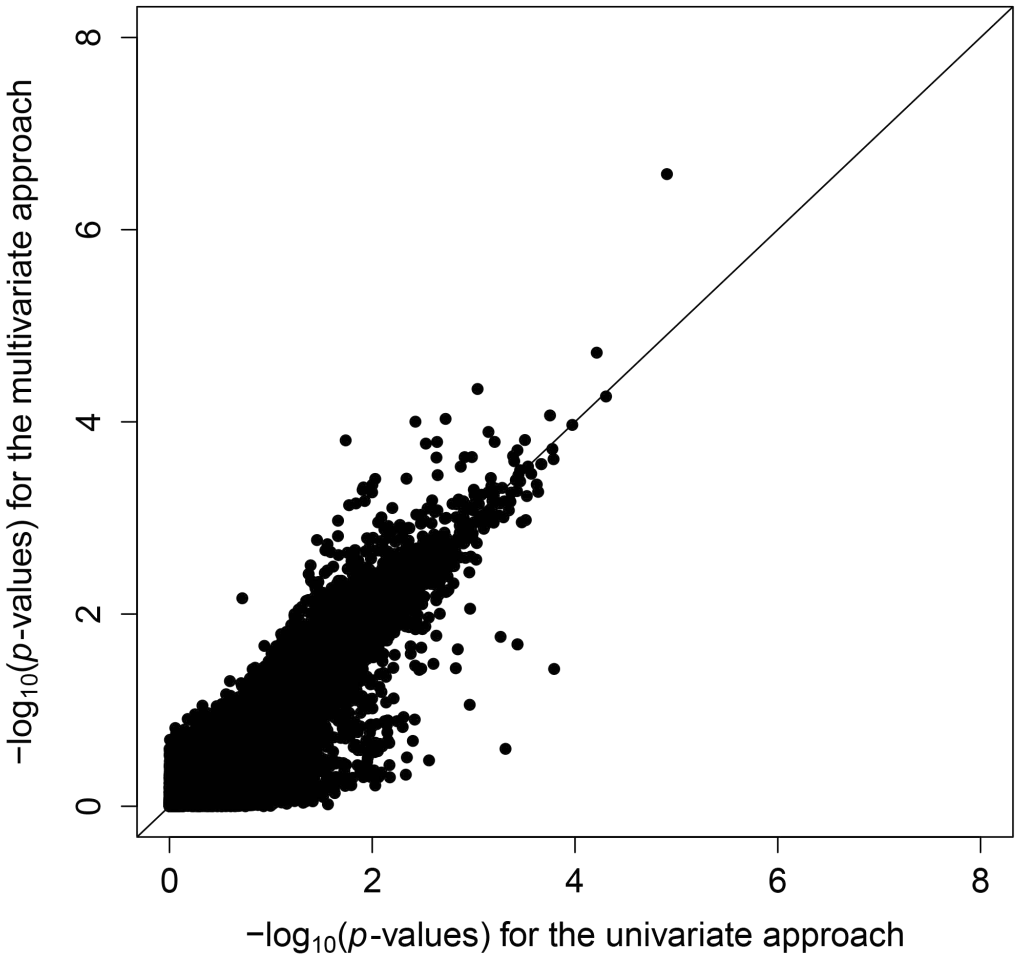

Figure 3.4: Plot of the p-values for the multivariate versus univariate methods in the analysis of the LDL data in the TDS study in the NHLBI ESP EA sample. SNPs with MACs ≥ 5 are included.

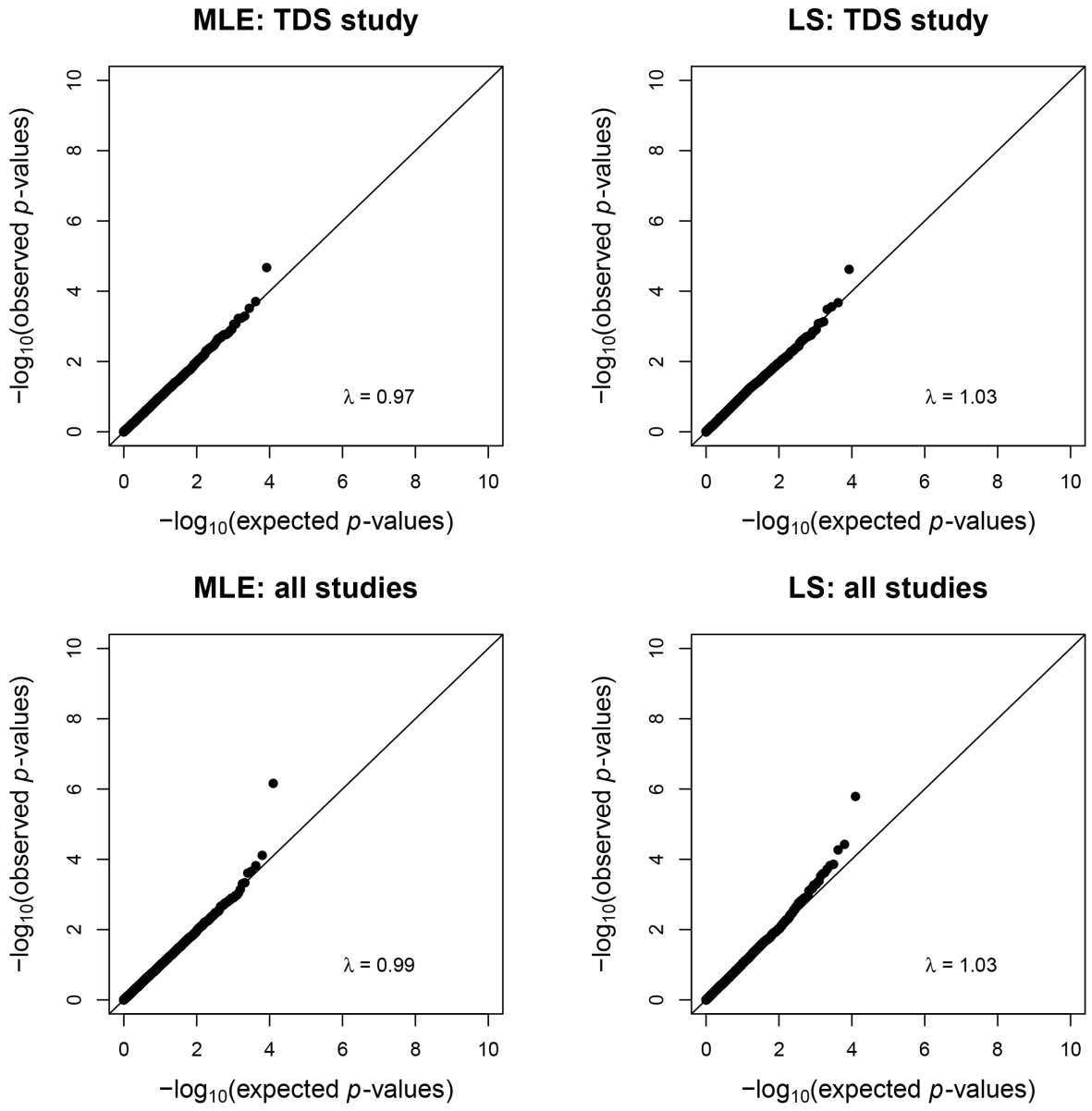

Figure 3.5: Quantile-quantile plots for the T1 tests based on the MLE and LS methods in the analysis of the LDL data in the TDS study only and in all four studies included in the NHLBI ESP EA sample. The values of the genomic control λ are also shown.

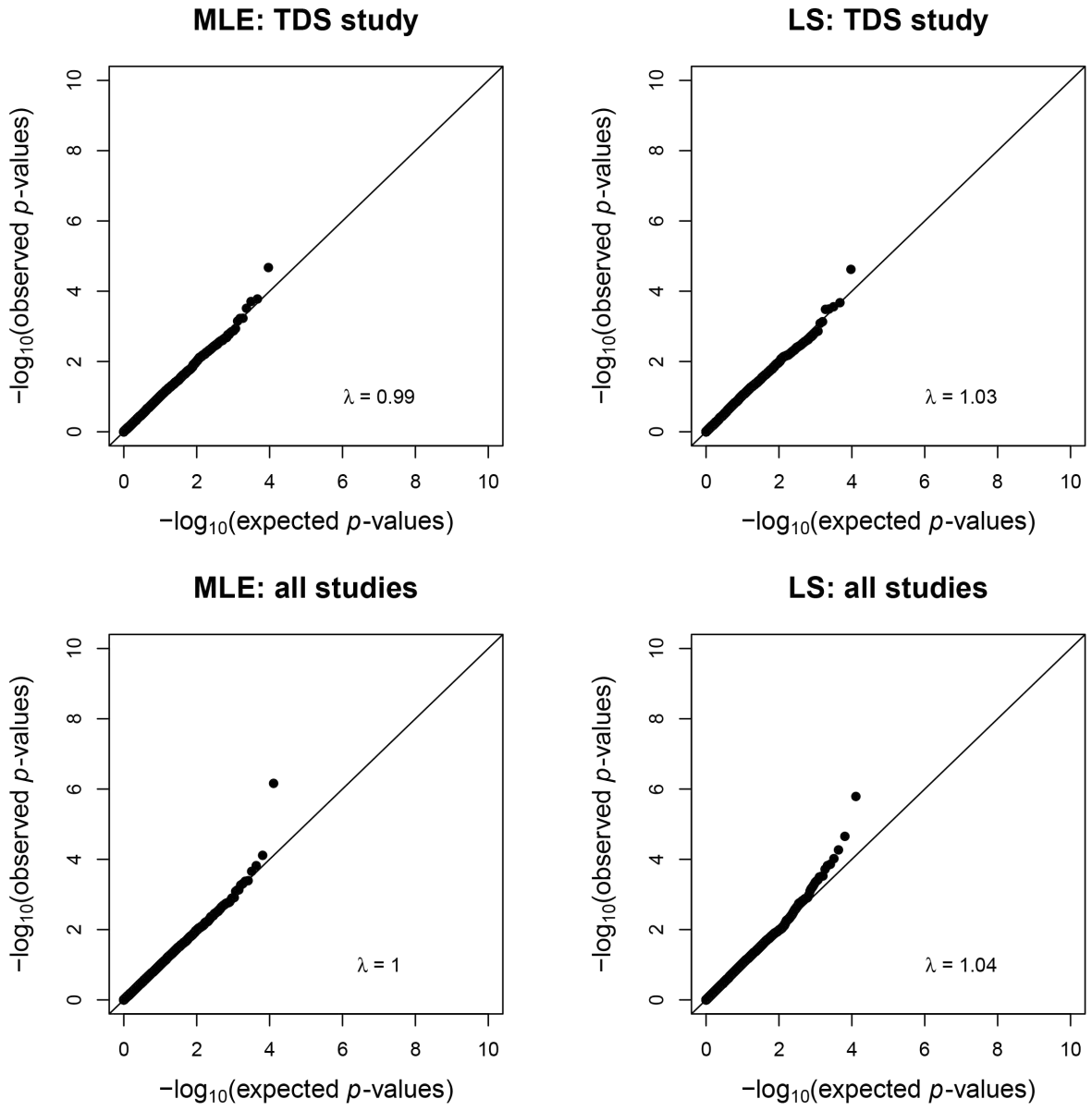

Figure 3.6: Quantile-quantile plots for the T5 tests based on the MLE and LS methods in the analysis of the LDL data in the TDS study only and in all four studies included in the NHLBI ESP EA sample. The values of the genomic control λ are also shown.

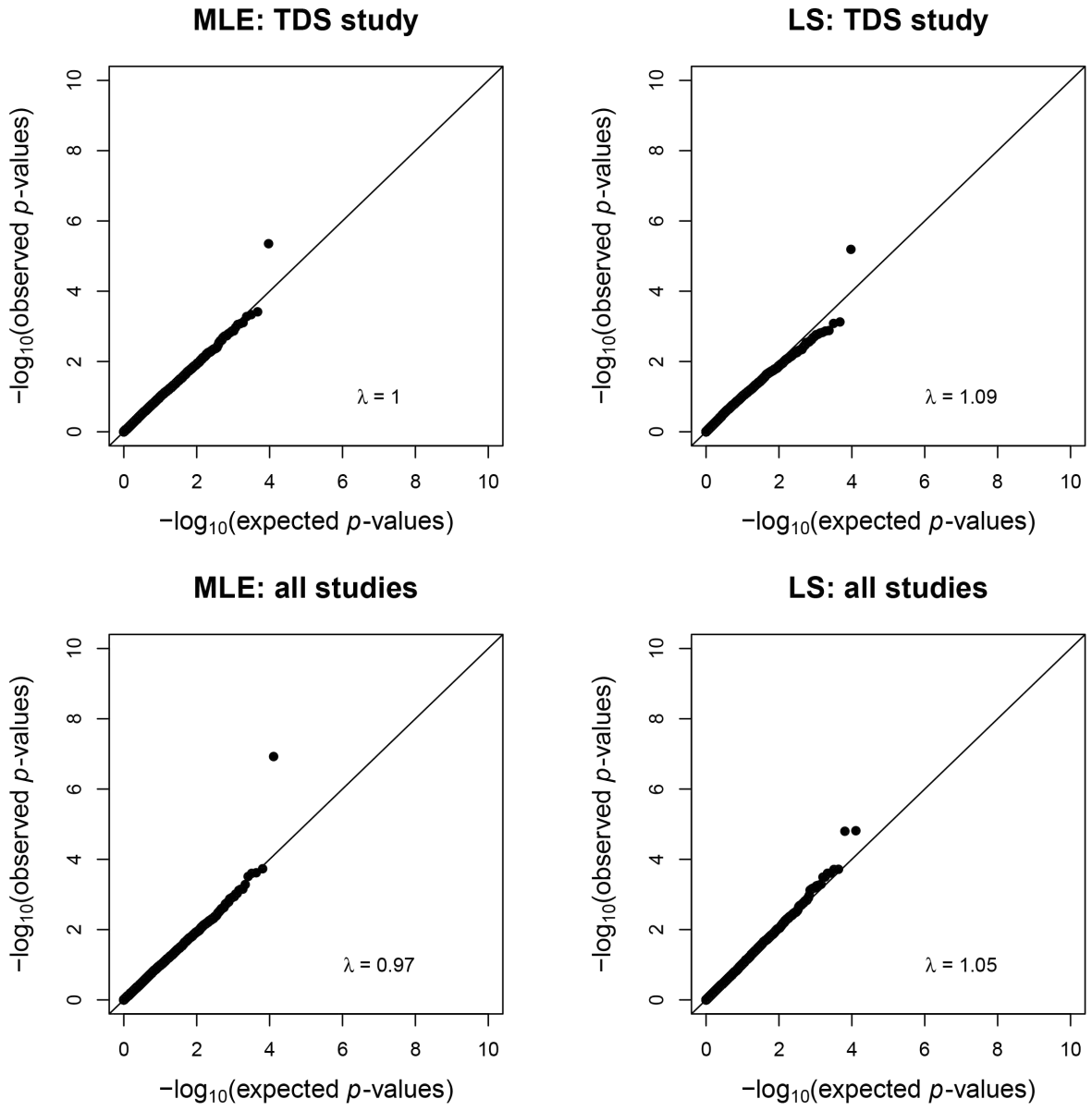

Figure 3.7: Quantile-quantile plots for the MB tests based on the MLE and LS methods in the analysis of the LDL data in the TDS study only and in all four studies included in the NHLBI ESP EA sample. The values of the genomic control λ are also shown.

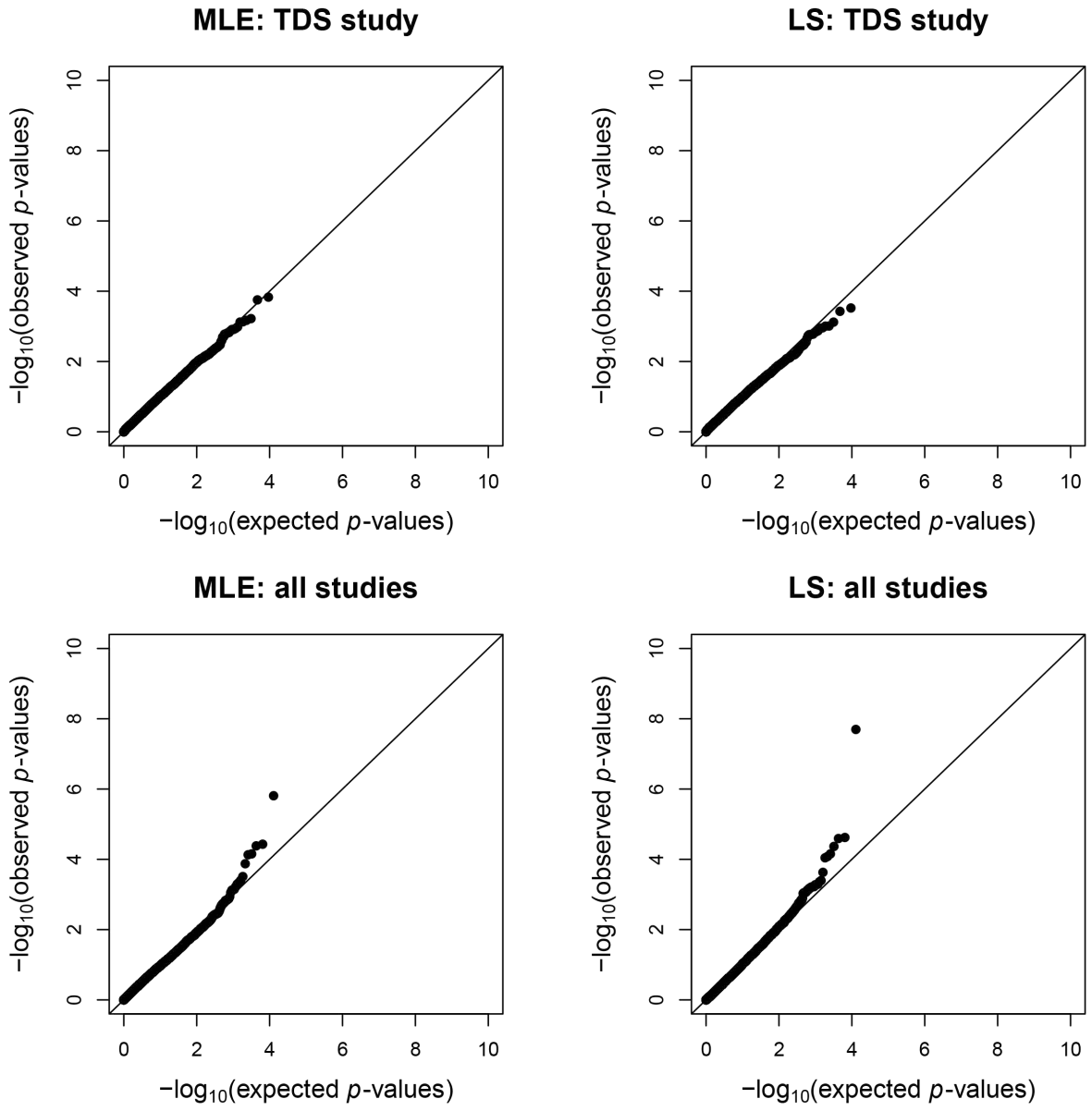

Figure 3.8: Quantile-quantile plots for the SKAT tests based on the MLE and LS methods in the analysis of the LDL data in the TDS study only and in all four studies included in the NHLBI ESP EA sample.

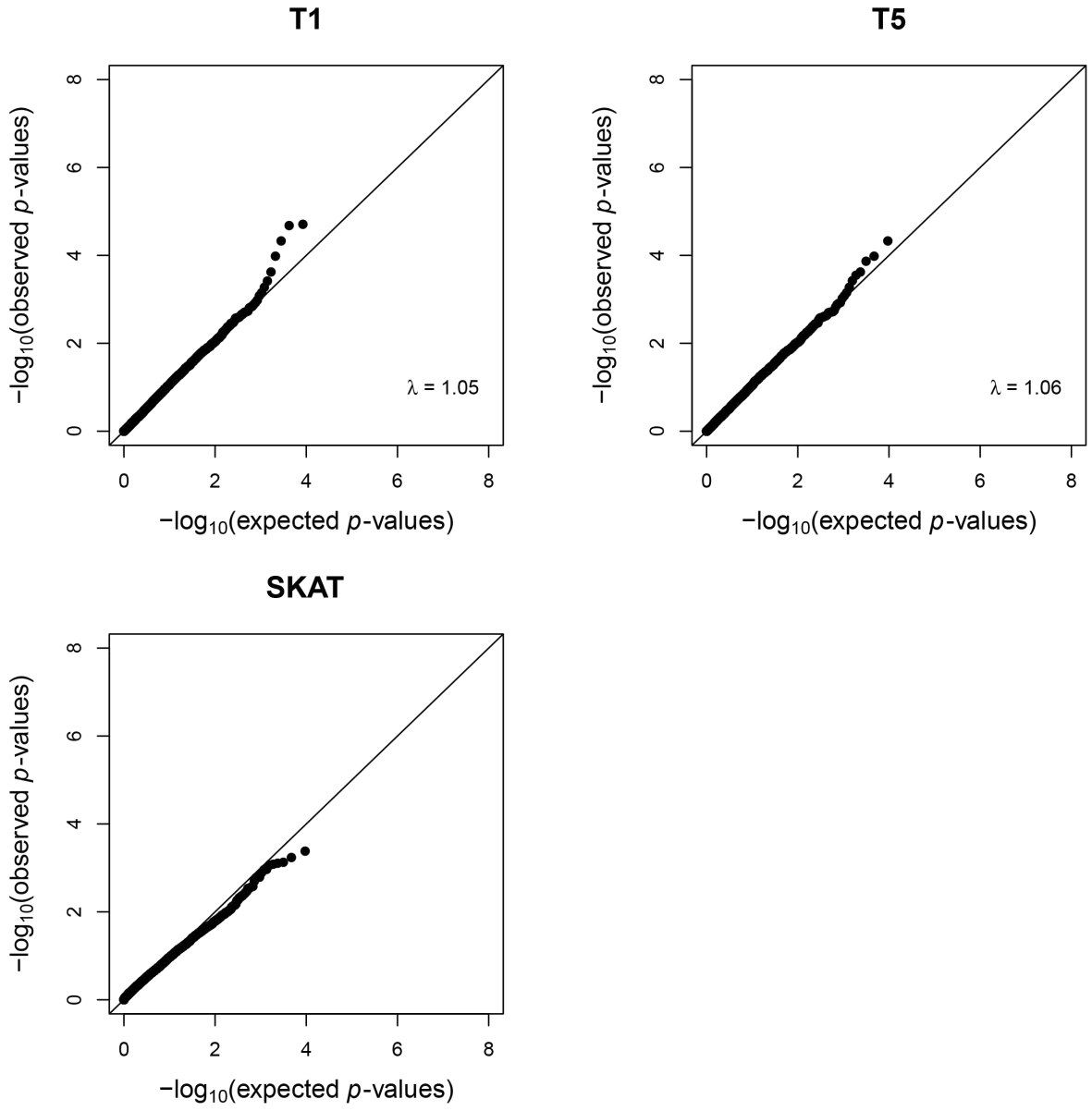

Figure 3.9: Quantile-quantile plots for the T1, T5, and SKAT tests of the global null hypothesis in the NHLBI ESP EA sample. The values of the genomic control λ are also shown for the T1 and T5 tests.

Table 3.16: Top 10 Genes for the T1 Tests in the Analysis of the LDL Data Using the MLE Method in the NHLBI ESP EA Sample

		MLE		LS	
Gene	MAC	ALL studies	TDS study	ALL studies	TDS study
LDLR	70	6.90E-07	2.12E-05	1.38E-04	2.39E-05
AZIN1	42	7.71E-05	1.96E-04	1.43E-02	2.12E-04
ACTL6A	54	1.52E-04	4.31E-03	1.74E-03	7.03E-03
PPP1R15A	59	2.20E-04	2.55E-03	4.53E-03	3.64E-03
ZFP91	14	2.46E-04	5.10E-04	5.06E-03	7.96E-04
MAGEB10	15	4.64E-04	2.30E-01	3.76E-05	2.28E-01
JAKMIP2	15	4.94E-04	5.07E-02	1.67E-03	6.11E-02
C14orf21	39	7.22E-04	4.64E-03	1.16E-02	5.69E-03
NCOA3	43	9.19E-04	2.29E-03	3.24E-02	3.01E-03
PHC2	28	1.06E-03	1.00E-01	1.37E-03	1.25E-01

3.6 Discussion

Multivariate TDS is a useful and cost-effective design when investigators are interested in multiple quantitative traits but cannot afford to sequence all cohort members. The CHARGE-TSS and NHLBI ESP are two recent examples of this design. It is not hard to envision that many large-scale whole-exome and whole-genome sequencing projects will adopt similar multivariate TDS designs. As demonstrated in the simulation studies and in the two real examples, standard linear regression without regard to the sampling design can result in estimation bias, type I error inflation, and power loss, and the existing methods for univariate TDS have important limitations.

In this paper, we propose for the first time a valid and efficient likelihood-based approach to making inferences under multivariate TDS, paying special attention to gene-level tests for rare variants. The methodology is very general and can be applied to both genetic and non-genetic studies. The proposed EM algorithm is stable and the software is available on our website.

Our approach is scalable to whole-exome and whole-genome sequencing studies. In our

Table 3.17: Top 10 Genes for the T5 Tests in the Analysis of the LDL Data Using the MLE Method in the NHLBI ESP EA Sample

		MLE		LS	
Gene	MAC	ALL studies	TDS study	ALL studies	TDS study
LDLR	70	6.90E-07	2.12E-05	1.38E-04	2.39E-05
AZIN1	42	7.71E-05	1.96E-04	1.43E-02	2.12E-04
ACTL6A	54	1.52E-04	4.31E-03	1.74E-03	7.03E-03
PPP1R15A	59	2.20E-04	2.55E-03	4.53E-03	3.64E-03
MAGEB10	17	4.05E-04	6.24E-01	2.22E-05	6.06E-01
IGSF1	117	4.21E-04	2.13E-02	3.18E-04	2.20E-02
JAKMIP2	15	4.94E-04	5.07E-02	1.67E-03	6.11E-02
C14orf21	41	5.42E-04	4.64E-03	8.33E-03	5.69E-03
TCF20	95	7.41E-04	5.62E-03	1.37E-02	5.90E-03
MACC1	143	7.54E-04	5.03E-03	1.07E-02	4.76E-03

Table 3.18: Top 10 Genes for the MB Tests in the Analysis of the LDL Data Using the MLE Method in the NHLBI ESP EA Sample

		MLE		LS	
Gene	MAC	ALL studies	TDS study	ALL studies	TDS studies
LDLR	70	1.19E-07	4.44E-06	1.59E-05	6.44E-06
SERPINB6	13	1.86E-04	4.43E-02	5.40E-04	7.81E-02
OSBPL11	11	2.41E-04	8.56E-04	1.45E-02	2.18E-03
ZFP91	123	2.51E-04	5.26E-04	6.39E-03	8.25E-04
EFEMP2	23	3.08E-04	8.67E-02	1.95E-04	8.92E-02
NLRC5	390	5.23E-04	1.11E-02	4.06E-03	1.52E-02
COBLL1	216	7.05E-04	1.88E-02	4.12E-03	2.72E-02
DSCC1	31	7.13E-04	8.94E-02	3.20E-04	1.08E-01
JAKMIP2	15	7.64E-04	5.07E-02	2.05E-03	6.11E-02
USP54	121	9.36E-04	1.07E-02	5.03E-03	1.09E-02

Table 3.19: Top 10 Genes for the SKAT Tests in the Analysis of the LDL Data Using the MLE Method in the NHLBI ESP EA Sample

		MLE		LS	
Gene	MAC	ALL studies	TDS study	ALL studies	TDS study
IL8	5	1.55E-06	1.92E-01	2.01E-08	2.16E-01
ECH1	6	3.72E-05	6.32E-04	5.09E-03	2.75E-03
MAGEB10	17	4.14E-05	2.42E-01	4.30E-05	2.29E-01
EGR1	12	7.00E-05	—	7.00E-05	—
PPP1R15A	59	7.42E-05	1.76E-04	8.23E-03	3.74E-04
CEP128	354	1.33E-04	8.76E-01	9.07E-05	8.66E-01
GRB14	100	3.09E-04	2.66E-02	6.60E-04	2.49E-02
GNA14	27	3.96E-04	9.00E-02	5.30E-04	1.03E-01
ACTL6A	54	4.60E-04	1.59E-03	1.79E-02	2.80E-03
MVK	17	5.10E-04	1.47E-04	3.28E-01	3.00E-04

Table 3.20: Top 10 Genes for the T1 Tests of the Global Null Hypothesis in the NHLBI ESP EA Sample

Gene	MAC	p-value
CCDC62	7	1.96E-05
CXCR5	5	2.10E-05
PLCG1	10	4.69E-05
LDLR	31	1.04E-04
EPHX1	13	2.39E-04
CHAF1A	12	3.83E-04
SFXN5	8	5.33E-04
AZIN1	15	7.03E-04
IGSF11	7	8.28E-04
PCK1	15	1.05E-03

Table 3.21: Top 10 Genes for the T5 Tests of the Global Null Hypothesis in the NHLBI ESP EA Sample

Gene	MAC	p-value
PLCG1	10	4.69E-05
LDLR	31	1.04E-04
AQP8	66	1.36E-04
EPHX1	13	2.39E-04
PHKB	66	2.83E-04
RETSAT	74	3.77E-04
SFXN5	8	5.33E-04
AZIN1	15	7.03E-04
IGSF11	7	8.28E-04
NSMAF	23	9.53E-04

Table 3.22: Top 10 Genes for the SKAT Tests of the Global Null Hypothesis in the NHLBI ESP EA Sample

Gene	MAC	p-value
RPP38	61	4.16E-04
PPP1R15A	21	5.80E-04
CKM	37	7.44E-04
C22orf31	37	7.83E-04
REV3L	103	8.25E-04
MRPS6	21	8.66E-04
MVK	6	1.08E-03
SPG7	29	1.10E-03
ARNTL2	119	1.29E-03
C7orf58	26	1.62E-03

single-variant analysis of the NHLBI ESP EA data, it took ∼5 seconds on an IBM HS21 machine to perform one association analysis. The computation time increases as the number of traits or the percentage of missing data increases. When there are no covariates or covariates are categorical (i.e. when m is small), the computation is fast. When there are continuous covariates, we recommend splitting the genome and using multiple CPUs.

As shown in the simulation studies, the MLE method has appreciable bias and inflated type I error when the normality assumption on $\boldsymbol{\epsilon}$ is severely violated. In practice, one should inspect the trait distributions and explore parametric transformations, such as the log transformation, or the rank-based inverse normal transformation. In genome-wide studies, a well-behaved quantile-quantile plot for the association tests would imply that non-normality has no undue influence on the type I error.

For single-variant analysis, we compared the MLE method with the univariate LS method. It is also possible to consider the multivariate LS method. If one is only interested in the marginal genetic effects on each trait and the traits are completely observed for all sequenced individuals, then univariate and multivariate LS methods yield the same results. If there is a small proportion of missingness, then the two methods should still yield similar results. If one is interested in the joint genetic effects on multiple traits, then a multivariate model is necessary. We adopt a multivariate model in our MLE approach primarily because the sampling scheme involves multiple traits. Our model is more elaborate than a univariate model, but it is the only approach that provides valid and efficient inferences for the multivariate TDS design.

In both the simulation studies and the real examples, all traits in the model are used in the sampling process. In practice, investigators may be interested in secondary quantitative traits which are not directly used for sampling but are correlated with the primary traits. (Note that standard linear regression is valid only when a secondary trait is independent of all primary traits, which is an unlikely scenario.) It is straightforward to analyze secondary

traits with our MLE method. Using a multivariate normal distribution for the primary and secondary traits, one can include each secondary trait of interest as an additional "primary" trait and use our MLE method with these $(K+1)$ traits.

Our approach does not require $\boldsymbol{Z}$ for nonsequenced individuals. In the NHLBI ESP, part of $\boldsymbol{Z}$ (sequencing centers/targets) is not available for nonsequencd individuals. In the CHARGE-TSS, $\boldsymbol{Z}$ is available for all individuals. Incorporating $\boldsymbol{Z}$ of nonsequenced individuals into the analysis has two advantages. First, it allows the selection of individuals for sequencing to depend on $\boldsymbol{Z}$. Second, it improves the efficiency of estimation. Then the likelihood involves the conditional distribution of $\boldsymbol{G}$ given $\boldsymbol{Z}^{(1)}$, which is the part of $\boldsymbol{Z}$ that is correlated with $\boldsymbol{G}$. We plan to incorporate kernel smoothing into the likelihood to handle continuous components in $\boldsymbol{Z}^{(1)}$. Table 3.23 shows the estimated distribution of $(\boldsymbol{Z}, \boldsymbol{G})$ in the analysis of the second most significant SNP in the NHLBI ESP EA sample; there is no strong evidence of correlation between $\boldsymbol{Z}$ and $\boldsymbol{G}$. A similar issue arises when some part of $\boldsymbol{Z}$ is subject to missingness. We denote that part of $\boldsymbol{Z}$ and $\boldsymbol{G}$ as $\widetilde{\boldsymbol{G}}$ and denote the rest of $\boldsymbol{Z}$ as $\widetilde{\boldsymbol{Z}}$. We plan to formulate the conditional distribution of $\widetilde{\boldsymbol{G}}$ given $\widetilde{\boldsymbol{Z}}$ through general odds ratio functions (Hu et al. 2010).

We have focused on the inference procedures rather than the design aspects. Although our simulation studies indicate that the two-tail design can be more efficient than the one-tail design, the optimal design remains unknown. It is unclear what the best sampling strategy is when multiple quantitative traits are of equal interest. Because our likelihood framework applies to any multivariate TDS, our variance formulas can be used to compare the efficiencies of different designs.

Table 3.23: Estimation of $f(Z)$, $f(Z,G)$, and $f(G|Z)$ in the Analysis of the Second Most Significant SNP of the LDL data in the NHLBI ESP EA Sample

Z			$\widehat{f}(Z,G)$			$\widehat{f}(G\|Z)$		
Center	Target	$\widehat{f}(Z)$	$G=0$	$G=1$	$G=2$	$G=0$	$G=1$	$G=2$
ARIC	broad_ESP_new	0.165	0.122	0.040	0.003	0.741	0.241	0.019
ARIC	uwrefseq_2009	0.317	0.204	0.104	0.009	0.645	0.327	0.029
CARDIA	broad_ESP_new	0.118	0.079	0.039	0.000	0.671	0.329	0.000
CARDIA	V2refseq2010	0.024	0.015	0.009	0.000	0.636	0.364	0.000
CHS	broad_ESP_new	0.005	0.005	0.000	0.000	1.000	0.000	0.000
CHS	uwrefseq_2009	0.027	0.019	0.008	0.000	0.693	0.307	0.000
FHS	broad_ESP_new	0.097	0.075	0.018	0.003	0.778	0.189	0.033
FHS	uwrefseq_2009	0.022	0.019	0.002	0.002	0.853	0.079	0.068
MESA	broad_ESP_new	0.032	0.017	0.015	0.000	0.523	0.477	0.000
MESA	V2refseq2010	0.097	0.068	0.027	0.002	0.702	0.282	0.016
WHI	broad_ESP_new	0.013	0.010	0.003	0.000	0.751	0.249	0.000
WHI	V2refseq2010	0.084	0.061	0.018	0.005	0.734	0.211	0.055

3.7 Theoretical Details

3.7.1 Derivation of the Observed-Data Likelihood

Let $\boldsymbol{V}_i \equiv (V_{i1}, \ldots, V_{iK})^{\mathrm{T}}$ be a $K \times 1$ vector of ones and zeros indicating which components of $\boldsymbol{Y}_i$ are observed or missing for the ith individual. Let R_i indicate, by the values 1 versus 0, whether the ith individual is selected for sequencing. We make the following assumptions:

Assumption 3.1. The conditional distribution of $\boldsymbol{V}_i$ given $(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i)$ is a function of $(\boldsymbol{Y}_i^{obs}, \boldsymbol{Z}_i, \boldsymbol{G}_i)$ for sequenced individuals and a function of $\boldsymbol{Y}_i^{obs}$ for nonsequenced individuals.

Assumption 3.2. The distribution of $\boldsymbol{R} \equiv (R_1, \ldots, R_N)$ depends on $(\boldsymbol{V}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{G}) \equiv \{(\boldsymbol{V}_1, \boldsymbol{Y}_1, \boldsymbol{Z}_1, \boldsymbol{G}_1), \ldots, (\boldsymbol{V}_N, \boldsymbol{Y}_N, \boldsymbol{Z}_N, \boldsymbol{G}_N)\}$ only through $\boldsymbol{V} \circ \boldsymbol{Y} \equiv (\boldsymbol{V}_1 \circ \boldsymbol{Y}_1, \ldots, \boldsymbol{V}_N \circ \boldsymbol{Y}_N)$, where "$\circ$" denotes component-wise product.

Assumption 3.3. $f(\boldsymbol{R}|\boldsymbol{V} \circ \boldsymbol{Y}) \prod_{i=1}^{n} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i) \prod_{i=n+1}^{N} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i)$ does not contain parameters $\boldsymbol{\theta}$ and F.

Under Assumptions 3.1-3.2, the complete-data density for the underlying variables $(R_i, \boldsymbol{V}_i,$

$\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i),\ i = 1, \ldots, N$, is

$$
\begin{aligned}
& f(\boldsymbol{R}, \boldsymbol{V}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{G}) \\
=& f(\boldsymbol{R}|\boldsymbol{V} \circ \boldsymbol{Y}) \prod_{i=1}^{N} f(\boldsymbol{V}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i) \\
=& f(\boldsymbol{R}|\boldsymbol{V} \circ \boldsymbol{Y}) \prod_{i=1}^{n} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i) f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{Z}_i, \boldsymbol{G}_i) f(\boldsymbol{Z}_i, \boldsymbol{G}_i) \\
& \times \prod_{i=n+1}^{N} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i) f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{Z}_i, \boldsymbol{G}_i) f(\boldsymbol{Z}_i, \boldsymbol{G}_i).
\end{aligned}
$$

The observed data are $(R_i, \boldsymbol{V}_i, \boldsymbol{V}_i \circ \boldsymbol{Y}_i, R_i\boldsymbol{Z}_i, R_i\boldsymbol{G}_i),\ i = 1, \ldots, N$, whose density is obtained by integrating over the unobserved variables in the complete-data density, i.e.,

$$
\begin{aligned}
& f(\boldsymbol{R}, \boldsymbol{V}, \boldsymbol{V} \circ \boldsymbol{Y}, \boldsymbol{R} \circ \boldsymbol{Z}, \boldsymbol{R} \circ \boldsymbol{G}) \\
=& f(\boldsymbol{R}|\boldsymbol{V} \circ \boldsymbol{Y}) \prod_{i=1}^{n} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i) \left\{ \int_{\boldsymbol{Y}^{mis}} f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{Z}_i, \boldsymbol{G}_i) d\boldsymbol{Y}^{mis} \right\} f(\boldsymbol{Z}_i, \boldsymbol{G}_i) \\
& \times \prod_{i=n+1}^{N} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i) \int_{\boldsymbol{z},\boldsymbol{g}} \left\{ \int_{\boldsymbol{Y}^{mis}} f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{z}, \boldsymbol{g}) d\boldsymbol{Y}^{mis} \right\} dF(\boldsymbol{z}, \boldsymbol{g}) \\
=& f(\boldsymbol{R}|\boldsymbol{V} \circ \boldsymbol{Y}) \prod_{i=1}^{n} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i) \prod_{i=n+1}^{N} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i) \\
& \times \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{Z}_i, \boldsymbol{G}_i) f(\boldsymbol{Z}_i, \boldsymbol{G}_i) \prod_{i=n+1}^{N} \int_{\boldsymbol{z},\boldsymbol{g}} f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{z}, \boldsymbol{g}) dF(\boldsymbol{z}, \boldsymbol{g}),
\end{aligned}
$$

where $\boldsymbol{R} \circ \boldsymbol{Z} = (R_1\boldsymbol{Z}_1, \ldots, R_N\boldsymbol{Z}_N)$, $\boldsymbol{R} \circ \boldsymbol{G} = (R_1\boldsymbol{G}_1, \ldots, R_N\boldsymbol{G}_N)$, and $\boldsymbol{Y}^{mis}$ is the missing part of $\boldsymbol{Y}$. We can ignore $f(\boldsymbol{R}|\boldsymbol{V} \circ \boldsymbol{Y}) \prod_{i=1}^{n} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i) \prod_{i=n+1}^{N} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i)$ because of Assumption 3.3. The remaining part of the above density is exactly the observed-data likelihood given in (3.20).

3.7.2 Estimation

To calculate the MLEs for (3.21), we use the EM algorithm in which missing data contain the partially missing $\boldsymbol{Y}_i$'s and the missing observations on $(\boldsymbol{Z}, \boldsymbol{G})$ for individuals not selected for sequencing. The complete-data log-likelihood function is

$$\sum_{i=1}^{N}\left[\sum_{j=1}^{m} I\{(\boldsymbol{Z}_i, \boldsymbol{G}_i) = (\boldsymbol{z}_j, \boldsymbol{g}_j)\}\left\{\log f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{z}_j, \boldsymbol{g}_j) + \log q_j\right\}\right].$$

At the tth iteration, the M-step maximizes

$$\sum_{i=1}^{N}\sum_{j=1}^{m} \widehat{\psi}_{ij}^{(t)}\left[E\{\log f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{z}_j, \boldsymbol{g}_j)|\boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j; \widehat{\boldsymbol{\theta}}^{(t)}\} + \log q_j\right],$$

where $E(\cdot|\boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j; \widehat{\boldsymbol{\theta}}^{(t)})$ is the conditional expectation given $\boldsymbol{Y}_i^{obs}$, $(\boldsymbol{Z}_i, \boldsymbol{G}_i) = (\boldsymbol{z}_j, \boldsymbol{g}_j)$, evaluated at $\widehat{\boldsymbol{\theta}}^{(t)}$, and $\widehat{\psi}_{ij}^{(t)}$ is the conditional probability of $I\{(\boldsymbol{Z}_i, \boldsymbol{G}_i) = (\boldsymbol{z}_j, \boldsymbol{g}_j)\} = 1$ given $\boldsymbol{Y}_i^{obs}$, $(\boldsymbol{z}_1, \boldsymbol{g}_1), \ldots, (\boldsymbol{z}_m, \boldsymbol{g}_m)$, evaluated at $\widehat{\boldsymbol{\theta}}^{(t)}$, $\widehat{q}_1^{(t)}, \ldots, \widehat{q}_m^{(t)}$. That is,

$$\widehat{\psi}_{ij}^{(t)} = \begin{cases} I\{(\boldsymbol{Z}_i, \boldsymbol{G}_i) = (\boldsymbol{z}_j, \boldsymbol{g}_j)\} & i = 1, \ldots, n; \\ \frac{f_{\widehat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{z}_j, \boldsymbol{g}_j)\widehat{q}_j^{(t)}}{\sum_{l=1}^{m} f_{\widehat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{z}_l, \boldsymbol{g}_l)\widehat{q}_l^{(t)}} & i = n+1, \ldots, N. \end{cases}$$

Write $\boldsymbol{W}_j = (\boldsymbol{g}_j^{\mathrm{T}}, \boldsymbol{z}_j^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. The M-step involves the following calculations:

$$(\widehat{\boldsymbol{\eta}}_k^{(t+1)})^{\mathrm{T}} = \left(\sum_{i=1}^{N}\sum_{j=1}^{m} \widehat{\psi}_{ij}^{(t)} \boldsymbol{W}_j^{\otimes 2}\right)^{-1}\left[\sum_{i=1}^{N}\sum_{j=1}^{m} \widehat{\psi}_{ij}^{(t)} E\{Y_{ki}|\boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j; \widehat{\boldsymbol{\theta}}^{(t)}\}\boldsymbol{W}_j\right], \ 1 \le k \le K,$$

$$\widehat{\boldsymbol{\Sigma}}^{(t+1)} = N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{m} \widehat{\psi}_{ij}^{(t)} E\left\{\left(\boldsymbol{Y}_i - \widehat{\boldsymbol{\eta}}^{(t+1)}\boldsymbol{W}_j\right)^{\otimes 2}|\boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j; \widehat{\boldsymbol{\theta}}^{(t)}\right\},$$

$$\widehat{q}_j^{(t+1)} = N^{-1}\sum_{i=1}^{N} \widehat{\psi}_{ij}^{(t)},$$

where $\boldsymbol{\eta}_k$ is the kth row of $\boldsymbol{\eta}$, and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$. We start with initial values $\widehat{\boldsymbol{\eta}}^{(0)} = \mathbf{0}$, $\widehat{\boldsymbol{\Sigma}}^{(0)}$ being the sample covariance matrix based on those $\boldsymbol{Y}_i$'s with complete observations, and $\widehat{q}_j^{(0)} = n^{-1}\sum_{i=1}^{n} I\{(\boldsymbol{Z}_i, \boldsymbol{G}_i) = (\boldsymbol{z}_j, \boldsymbol{g}_j)\}$, $j = 1, \ldots, m$, and iterate until convergence to obtain the MLEs $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\Sigma}}, \widehat{q}_1, \cdots, \widehat{q}_m)$. In the above expressions, the conditional expectations can be evaluated by using the fact that the missing part of $\boldsymbol{Y}_i$, denoted by $\boldsymbol{Y}_i^{mis}$, given $\boldsymbol{Y}_i^{obs}$ and $(\boldsymbol{z}_j, \boldsymbol{g}_j)$, follows a normal distribution with mean $\boldsymbol{\beta}_i^{mis}\boldsymbol{g}_j + \boldsymbol{\gamma}_i^{mis}\boldsymbol{z}_j + \boldsymbol{\Sigma}_i^{mo}\{\boldsymbol{\Sigma}_i^{oo}\}^{-1}(\boldsymbol{Y}_i^{obs} - \boldsymbol{\beta}_i^{obs}\boldsymbol{g}_j - \boldsymbol{\gamma}_i^{obs}\boldsymbol{z}_j)$ and variance $\boldsymbol{\Sigma}_i^{mm} - \boldsymbol{\Sigma}_i^{mo}\{\boldsymbol{\Sigma}_i^{oo}\}^{-1}\{\boldsymbol{\Sigma}_i^{mo}\}^{\mathrm{T}}$, where $\boldsymbol{\beta}_i^{mis}$ and $\boldsymbol{\beta}_i^{obs}$ are the corresponding parts for $\boldsymbol{Y}_i^{mis}$ and $\boldsymbol{Y}_i^{obs}$ in $\boldsymbol{\beta}$, and the same partitions apply to $\boldsymbol{\gamma}$ to yield $\boldsymbol{\gamma}_i^{mis}$ and $\boldsymbol{\gamma}_i^{obs}$ and to $\boldsymbol{\Sigma}$ to yield $\boldsymbol{\Sigma}_i^{mm}, \boldsymbol{\Sigma}_i^{mo}$, and $\boldsymbol{\Sigma}_i^{oo}$.

We estimate the asymptotic covariance matrix of the MLEs by the Louis formula (Louis 1982). We use A_{kl} to denote the (k,l)th element of any matrix $\boldsymbol{A}$. For $i = 1, \ldots, N$ and $j = 1, \ldots, m$, we calculate the derivatives of $\log f(\boldsymbol{Y}_i|\boldsymbol{z}_j, \boldsymbol{g}_j) + \log q_j$ to obtain the $\{K(p+d) + K(K+1)/2 + m\} \times 1$ complete-data score vector

$$\boldsymbol{l}_{1ij} = \left[\boldsymbol{S}_{1ij}^{\mathrm{T}}, \ldots, \boldsymbol{S}_{Kij}^{\mathrm{T}}, T_{11ij}, T_{12ij}, \ldots, T_{KKij}, \boldsymbol{P}_{ij}^{\mathrm{T}}\right]^{\mathrm{T}},$$

where $\boldsymbol{S}_{kij} = \boldsymbol{W}_j\boldsymbol{e}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\eta}}\boldsymbol{W}_j)$, with $\boldsymbol{e}_k$ being the kth canonical vector of length K, i.e. with 1 in the kth position and 0 in all the other positions,

$$\begin{aligned} T_{klij} = &-\frac{1}{2}\{1 + I(k \neq l)\}(\widehat{\boldsymbol{\Sigma}}^{-1})_{kl} \\ &+ \frac{1}{4}\{1 + I(k \neq l)\}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\eta}}\boldsymbol{W}_j)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{e}_{kl} + \boldsymbol{e}_{lk})\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\eta}}\boldsymbol{W}_j), \quad k \leq l, \end{aligned}$$

with $\boldsymbol{e}_{kl} = \boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}$ and $\boldsymbol{P}_{ij} = (0, \ldots, 0, 1/\widehat{q}_j, 0, \ldots, 0)^{\mathrm{T}}$. We also calculate the second derivatives as a $\{K(p+d) + K(K+1)/2 + m\} \times \{K(p+d) + K(K+1)/2 + m\}$ matrix, which is the

block diagonal matrix

$$\boldsymbol{l}_{2ij} = \begin{bmatrix} \boldsymbol{l}_{11ij} & \mathbf{0}_{\{K(p+d)+K(K+1)/2\}\times m} \\ \mathbf{0}_{m\times\{K(p+d)+K(K+1)/2\}} & \boldsymbol{l}_{22ij} \end{bmatrix},$$

where

$$\boldsymbol{l}_{11ij} = \begin{bmatrix} \frac{\partial \boldsymbol{S}_{1ij}}{\partial \boldsymbol{\eta}_1} & \dots & \frac{\partial \boldsymbol{S}_{1ij}}{\partial \boldsymbol{\eta}_K} & \frac{\partial \boldsymbol{S}_{1ij}}{\partial \Sigma_{11}} & \dots & \frac{\partial \boldsymbol{S}_{1ij}}{\partial \Sigma_{KK}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \boldsymbol{S}_{Kij}}{\partial \boldsymbol{\eta}_1} & \dots & \frac{\partial \boldsymbol{S}_{Kij}}{\partial \boldsymbol{\eta}_K} & \frac{\partial \boldsymbol{S}_{Kij}}{\partial \Sigma_{11}} & \dots & \frac{\partial \boldsymbol{S}_{Kij}}{\partial \Sigma_{KK}} \\ \frac{\partial \boldsymbol{S}_{1ij}}{\partial \Sigma_{11}}^{\mathrm{T}} & \dots & \frac{\partial \boldsymbol{S}_{Kij}}{\partial \Sigma_{11}}^{\mathrm{T}} & \frac{\partial T_{11ij}}{\partial \Sigma_{11}} & \dots & \frac{\partial T_{11ij}}{\partial \Sigma_{KK}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \boldsymbol{S}_{1ij}}{\partial \Sigma_{KK}}^{\mathrm{T}} & \dots & \frac{\partial \boldsymbol{S}_{Kij}}{\partial \Sigma_{KK}}^{\mathrm{T}} & \frac{\partial T_{KKij}}{\partial \Sigma_{11}} & \dots & \frac{\partial T_{KKij}}{\partial \Sigma_{KK}} \end{bmatrix},$$

and $\boldsymbol{l}_{22ij}$ is a diagonal matrix with diagonal elements $\{0,\dots,0,-1/\widehat{q}_j^2,0,\dots,0\}$. In the above matrix,

$$\frac{\partial \boldsymbol{S}_{kij}}{\partial \boldsymbol{\eta}_l} = -\boldsymbol{W}_j\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{e}_k{}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{e}_l,$$

$$\begin{aligned}
\frac{\partial \boldsymbol{S}_{kij}}{\partial \Sigma_{k'l'}} =& -\frac{1}{2}\{1+I(k'\neq l')\}\boldsymbol{W}_j\boldsymbol{e}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{e}_{k'l'}+\boldsymbol{e}_{l'k'})\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}_i-\widehat{\boldsymbol{\eta}}\boldsymbol{W}_j), \\
\frac{\partial T_{klij}}{\partial \Sigma_{k'l'ij}} =& \frac{1}{4}\{1+I(k\neq l)\}\{1+I(k'\neq l')\}\left\{\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{e}_{k'l'}+\boldsymbol{e}_{l'k'})\widehat{\boldsymbol{\Sigma}}^{-1}\right\}_{kl} \\
& -\frac{1}{8}\{1+I(k\neq l)\}\{1+I(k'\neq l')\}(\boldsymbol{Y}_i-\widehat{\boldsymbol{\eta}}\boldsymbol{W}_j)^{\mathrm{T}} \\
& \quad\left\{\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{e}_{k'l'}+\boldsymbol{e}_{l'k'})\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{e}_{kl}+\boldsymbol{e}_{lk})\widehat{\boldsymbol{\Sigma}}^{-1}\right\}(\boldsymbol{Y}_i-\widehat{\boldsymbol{\eta}}\boldsymbol{W}_j) \\
& -\frac{1}{8}\{1+I(k\neq l)\}\{1+I(k'\neq l')\}(\boldsymbol{Y}_i-\widehat{\boldsymbol{\eta}}\boldsymbol{W}_j)^{\mathrm{T}} \\
& \quad\left\{\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{e}_{kl}+\boldsymbol{e}_{lk})\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{e}_{k'l'}+\boldsymbol{e}_{l'k'})\widehat{\boldsymbol{\Sigma}}^{-1}\right\}(\boldsymbol{Y}_i-\widehat{\boldsymbol{\eta}}\boldsymbol{W}_j).
\end{aligned}$$

We then calculate the information matrix as

$$\boldsymbol{Q} = -\sum_{i=1}^{N}\sum_{j=1}^{m}\widehat{\psi_{ij}}E\{\boldsymbol{l}_{2ij}|\boldsymbol{Y}_i^{obs},\boldsymbol{z}_j,\boldsymbol{g}_j\} - \sum_{i=1}^{N}\left[\sum_{j=1}^{m}\widehat{\psi_{ij}}E\{\boldsymbol{l}_{1ij}^{\otimes 2}|\boldsymbol{Y}_i^{obs},\boldsymbol{z}_j,\boldsymbol{g}_j\} - \left(\sum_{j=1}^{m}\widehat{\psi_{ij}}E\{\boldsymbol{l}_{1ij}|\boldsymbol{Y}_i^{obs},\boldsymbol{z}_j,\boldsymbol{g}_j\}\right)^{\otimes 2}\right].$$

To account for the constraint that $\sum_{j=1}^{m} q_j = 1$, we define $\boldsymbol{D}$ to be the derivative matrix of $(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\Sigma},q_1,\cdots,q_m)$ with respect to $(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\Sigma},q_1,\cdots,q_{m-1})$. Then, the covariance matrix for $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\Sigma}},\widehat{q_1},\cdots,\widehat{q_{m-1}})$ is estimated by $\boldsymbol{\Omega} = \boldsymbol{F}^{-1}$, where $\boldsymbol{F} = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{D}$.

3.7.3 Asymptotic Properties

Let $\boldsymbol{\Theta}$ denote the parameter space of $\boldsymbol{\theta}$, which is a bounded open set in the interior of the domain of $\boldsymbol{\theta}$, and $\mathcal{F}$ denote the space of the joint distributions of $(\boldsymbol{Z},\boldsymbol{G})$. Let $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ and $F_0 \in \mathcal{F}$ denote the true values of $\boldsymbol{\theta}$ and F. We impose the following regularity conditions and state the asymptotic results in Theorem 1.

Assumption 3.4. With probability one, $\Pr(R=1, V_k = V_l = 1|\boldsymbol{V}\circ\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{G})$ is bounded away from zero, for each pair of k and $l \in \{1,\ldots,K\}$.

Assumption 3.5. For any nonzero $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, $\Pr(\boldsymbol{\beta}\boldsymbol{G}+\boldsymbol{\gamma}\boldsymbol{Z}=\boldsymbol{0}) < 1$.

Assumption 3.6. The density function of F_0 is positive in its support and continuously differentiable with respect to a suitable measure.

Theorem 3.1. Under Assumptions 3.1–3.6, $\widehat{\boldsymbol{\theta}}$ and $\widehat{F}(\cdot,\cdot)$ are consistent in that $|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0| + \sup_{\boldsymbol{z},\boldsymbol{g}}|\widehat{F}(\boldsymbol{z},\boldsymbol{g})-F_0(\boldsymbol{z},\boldsymbol{g})| \to 0$ almost surely. In addition, $\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

Proof. The observed-data likelihood given in (3.20) is similar to the likelihood given in (6) of Lin and Zeng (2006), which pertains to haplotype rather than genotype effects. In (3.20), $f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs}|\boldsymbol{Z},\boldsymbol{G})$ is the density of a multivariate linear regression model with partial missingness

in $\boldsymbol{Y}$, whereas in (6) of Lin and Zeng (2006), $m_g(Y, \boldsymbol{X}; \boldsymbol{\theta})$, which reduces to $P_{\alpha,\boldsymbol{\beta},\boldsymbol{\xi}}(Y|\boldsymbol{X})$ when haplotypes are replaced by genotypes, is the density of a univariate generalized linear model with Y being always observed. If we can verify that Conditions 1–3 for $P_{\alpha,\boldsymbol{\beta},\boldsymbol{\xi}}(\boldsymbol{Y}|\boldsymbol{X})$ in Lin and Zeng (2006) are satisfied by $f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs}|\boldsymbol{Z}, \boldsymbol{G})$, we can use Theorem 1 of Lin and Zeng (2006) to show the consistency, asymptotic normality, and asymptotic efficiency of our estimators.

Before verifying Conditions 1-3 in Lin and Zeng (2006), we need some additional notation. Suppose that there are s distinct missing patterns in $\boldsymbol{Y}$, each with a positive probability of being observed. Let δ_t be the indicator of the tth missing pattern. Let $\boldsymbol{Y}^{obs(t)}$ and $\boldsymbol{Y}^{mis(t)}$ denote the observed and missing parts of $\boldsymbol{Y}$ for the tth missing pattern, $t = 1, \ldots, s$. Then $f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs}|\boldsymbol{Z}, \boldsymbol{G})$ can be rewritten as $\prod_{t=1}^{s} \left\{ f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs(t)}|\boldsymbol{Z}, \boldsymbol{G}) \right\}^{\delta_t}$.

Condition 1 in Lin and Zeng (2006) pertains to the identifiability of the regression model. Suppose that two sets of parameters $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$ yield the same likelihood value. Then $\prod_{t=1}^{s} \left\{ f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs(t)}|\boldsymbol{Z}, \boldsymbol{G}) \right\}^{\delta_t} = \prod_{t=1}^{s} \left\{ f_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{Y}^{obs(t)}|\boldsymbol{Z}, \boldsymbol{G}) \right\}^{\delta_t}$ for sequenced individuals. By Assumption 3.4, we can find, for each pair of k and $l \in \{1, \ldots, K\}$, some $t_0 \in \{1, \ldots, s\}$, such that Y_k and Y_l are observed in the t_0th missing pattern. Setting $\delta_{t_0} = 1$, $\delta_t = 0$, and $t \neq t_0$, we have $f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs(t_0)}|\boldsymbol{Z}, \boldsymbol{G}) = f_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{Y}^{obs(t_0)}|\boldsymbol{Z}, \boldsymbol{G})$, where both sides are multivariate normal densities. Because Y_k and Y_l are components of $\boldsymbol{Y}^{obs(t_0)}$, we have $\boldsymbol{\eta}^k = \widetilde{\boldsymbol{\eta}}^k$, $\boldsymbol{\eta}^l = \widetilde{\boldsymbol{\eta}}^l$, $\Sigma_{kk} = \widetilde{\Sigma}_{kk}$, $\Sigma_{ll} = \widetilde{\Sigma}_{ll}$, and $\Sigma_{kl} = \widetilde{\Sigma}_{kl}$. Condition 1 in Lin and Zeng (2006) is verified.

Conditions 2 and 3 in Lin and Zeng (2006) are the same if we replace haplotypes by genotypes. Thus, it remains to show that the information operator for $\boldsymbol{\theta}$ and F is continuously invertible at the true parameter values. This is tantamount to showing that the score function at any non-trivial submodel is non-zero because the information operator is the sum of an invertible operator and a compact operator mapping the score space of $(\boldsymbol{\theta}_0, F_0)$ to

itself. To this end, suppose that there exists a constant vector $\boldsymbol{u}$, such that

$$\boldsymbol{u}^{\mathrm{T}} \left\{ \sum_{t=1}^{s} \delta_t \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs(t)} | \boldsymbol{Z}, \boldsymbol{G}) \right\} = 0. \tag{3.22}$$

Let $\boldsymbol{b}^{(t)} \equiv (b_1^{(t)}, \ldots, b_K^{(t)})^{\mathrm{T}} = \{D(\boldsymbol{V}^{(t)})\boldsymbol{\Sigma}D(\boldsymbol{V}^{(t)})\}^{+}\{\boldsymbol{V}^{(t)} \circ (\boldsymbol{Y} - \boldsymbol{\eta}\boldsymbol{W})\}$, where $\boldsymbol{V}^{(t)}$ represents $\boldsymbol{V}$ in the tth missing pattern, $D(\boldsymbol{V}^{(t)})$ represents the diagonal matrix with the diagonal vector being $\boldsymbol{V}^{(t)}$, and $\boldsymbol{A}^{+}$ represents the Moore-Penrose generalized inverse of any square matrix $\boldsymbol{A}$. Then

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs(t)} | \boldsymbol{Z}, \boldsymbol{G}) = \left[(\boldsymbol{S}_1^{(t)})^{\mathrm{T}}, \ldots, (\boldsymbol{S}_K^{(t)})^{\mathrm{T}}, T_{11}^{(t)}, T_{12}^{(t)}, \ldots, T_{KK}^{(t)} \right]^{\mathrm{T}},$$

where $\boldsymbol{S}_k^{(t)} = \boldsymbol{W} b_k^{(t)}$, and

$$\begin{aligned} T_{kl}^{(t)} = & -\frac{1}{2}\{1 + I(k \neq l)\}[\{D(\boldsymbol{V}^{(t)})\boldsymbol{\Sigma}D(\boldsymbol{V}^{(t)})\}^{+}]_{kl} \\ & + \frac{1}{2}\{1 + I(k \neq l)\} b_k^{(t)} b_l^{(t)}, \quad k \leq l. \end{aligned}$$

By Assumption 3.4, we can find, for each pair of k and $l \in \{1, \ldots, K\}$, $k \leq l$, some $t_0 \in \{1, \ldots, s\}$, such that $V_k^{(t_0)} = V_l^{(t_0)} = 1$. Set $\delta_{t_0} = 1$, $\delta_t = 0$, and $t \neq t_0$. Since Y_k and Y_l can take arbitrary values and $b_k^{(t_0)}$ and $b_l^{(t_0)}$ are non-degenerate linear functions of Y_k and Y_l, we see that $b_k^{(t_0)}$ and $b_l^{(t_0)}$ can take arbitrary values. By examining the linear and quadratic terms of $b_k^{(t_0)}$ and $b_l^{(t_0)}$ in equation (3.22), we conclude that their corresponding coefficients must be zero. That is, $\boldsymbol{u}_k^{\mathrm{T}}\boldsymbol{W} = 0$, $\boldsymbol{u}_l^{\mathrm{T}}\boldsymbol{W} = 0$, and $u_{kl} = 0$, where $\boldsymbol{u}_k$, $\boldsymbol{u}_l$, and u_{kl} are the components of $\boldsymbol{u}$ associated with $\boldsymbol{S}_k^{(t)}$, $\boldsymbol{S}_l^{(t)}$, and $T_{kl}^{(t)}$, respectively. By Assumption 3.5, $\boldsymbol{u}_k = \boldsymbol{0}$ and $\boldsymbol{u}_l = \boldsymbol{0}$. It follows that $\boldsymbol{u} = \boldsymbol{0}$. Thus, the score function is non-zero at any non-trivial submodel, and Conditions 2 and 3 in Lin and Zeng (2006) hold.

Remark. Assumption 3.4 suggests that we need to observe with positive probability each pair of components of $\boldsymbol{Y}$ in some individuals selected for sequencing in order for the MLE

method to be applicable. We do not require a fully-observed $\boldsymbol{Y}$ for any individual. On the other hand, both the CHARGE-TSS ARIC data and NHLBI ESP EA data contain a large proportion of sequenced individuals with fully-observed $\boldsymbol{Y}$. Thus, Condition 1 is not an issue but mainly serves theoretical purposes.

3.7.4 Association Tests

For Wald tests employed in single-variant analysis, we estimate all parameters under the alternative hypothesis. Suppose that we decompose $\boldsymbol{\beta}$ into $(\boldsymbol{\beta}_a^{\mathrm{T}}, \boldsymbol{\beta}_b^{\mathrm{T}})^{\mathrm{T}}$ and wish to test the null hypothesis $H_0^a : \boldsymbol{\beta}_a = \mathbf{0}$. The Wald test statistic is $T_a \equiv \widehat{\boldsymbol{\beta}}_a^{\mathrm{T}} \boldsymbol{\Omega}_{aa}^{-1} \widehat{\boldsymbol{\beta}}_a$, where $\widehat{\boldsymbol{\beta}}_a$ is the MLE of $\boldsymbol{\beta}_a$, and $\boldsymbol{\Omega}_{aa}$ is the covariance matrix of $\widehat{\boldsymbol{\beta}}_a$, which is the submatrix of $\boldsymbol{\Omega}$ corresponding to $\boldsymbol{\beta}_a$. We refer T_a to the $\chi^2_{d_a}$ distribution, with the degree of freedom d_a being the dimension of $\boldsymbol{\beta}_a$. In particular, to test the genetic effect on each trait, we consider the null hypothesis $H_0^{(k)} : \beta_k = 0$ for $k = 1, \ldots, K$. The test statistic is $T_k \equiv \widehat{\beta}_k^2 / \Omega_{kk}$, where Ω_{kk} is the variance estimate of $\widehat{\beta}_k$. We refer T_k to the χ^2_1 distribution.

Gene-level tests for rare variants rely on score statistics. To test the global null hypothesis that there is no genetic effect on any trait, i.e. $H_0 : \boldsymbol{\beta} = \mathbf{0}$, we calculate the restricted MLE of $(\boldsymbol{\gamma}, \boldsymbol{\Sigma}, q_1, \cdots, q_{m-1})$ under H_0. This is accomplished through the above EM algorithm in which $\boldsymbol{\beta}$ is set to $\mathbf{0}$ and only $(\boldsymbol{\gamma}, \boldsymbol{\Sigma}, q_1, \cdots, q_{m-1})$ is estimated. The score statistic for testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is $\boldsymbol{U}_1 \equiv \sum_{i=1}^{N} \sum_{j=1}^{m} \widehat{\psi}_{ij} \boldsymbol{l}_{1ij}^{(1)}$, where $\boldsymbol{l}_{1ij}^{(1)}$ is the subvector of $\boldsymbol{l}_{1ij}$ corresponding to $\boldsymbol{\beta}$. It can be shown that $\boldsymbol{U}_1$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{V}_1 = \boldsymbol{F}_{11} - \boldsymbol{F}_{12} \boldsymbol{F}_{22}^{-1} \boldsymbol{F}_{21}$, where $\begin{bmatrix} \boldsymbol{F}_{11} & \boldsymbol{F}_{12} \\ \boldsymbol{F}_{21} & \boldsymbol{F}_{22} \end{bmatrix}$ is the partition of $\boldsymbol{F}$ with respect to $\boldsymbol{\beta}$ and the other parameters.

For T1 and T5 tests, G is the total number of mutations among variants whose MAFs are below 1% and 5%, respectively. For the MB test, G is the weighted sum of mutations with weights defined as $\{\mathrm{MAF}(1 - \mathrm{MAF})\}^{-1/2}$ for each variant (Madsen and Browning 2009). For the above three tests, G is a scalar, and $d = 1$. The test statistic for testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is

$T_{(1)} \equiv \boldsymbol{U}_1^{\mathrm{T}} \boldsymbol{V}_1^{-1} \boldsymbol{U}_1$. We refer $T_{(1)}$ to the χ^2_K distribution.

For SKAT, $\boldsymbol{G}$ is a vector of the genotypes of individual variants within a gene. A SKAT-type statistic can be defined as $\boldsymbol{Q}_2 \equiv \boldsymbol{U}_1^{\mathrm{T}} \boldsymbol{B} \boldsymbol{U}_1$, where $\boldsymbol{B}$ is a diagonal matrix of weights that depend on the MAFs through a beta function. The null distribution of $\boldsymbol{Q}_2$ is approximated by $\sum_{j=1}^{Kd} \lambda_j \chi^2_{1,j}$, where $(\lambda_1, \ldots, \lambda_{Kd})$ are the eigenvalues of $\boldsymbol{V}_1^{1/2} \boldsymbol{B} \boldsymbol{V}_1^{1/2}$, and $(\chi^2_{1,1}, \ldots, \chi^2_{1,Kd})$ are independent χ^2_1 random variables (Wu et al. 2011).

To test the genetic effect on a particular trait, say, the k_0th trait, i.e. $H_0 : \boldsymbol{\beta}_{k_0} = \boldsymbol{0}$, where $\boldsymbol{\beta}_{k_0}$ is the k_0th row of $\boldsymbol{\beta}$ reflecting the genetic effect on the k_0th trait, we estimate $(\{\boldsymbol{\eta}_k\}_{k=1,\ldots,K, k\neq k_0}, \boldsymbol{\gamma}_{k_0}, \boldsymbol{\Sigma}, q_1, \cdots, q_{m-1})$ under H_0. This is accomplished through the above EM algorithm (with a modified M-step) in which $\boldsymbol{\beta}_{k_0}$ is set to $\boldsymbol{0}$ and only $(\{\boldsymbol{\eta}_k\}_{k=1,\ldots,K, k\neq k_0}$, $\boldsymbol{\gamma}_{k_0}, \boldsymbol{\Sigma}, q_1, \cdots, q_{m-1})$ is estimated. The M-step for estimating $\boldsymbol{\eta}$ is

$$\left[\widehat{\boldsymbol{\eta}}_1^{(t+1)}, \ldots, \widehat{\boldsymbol{\gamma}}_{k_0}^{(t+1)}, \ldots, \widehat{\boldsymbol{\eta}}_K^{(t+1)}\right]^{\mathrm{T}} = \left[\boldsymbol{A}^{\mathrm{T}} \left\{ \left(\widehat{\boldsymbol{\Sigma}}^{(t)}\right)^{-1} \otimes \left(\sum_{i=1}^{N} \sum_{j=1}^{m} \widehat{\psi}_{ij}^{(t)} \boldsymbol{W}_j^{\otimes 2} \right) \right\} \boldsymbol{A} \right]^{-1}$$
$$\boldsymbol{A}^{\mathrm{T}} \left[\sum_{i=1}^{N} \sum_{j=1}^{m} \widehat{\psi}_{ij}^{(t)} \left\{ \left(\widehat{\boldsymbol{\Sigma}}^{(t)}\right)^{-1} \otimes \boldsymbol{W}_j \right\} E\{\boldsymbol{Y}_i | \boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j; \widehat{\boldsymbol{\theta}}^{(t)}\} \right],$$

where $\boldsymbol{A}$ is a $pK \times (pK - 1)$ matrix constructed by deleting the $\{p(k_0 - 1) + 1\}$th column of the $pK \times pK$ identity matrix $\boldsymbol{I}_{pK}$, and $\boldsymbol{A} \otimes \boldsymbol{B}$ denotes the Kronecker product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. The score statistic for testing $H_0 : \boldsymbol{\beta}_{k_0} = \boldsymbol{0}$ is $\boldsymbol{U}_2 \equiv \sum_{i=1}^{N} \sum_{j=1}^{m} \widehat{\psi}_{ij} \boldsymbol{l}_{1ij}^{(21)}$, where $\begin{bmatrix} \boldsymbol{l}_{1ij}^{(21)} \\ \boldsymbol{l}_{1ij}^{(22)} \end{bmatrix}$ and $\begin{bmatrix} \boldsymbol{F}_{11}^{(2)} & \boldsymbol{F}_{12}^{(2)} \\ \boldsymbol{F}_{21}^{(2)} & \boldsymbol{F}_{22}^{(2)} \end{bmatrix}$ are the partitions of $\boldsymbol{l}_{1ij}$ and $\boldsymbol{F}$ with respect to $\boldsymbol{\beta}_{k_0}$ and the other parameters. It can be shown that $\boldsymbol{U}_2$ is asymptotically normal with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{V}_2 \equiv \boldsymbol{F}_{11}^{(2)} - \boldsymbol{F}_{12}^{(2)} \left(\boldsymbol{F}_{22}^{(2)}\right)^{-1} \boldsymbol{F}_{21}^{(2)}$. All tests of $H_0 : \boldsymbol{\beta}_{k_0} = \boldsymbol{0}$ can be constructed in a similar manner. For SKAT tests, we use the vector of genotypes of individual variants as the genetic variables for the k_0th trait and use the burden scores for other traits to ensure numerical stability.

CHAPTER 4: EFFICIENT SEMIPARAMETRIC INFERENCE UNDER TWO-PHASE, OUTCOME-DEPENDENT SAMPLING

4.1 Introduction

In epidemiological studies, the outcomes of interest (e.g, anthropometry measurements, lipids levels, or disease status) and demographical and environmental variables (e.g., age, gender, and smoking status) are typically available for all subjects. However, the covariates of main interest often involve genotyping, biomarker assay, or medical imaging and thus are prohibitively expensive to measure for all subjects, especially in a large study. If disease status or another discrete outcome is of primary interest, then the case-control design with an equal number of cases and controls is the most efficient one (Scott and Wild 1997). If a continuous outcome such as height is of primary interest, then a cost-effective strategy is the "extreme-tail" sampling design, whereby one selectively measures the "expensive covariates" only for subjects with extreme values of the primary outcome measure (Lin et al. 2013). In either case, the efficiency of the design can be improved by stratifying on the "inexpensive covariates".

The case-control and extreme-tail sampling designs can be viewed as special cases of the two-phase, outcome-dependent design, which was first introduced by White (1982). In the first phase of this design, the outcome of interest and inexpensive covariates are observed for all study subjects; the information collected during the first phase is then used to determine which subjects to include for measurements on expensive covariates during the second phase. This design greatly reduces the cost and other practical burdens associated with the collection of expensive covariate data and thus has been widely used in large epidemiological studies.

One recent example of the two-phase design is the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP), where 4494 subjects from seven cohorts were selected for whole-exome sequencing (Lin et al. 2013). Among these subjects, 659, 806, and 657 were selected because of extremely high or low values of body mass index (BMI), blood pressure (BP) adjusted for age, gender, race, BMI, and anti-hypertensive medication, and low-density lipoprotein (LDL) adjusted for age, gender, race, and lipid medication, respectively.

Several methods have been developed for regression analysis of two-phase studies. Semiparametric methods, which specify a parametric form for the regression model but allow for an arbitrary covariate distribution, are particularly appealing. In particular, Robins et al. (1995) proposed a semiparametric estimator based on inverse probability of inclusion weighting. Their approach requires every study subject to have a positive probability of being selected in the second phase and thus cannot be applied to the extreme-tail design adopted by the NHLBI ESP. In addition, their estimator can be difficult to implement in practice because it involves numerical solution of an infinite-dimensional integral equation when the outcome of interest is continuous. Lawless et al. (1999) suggested to discretize the continuous first-phase data into a small number of strata and then use the stratum membership to select subjects in the second phase. For subjects not selected in the second phase, only the stratum membership is used in the inference. Breslow et al. (2003) established the asymptotic properties of the corresponding maximum likelihood estimator (MLE). Such data discretization entails a substantial loss of information and may even bias parameter estimation.

To improve efficiency, Chatterjee et al. (2003) proposed a pseudo-score estimator (PSE), and Weaver and Zhou (2005) proposed a maximum estimated likelihood estimator (MELE). Both methods allow the outcome of interest to be continuous but require the inexpensive covariates to be discrete. Chatterjee and Chen (2007) extended the PSE method to allow

for continuous inexpensive covariates in the regression analysis by using kernel smoothing but required the second-phase selection to depend on only discrete covariates. Both the PSE and MELE methods are statistically inefficient. Song et al. (2009) and Lin et al. (2013) considered efficient estimation for two-phase studies without inexpensive covariates. When the inexpensive covariates are available, however, this approach is inefficient because it disregards the inexpensive covariates for subjects not selected in the second phase. More important, this approach may yield biased estimators if the second-phase selection depends on the inexpensive covariates.

In this paper, we study efficient semiparametric estimation for regression models under general two-phase designs such that the sampling in the second phase can depend on the first-phase data in any manner. We allow the outcome variable to be discrete or continuous, and we accommodate inexpensive covariates. Efficient estimation under such general designs has not been pursued previously. We stress the importance of using inexpensive covariates, which are available in virtually all epidemiological studies, to improve the efficiency of the second-phase sampling, control for confounding, and evaluate interactions among the expensive and inexpensive covariates. We allow inexpensive covariates to be continuous and correlated with expensive covariates, and we do not parametrize the distribution of covariates. Dealing with this general situation is very challenging because the likelihood function involves the conditional density functions of expensive covariates given continuous inexpensive covariates. We overcome this difficulty by incorporating sieve approximations (Grenander 1981) of the conditional density functions into the nonparametric likelihood function. We develop a computationally efficient and numerically stable expectation-maximization (EM) algorithm to maximize the sieve likelihood. We establish the consistency, asymptotic normality, and asymptotic efficiency of the resulting estimators through a novel combination of modern empirical process theory and sieve approximation theory. We demonstrate the superiority of the proposed methods over the existing ones through extensive simulation studies. Finally,

we provide applications to the aforementioned NHLBI ESP.

4.2 Methods

4.2.1 Sieve Maximum Likelihood Estimation

Let Y denote the outcome of interest, $\boldsymbol{X}$ denote the vector of expensive covariates that is measured on a fraction of subjects in the study, $\boldsymbol{Z}$ denote the vector of inexpensive covariates that is potentially correlated with $\boldsymbol{X}$, and $\boldsymbol{W}$ denote the vector of inexpensive covariates that is known to be independent of $\boldsymbol{X}$ given $\boldsymbol{Z}$. The data $(Y, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$ are assumed to be generated from the joint distribution $P_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})P(\boldsymbol{X}|\boldsymbol{Z})P(\boldsymbol{Z}, \boldsymbol{W})$, where $P_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$ is a parametric regression model indexed by parameter $\boldsymbol{\theta}$, $P(\boldsymbol{X}|\boldsymbol{Z})$ is the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Z}$, and $P(\boldsymbol{Z}, \boldsymbol{W})$ is the joint distribution of $\boldsymbol{Z}$ and $\boldsymbol{W}$. For linear regression,

$$P_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y - \alpha - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z} - \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{W})^2}{2\sigma^2}\right\},$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\eta}^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$; for logistic regression,

$$P_{\boldsymbol{\theta}}(Y = 1|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = \left[1 + \exp\left\{-(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z} + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{W})\right\}\right]^{-1},$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\eta}^{\mathrm{T}})^{\mathrm{T}}$. The linear predictors can be modified to include the interaction terms among $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$.

Under the two-phase design, (Y, $\boldsymbol{Z}$, $\boldsymbol{W}$) is measured for all n subjects in the first phase, and $\boldsymbol{X}$ is measured for a sub-sample of size n_2 in the second phase. Let R indicate, by the values 1 versus 0, whether the subject is selected for the measurement of $\boldsymbol{X}$ in the second phase. We make the following assumption:

(A.1) The distribution of R depends on $(Y, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$ only through the first-phase data $(Y, \boldsymbol{Z}, \boldsymbol{W})$.

Under Assumption (A.1), the data on $\boldsymbol{X}$ are missing at random, such that the sampling indicators $(R_1, \ldots, R_n)$ can be omitted from the likelihood function when estimating $\boldsymbol{\theta}$. Thus, the observed-data log-likelihood takes the form

$$\sum_{i=1}^{n} R_i \left\{ \log P_{\boldsymbol{\theta}}(Y_i | \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) + \log P(\boldsymbol{X}_i | \boldsymbol{Z}_i) \right\}$$
$$+ \sum_{i=1}^{n} (1 - R_i) \log \int P_{\boldsymbol{\theta}}(Y_i | \boldsymbol{x}, \boldsymbol{Z}_i, \boldsymbol{W}_i) P(\boldsymbol{x} | \boldsymbol{Z}_i) d\boldsymbol{x}. \tag{4.23}$$

We maximize expression (4.23) using the nonparametric maximum likelihood estimation (NPMLE). For each distinct observed $\boldsymbol{z}$, we estimate $P(\boldsymbol{X}|\boldsymbol{z})$ by a discrete probability function on the distinct observed values of $\boldsymbol{X}$, denoted by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ $(m \leq n_2)$. Even with this discretization, maximization of expression (4.23) is not feasible when $\boldsymbol{Z}$ contains continuous components because then only a small number of the observations on $\boldsymbol{X}$ are associated with each distinct observed $\boldsymbol{z}$.

To tackle this challenge, we approximate $P(\boldsymbol{X}|\boldsymbol{z})$ by the method of sieves (Grenander 1981). Specifically, we use the B-spline basis (Schumaker 1981) to construct the approximating functions. Assuming that $\boldsymbol{Z}$ has bounded support, we center and rescale each component of $\boldsymbol{Z}$ such that it has support on $[0, 1]$. We then partition the interval $[0, 1]$ as $\Delta \equiv \{t_{-q+1} = \ldots = t_{-1} = 0 = t_0 < t_1 < \ldots < t_{b_n+1} = 1 = \ldots = t_{q+b_n}\}$, where $\{t_l \colon l = -q+1, \ldots, q+b_n\}$ are the knots, q is the order of the B-spline basis, and b_n is the number of interior knots. The number b_n is determined by the first-phase sample size n. For ease of implementation, we choose the interior knots as evenly spaced partitions in $[0, 1]$ with gap $1/(b_n + 1)$. Let $\{N_l^q(z)\}_{l=-q+1}^{b_n}$ be a one-dimensional normalized B-spline basis of order q associated with Δ. We construct $N_l^q(z)$ from the recursive formula

$$N_l^q(z) = \frac{z - t_l}{t_{l+q-1} - t_l} N_l^{q-1}(z) + \frac{t_{l+q} - z}{t_{l+q} - t_{l+1}} N_{l+1}^{q-1}(z), \quad l = -q+1, \ldots, b_n,$$

where $N_l^1(z) = I(t_l \le z \le t_{l+1})$, $l = 0, \ldots, b_n$. We refer to $\{N_l^1(z)\}_{l=0}^{b_n}$ as the histogram basis. We then construct the multivariate B-spline basis on the support of $\boldsymbol{Z}$ as

$$\{B_{\boldsymbol{l}}^q(\boldsymbol{Z})\colon\ B_{\boldsymbol{l}}^q(\boldsymbol{Z}) = N_{l_1}^q(Z_1)\cdots N_{l_{d_{\boldsymbol{z}}}}^q(Z_{d_{\boldsymbol{z}}}),\ \boldsymbol{l} = (l_1, \ldots, l_{d_{\boldsymbol{z}}})^{\mathrm{T}},\ l_1, \ldots, l_{d_{\boldsymbol{z}}} = -q+1, \ldots, b_n\},$$

where Z_v is the vth component of $\boldsymbol{Z}$, and $d_{\boldsymbol{z}}$ is the dimension of $\boldsymbol{Z}$. To simplify notation, we order the $(b_n+q)^{d_{\boldsymbol{z}}}$ multivariate basis functions as $B^q_{(-q+1,\ldots,-q+1)}(\boldsymbol{Z})$, $\ldots$, $B^q_{(b_n,\ldots,b_n)}(\boldsymbol{Z})$ and then re-index them with $j = 1, \ldots, (b_n+q)^{d_{\boldsymbol{z}}}$. Because the B-spline basis functions have local support, we approximate $\log P(\boldsymbol{X}_i|\boldsymbol{Z}_i)$ and $P(\boldsymbol{x}|\boldsymbol{Z}_i)$ in expression (4.23) by $\sum_{k=1}^m I(\boldsymbol{X}_i = \boldsymbol{x}_k)\sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i)\log p_{kj}$ and $\sum_{k=1}^m I(\boldsymbol{x} = \boldsymbol{x}_k)\sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i)p_{kj}$, respectively, where $s_n = (b_n + q)^{d_{\boldsymbol{z}}}$, and $p_{kj} = s_n \int P(\boldsymbol{x}_k|\boldsymbol{z})B_j^q(\boldsymbol{z})d\boldsymbol{z}$.

We aim to maximize the following function

$$\begin{aligned} l_n(\boldsymbol{\theta}, \{p_{kj}\}) = &\sum_{i=1}^n R_i \left\{ \log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) + \sum_{k=1}^m \sum_{j=1}^{s_n} I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{Z}_i)\log p_{kj} \right\} \\ &+ \sum_{i=1}^n (1-R_i)\log\left\{ \sum_{k=1}^m P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_k, \boldsymbol{Z}_i, \boldsymbol{W}_i) \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i)p_{kj} \right\} \end{aligned} \tag{4.24}$$

under the constraints of $\sum_{k=1}^m p_{kj} = 1$ and $p_{kj} \ge 0$ $(k = 1, \ldots, m;\ j = 1, \ldots, s_n)$. With the use of the empirical distribution function of $\boldsymbol{X}$ given $\boldsymbol{Z}$, parameter estimation based on the maximization of expression (4.24) is feasible even when $\boldsymbol{X}$ is multidimensional.

Remark 4.1. If there are no inexpensive covariates $\boldsymbol{Z}$ and $\boldsymbol{W}$, then the observed-data log-likelihood (4.23) reduces to

$$\sum_{i=1}^n R_i \left\{\log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i) + \log P(\boldsymbol{X}_i)\right\} + \sum_{i=1}^n (1-R_i)\log \int P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x})P(\boldsymbol{x})d\boldsymbol{x}. \tag{4.25}$$

Song et al. (2009) and Lin et al. (2013) maximized expression (4.25) using the NPMLE, where $P(\boldsymbol{X})$ is estimated by the discrete probabilities at the observed values of $\boldsymbol{X}$. This MLE approach, denoted by MLE_0 hereafter, can be regarded as a special case of our proposed

SMLE approach. If the inexpensive covariates are available for all subjects but the second-phase selection does not depend on either $\boldsymbol{Z}$ or $\boldsymbol{W}$, then the MLE_0 method can be adopted by redefining the "expensive covariates" as $(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}}, \boldsymbol{W}^{\mathrm{T}})^{\mathrm{T}}$ and disregarding $\boldsymbol{Z}$ and $\boldsymbol{W}$ for subjects not selected in the second phase. This data reduction approach may entail a substantial loss of information. If the second-phase selection does depend on $\boldsymbol{Z}$ and $\boldsymbol{W}$, then expression (4.25) no longer correctly reflects the sampling mechanism, and the MLE_0 method is generally biased.

4.2.2 EM Algorithm

Direct maximization of expression (4.24) is difficult due to the intractable form of the second term. To make the problem more tractable, we artificially create a latent variable U for subjects with $R = 0$ such that U takes values on $1/s_n, \ldots, 1$ and satisfies the equations $P(U = j/s_n|\boldsymbol{Z}, \boldsymbol{W}) = B_j^q(\boldsymbol{Z})$, $P(\boldsymbol{X} = \boldsymbol{x}_k|\boldsymbol{Z}, \boldsymbol{W}, U = j/s_n) = P(\boldsymbol{X} = \boldsymbol{x}_k|U = j/s_n) = p_{kj}$, and $P(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}, U) = P(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$. Consequently, $P(\boldsymbol{X} = \boldsymbol{x}_k|\boldsymbol{Z}) = \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}) p_{kj}$ for subjects with $R = 0$, and the second term in expression (4.24) is equivalent to the log-likelihood of $(Y_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)$, assuming that the complete data consist of $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i, U_i)$ but with both $\boldsymbol{X}_i$ and U_i missing.

We devise an EM-type algorithm to maximize expression (4.24) by treating $(\boldsymbol{X}, U)$ for subjects with $R = 0$ as missing. The complete-data log-likelihood is

$$\begin{aligned}
&\sum_{i=1}^{n} R_i \left\{ \log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) + \sum_{k=1}^{m} \sum_{j=1}^{s_n} I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{Z}_i) \log p_{kj} \right\} \\
&+ \sum_{i=1}^{n} (1 - R_i) \sum_{k=1}^{m} I(\boldsymbol{X}_i = \boldsymbol{x}_k) \log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_k, \boldsymbol{Z}_i, \boldsymbol{W}_i) \\
&+ \sum_{i=1}^{n} (1 - R_i) \sum_{k=1}^{m} \sum_{j=1}^{s_n} I(\boldsymbol{X}_i = \boldsymbol{x}_k, U_i = j/s_n) \log p_{kj}.
\end{aligned}$$

In the E-step, we calculate the conditional expectations of $I(\boldsymbol{X}_i = \boldsymbol{x}_k)$ and $I(\boldsymbol{X}_i = \boldsymbol{x}_k, U_i =$

j/s_n) given the observed data for the ith subject with $R_i = 0$ as

$$\widehat{q}_{ik} = \frac{P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_k, \boldsymbol{Z}_i, \boldsymbol{W}_i) \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i) p_{kj}}{\sum_{k'=1}^{m} P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_{k'}, \boldsymbol{Z}_i, \boldsymbol{W}_i) \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i) p_{k'j}}, \quad k = 1, \ldots, m,$$

and

$$\widehat{\psi}_{kji} = \frac{B_j^q(\boldsymbol{Z}_i) p_{kj}}{\sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i) p_{k'j}} \widehat{q}_{ik}, \quad k = 1, \ldots, m, \ j = 1, \ldots, s_n,$$

respectively. In the M-step, we update $\boldsymbol{\theta}$ by maximizing

$$\sum_{i=1}^{n} R_i \log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i) + \sum_{i=1}^{n} (1 - R_i) \sum_{k=1}^{m} \widehat{q}_{ik} \log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_k, \boldsymbol{Z}_i, \boldsymbol{W}_i). \tag{4.26}$$

Expression (4.26) is a weighted sum of the log-likelihood functions for the regression model $P_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$. Thus, we can use existing algorithms for weighted regression to maximize expression (4.26). We update p_{kj} ($k = 1, \ldots, m$; $j = 1, \ldots, s_n$) by maximizing

$$\sum_{i=1}^{n} R_i \sum_{k=1}^{m} \sum_{j=1}^{s_n} I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{Z}_i) \log p_{kj} + \sum_{i=1}^{n} (1 - R_i) \sum_{k=1}^{m} \sum_{j=1}^{s_n} \widehat{\psi}_{kji} \log p_{kj}$$

such that

$$p_{kj} = \frac{\sum_{i=1}^{n} \left\{ R_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{Z}_i) + (1 - R_i) \widehat{\psi}_{kji} \right\}}{\sum_{k=1}^{m} \sum_{i=1}^{n} \left\{ R_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{Z}_i) + (1 - R_i) \widehat{\psi}_{kji} \right\}}.$$

We start with initial values $\widehat{\alpha}^{(0)} = 0$, $\widehat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, $\widehat{\boldsymbol{\gamma}}^{(0)} = \mathbf{0}$, $\widehat{\boldsymbol{\eta}}^{(0)} = \mathbf{0}$, $\widehat{\sigma^2}^{(0)}$ being the sample variance of Y (in linear regression), and $\widehat{p}_{kj}^{(0)} = \sum_{i=1}^{n} R_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{Z}_i) / \sum_{i=1}^{n} R_i$ $B_j^q(\boldsymbol{Z}_i)$, and we iterate until convergence to obtain the sieve maximum likelihood estimators (SMLEs) $\widehat{\boldsymbol{\theta}}$ and $\widehat{p}_{kj}$ ($k = 1, \ldots, m$; $j = 1, \ldots, s_n$). Because the MLE for the distribution function of $\boldsymbol{Z}$ is the empirical distribution function based on $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, the joint distribution

function of $(\boldsymbol{X}, \boldsymbol{Z})$, denoted by $F(\cdot, \cdot)$, can be estimated by

$$\widehat{F}(\boldsymbol{x}, \boldsymbol{z}) = n^{-1} \sum_{k=1}^{m} \sum_{i=1}^{n} I(\boldsymbol{x}_k \leq \boldsymbol{x}, \boldsymbol{Z}_i \leq \boldsymbol{z}) \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i) \widehat{p}_{kj}.$$

Remark 4.2. When Z is a scalar, we can use the histogram basis $\{B_j^1(z)\}_{j=1}^{b_n+1}$ to estimate $P(\boldsymbol{X}|Z)$ (see Section 4.2.3). In this case, the artificial latent variable U is not needed, and the EM algorithm can be greatly simplified. The complete-data log-likelihood becomes

$$\begin{aligned} &\sum_{i=1}^{n} R_i \left\{ \log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i, Z_i, \boldsymbol{W}_i) + \sum_{k=1}^{m} \sum_{j=1}^{s_n} I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^1(Z_i) \log p_{kj} \right\} \\ &+ \sum_{i=1}^{n} (1 - R_i) \sum_{k=1}^{m} I(\boldsymbol{X}_i = \boldsymbol{x}_k) \left\{ \log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_k, Z_i, \boldsymbol{W}_i) + \sum_{j=1}^{s_n} B_j^1(Z_i) \log p_{kj} \right\}. \end{aligned}$$

Consequently, in the E-step, we only need to calculate $\widehat{q}_{ik}$ for the ith subject with $R_i = 0$ as

$$\widehat{q}_{ik} = \sum_{j=1}^{s_n} B_j^1(Z_i) \frac{P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_k, Z_i, \boldsymbol{W}_i) p_{kj}}{\sum_{k'=1}^{m} P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_{k'}, Z_i, \boldsymbol{W}_i) p_{k'j}}, \quad k = 1, \ldots, m.$$

In the M-step, we update $\boldsymbol{\theta}$ by maximizing expression (4.26) and update p_{kj} ($k = 1, \ldots, m$; $j = 1, \ldots, s_n$) by the following simple formula

$$p_{kj} = \frac{\sum_{i=1}^{n} \left\{ R_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^1(Z_i) + (1 - R_i) B_j^1(Z_i) \widehat{q}_{ik} \right\}}{\sum_{i=1}^{n} B_j^1(Z_i)}.$$

4.2.3 Asymptotic Properties

Let $\boldsymbol{\Theta}$ denote the parameter space of $\boldsymbol{\theta}$, which is a bounded open set in the interior of the domain of $\boldsymbol{\theta}$, and let $\mathcal{F}$ denote the space of the joint distributions of $(\boldsymbol{X}, \boldsymbol{Z})$. Let $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ and $F_0 \in \mathcal{F}$ denote the true values of $\boldsymbol{\theta}$ and F, respectively. We impose the following regularity conditions:

(C.1) The set of covariates $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$ has bounded support.

(C.2) If there exist two sets of parameters $(\boldsymbol{\theta}_1, F_1)$ and $(\boldsymbol{\theta}_2, F_2)$ such that

$$P_{\boldsymbol{\theta}_1}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})F_1(\boldsymbol{X}, \boldsymbol{Z}) = P_{\boldsymbol{\theta}_2}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})F_2(\boldsymbol{X}, \boldsymbol{Z}),$$

where $(Y, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) \in \mathcal{C} \equiv \{(y, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$: $P(R = 1|y, \boldsymbol{z}, \boldsymbol{w}) \geq q_0\}$, and q_0 is a positive constant, then $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ and $F_1 = F_2$. In addition, if there exists a constant vector $\boldsymbol{v}$ such that $[\partial \log\{P_{\boldsymbol{\theta}_0}(y_1|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}_1)/P_{\boldsymbol{\theta}_0}(y_2|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}_2)\}/\partial\boldsymbol{\theta}]^{\mathrm{T}}\boldsymbol{v} = 0$ for any $(y_i, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}_i) \in \mathcal{C}$, $i = 1, 2$, then $\boldsymbol{v} = \boldsymbol{0}$.

(C.3) The density function of F_0 is positive in its support and q-times continuously differentiable with respect to a suitable measure.

(C.4) The function $\mathrm{E}(R|\boldsymbol{X}, \boldsymbol{Z})$ is q-times continuously differentiable with respect to $\boldsymbol{X}$ and $\boldsymbol{Z}$.

(C.5) As $n \to \infty$, $s_n \to \infty$, and $n^{1/2}s_n^{-q/d_{\boldsymbol{z}}} \to 0$.

Remark 4.3. The first part of Condition (C.2) pertains to model identifiability with complete data. For many commonly used regression models, the set $\mathcal{C}$, where $P(R = 1|y, \boldsymbol{z}, \boldsymbol{w}) \geq q_0$, does not necessarily need to cover the entire support of $(Y, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$. For example, in linear regression, $\mathcal{C}$ can consist of data points with extremely large or small values of Y only. The second part of Condition (C.2) ensures that the score functions for $\boldsymbol{\theta}$ are of full rank on $\mathcal{C}$. For linear regression, this condition follows from the linear independence of the covariates $(1, \boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}}, \boldsymbol{W}^{\mathrm{T}})^{\mathrm{T}}$. Condition (C.3) pertains to the smoothness of the joint distribution function of $(\boldsymbol{X}, \boldsymbol{Z})$. Condition (C.4) holds for all commonly used two-phase designs, including the extreme-tail design adopted by the NHLBI ESP. Under Condition (C.5), the order of the B-spline basis q should be greater than $d_{\boldsymbol{z}}/2$. Consequently, when $d_{\boldsymbol{z}} = 1$, we can choose $q = 1$ and use the histogram basis $\{B_j^1(z)\}_{j=1}^{b_n+1}$ to estimate $P(\boldsymbol{X}|Z)$.

We state the asymptotic results in two theorems, whose proofs are given in the Appendix.

Theorem 4.1. Under Assumption (A.1) and Conditions (C.1)–(C.5), $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| + \sup_{\boldsymbol{x},\boldsymbol{z}}|\widehat{F}(\boldsymbol{x}, \boldsymbol{z}) - F_0(\boldsymbol{x}, \boldsymbol{z})| \to 0$ almost surely.

Theorem 4.2. Under Assumption (A.1) and Conditions (C.1)–(C.5), $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

The profile log-likelihood function for $\boldsymbol{\theta}$ is $pl(\boldsymbol{\theta}) \equiv \max_{\{p_{kj}\}} l_n(\boldsymbol{\theta}, \{p_{kj}\})$. As justified at the end of the Appendix, we can estimate the limiting covariance matrix of $\widehat{\boldsymbol{\theta}}$ by the negative inverse of the Hessian matrix of $pl(\widehat{\boldsymbol{\theta}})$. Specifically, we obtain the value of $pl(\boldsymbol{\theta})$ by holding $\boldsymbol{\theta}$ fixed in the EM algorithm and obtaining the value of $l_n(\boldsymbol{\theta}, \{p_{kj}\})$ at convergence. Then, we estimate the covariance matrix of $\boldsymbol{\theta}$ by the negative inverse of the matrix whose (k,l)th element is $h_n^{-2}\left\{pf(\widehat{\boldsymbol{\theta}} + \boldsymbol{e}_k h_n + \boldsymbol{e}_l h_n) - pf(\widehat{\boldsymbol{\theta}} + \boldsymbol{e}_k h_n) - pf(\widehat{\boldsymbol{\theta}} + \boldsymbol{e}_l h_n) + pf(\widehat{\boldsymbol{\theta}})\right\}$, where $\boldsymbol{e}_k$ is the kth canonical vector, and h_n is a constant of the order $n^{-1/2}$.

4.3 Simulation Studies

We conducted extensive simulation studies to compare the performance of the SMLE and MLE_0 methods in realistic situations. In the first set of studies, we set $X = U_1$, $Z = rU_1 + U_2$, and $W = U_3$, where U_1, U_2, and U_3 are independent Uniform(0,1) variables, and r is a parameter controlling the correlation between X and Z. We generated the outcome from the linear model: $Y = 0.5X + 0.5Z + 0.5W + \epsilon$, where ϵ is a standard normal random variable independent of U_1, U_2, and U_3. We let $n = 2000$ and selected 150 subjects with the highest and 150 subjects with the lowest values of Y in the second phase. For the subjects selected in the second phase, the data consist of (Y, X, Z, W); for those not selected in the second phase, the data utilized by the SMLE and MLE_0 methods consist of (Y, Z, W) and Y, respectively. In the SMLE method, we estimated $P(X|Z)$ using the histogram basis. We partitioned the domain of Z using evenly-spaced quantiles and varied the number of regions s_n from 5 to 15 to assess its effects on model-fitting. The results with different s_n are very similar. The maximum difference in the coverage probability of the 95% confidence interval for any parameter is only 0.5%. Therefore, we only report the results for $s_n = 10$. We

estimated the covariance matrix of $\widehat{\boldsymbol{\theta}}$ by the profile likelihood method with step size of $n^{-1/2}$.

The results of the simulation studies are summarized in Table 4.24. Both the SMLE and MLE_0 parameter estimators are virtually unbiased. The SMLE variance estimator accurately reflects the true variation, and the corresponding confidence intervals have reasonable coverage probabilities. The SMLE estimator is much more efficient than the MLE_0 estimator for Z and W because the SMLE method utilizes additional data on (Z, W) for those subjects not selected in the second phase. The SMLE estimator is also more efficient than the MLE_0 estimator for X, and the efficiency gain increases as the correlation between X and Z increases.

Table 4.24: Simulation Results Under the Model $Y = 0.5X + 0.5Z + 0.5W + \epsilon$ With the Second-Phase Sample Selection Depending Only on Y

		SMLE					MLE_0	
r	Covariate	Bias	SE	SEE	CP	RE	Bias	SE
0.0	X	0.004	0.112	0.108	0.943	1.029	0.005	0.114
	Z	0.001	0.082	0.083	0.951	1.923	0.006	0.114
	W	−0.001	0.078	0.078	0.952	2.126	0.005	0.114
0.1	X	0.005	0.112	0.109	0.941	1.036	0.004	0.114
	Z	0.004	0.081	0.082	0.951	1.973	0.006	0.114
	W	−0.001	0.078	0.078	0.952	2.153	0.005	0.115
0.2	X	0.005	0.112	0.109	0.945	1.077	0.004	0.116
	Z	0.005	0.081	0.082	0.952	2.029	0.006	0.115
	W	−0.001	0.078	0.078	0.952	2.167	0.005	0.115
0.3	X	0.004	0.114	0.111	0.945	1.104	0.005	0.119
	Z	0.005	0.081	0.082	0.952	2.056	0.006	0.116
	W	−0.001	0.078	0.078	0.953	2.189	0.005	0.115

NOTE: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; SEE is the empirical mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval; RE is the empirical variance of MLE_0 over that of SMLE. Each entry is based on 10,000 replicates.

In the second set of simulation studies, we generated the data from the model $Y = 0.5X + 0.5Z + 0.5W + 0.4XW + \epsilon$. The results are summarized in Table 4.25. The SMLE

estimator is much more efficient than the MLE_0 estimator for all covariates. In addition, the relative efficiency of the SMLE estimator to the MLE_0 estimator for X is much higher with the interaction term than without the interaction term in the regression model.

Table 4.25: Simulation Results Under the Model $Y = 0.5X + 0.5Z + 0.5W + 0.4XW + \epsilon$ With the Second-Phase Sample Selection Depending Only on Y

		SMLE					MLE_0	
r	Covariate	Bias	SE	SEE	CP	RE	Bias	SE
0.0	X	0.009	0.225	0.214	0.935	1.207	0.010	0.248
	Z	0.001	0.087	0.087	0.950	1.885	0.008	0.120
	W	0.005	0.208	0.199	0.941	1.400	0.011	0.246
	XW	−0.008	0.388	0.374	0.941	1.275	0.001	0.438
0.1	X	0.012	0.224	0.213	0.934	1.244	0.011	0.250
	Z	0.006	0.086	0.086	0.951	1.972	0.007	0.121
	W	0.005	0.206	0.199	0.944	1.454	0.012	0.248
	XW	−0.009	0.385	0.372	0.940	1.321	0.000	0.443
0.2	X	0.011	0.220	0.211	0.935	1.316	0.011	0.252
	Z	0.007	0.084	0.085	0.949	2.082	0.007	0.122
	W	0.005	0.202	0.196	0.943	1.535	0.012	0.251
	XW	−0.009	0.377	0.365	0.941	1.396	0.000	0.446
0.3	X	0.009	0.218	0.209	0.938	1.387	0.011	0.256
	Z	0.007	0.083	0.084	0.950	2.147	0.007	0.122
	W	0.004	0.199	0.192	0.944	1.635	0.012	0.254
	XW	−0.008	0.369	0.358	0.942	1.494	0.000	0.451

NOTE: See the Note to Table 1.

In the above two sets of simulation studies, the second-phase selection depends on the outcome only such that MLE_0 provides unbiased estimation of all parameters. If the second-phase selection depends on both the outcome and inexpensive covariates, then MLE_0 may be biased, whereas PSE (Chatterjee et al. 2003, Chatterjee and Chen 2007) can still be adopted provided that the sampling depends on only discrete covariates. In a third set of simulations, we compared the SMLE, MLE_0, and PSE methods in this scenario. Specifically, we set $X = I(U_1 > 0.8)$ and $Z = I(\widetilde{Z} > \widetilde{z}_{0.8})$, where $\widetilde{Z} = rX + U_2$, r is a parameter controlling the correlation between X and Z, U_1 and U_2 are independent Uniform(0,1), and

$\widetilde{z}_{0.8}$ is the 80% quantile of $\widetilde{Z}$. We generated the outcome from the model $Y = X + Z + \epsilon$, where ϵ is a standard normal random variable independent of U_1 and U_2. In the first phase, we simulated a cohort of 4000 subjects and defined six strata according to the values of Z and Y. That is, for subjects with $Z = 0$, we defined three strata according to whether their values of Y are less than the 5% quantile, greater than the 95% quantile, or between these two quantiles; for subjects with $Z = 1$, we defined another three strata according to whether their values of Y are less than the 20% quantile, greater than the 80% quantile, or between these two quantiles. The quantiles were chosen such that each of the extreme-tail strata contained ∼160 subjects. In the second phase, we only included subjects with values of Y in the four extreme-tail strata such that $n_2 \approx 640$. Because Z is binary, for the SMLE method we estimated $P(X|Z)$ by the empirical probability of X given Z. As shown in Table 4.26, the SMLE method is much more efficient than the PSE method, and the efficiency gain increases as the correlation between X and Z decreases. The MLE_0 parameter estimators are severely biased whether X and Z are correlated or not.

Table 4.26: Simulation Results When the Second-Phase Sample Selection Depends on Both Y and Z

		SMLE					MLE_0		PSE	
r	Covariate	Bias	SE	SEE	CP	RE	Bias	SE	Bias	SE
0.0	X	0.005	0.074	0.073	0.952	1.307	0.291	0.096	0.006	0.085
	Z	0.000	0.047	0.047	0.947	1.146	−0.499	0.044	0.000	0.051
0.1	X	0.004	0.070	0.070	0.952	1.220	0.267	0.093	0.004	0.078
	Z	0.000	0.049	0.049	0.945	1.123	−0.556	0.041	0.001	0.052
0.2	X	0.003	0.067	0.067	0.952	1.154	0.254	0.090	0.002	0.072
	Z	0.000	0.052	0.051	0.944	1.106	−0.609	0.039	0.000	0.055
0.3	X	0.003	0.066	0.066	0.950	1.118	0.241	0.089	0.002	0.070
	Z	0.000	0.056	0.055	0.945	1.092	−0.658	0.038	0.000	0.059

NOTE: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; SEE is the empirical mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval; RE is the empirical variance of PSE over that of SMLE. Each entry is based on 10,000 replicates.

4.4 NHLBI ESP

The NHLBI ESP was designed to identify genetic variants in all protein-coding regions of the human genome that are associated with heart, lung, and blood diseases. It involves seven cohorts: the Atherosclerosis Risk in Communities (ARIC) study (The ARIC Investigators 1989); Coronary Artery Risk Development in Young Adults (CARDIA) study (Friedman et al. 1988); Cardiovascular Health Study (CHS) (Fried et al. 1991); Framingham Heart Study (FHS) (Dawber et al. 1951); Jackson Heart Study (Taylor Jr et al. 2005); Multi-Ethnic Study of Atherosclerosis (MESA) (Bild et al. 2002); and Women's Health Initiative (WHI) (The Women's Health Initiative Study Group 1998). As mentioned in Section 1, the NHLBI ESP consisted of multiple studies, some of which employed two-phase designs. Exome sequencing was performed on the selected subjects at the University of Washington and the Broad Institute. Details for the design, sample selection criteria, genotype quality control, and annotation can be found in Lin et al. (2013). We provide applications to the BP and LDL studies in the NHLBI ESP.

4.4.1 BP Study

We considered the BP study in the NHLBI ESP. The first phase was comprised of 28,202 subjects from the ARIC, CARDIA, CHS, FHS, JHS, and MESA cohorts. In the second phase, 253 and 245 subjects from the upper and lower tails of the BP distribution, respectively, were selected by the NHLBI ESP investigators for sequencing. The selection was not based on the original BP values, but rather the average residuals from the linear models relating diastolic and systolic BP values to age, gender, race, BMI, and anti-hypertensive medication. In addition to the 498 subjects selected from the two tails of the BP distribution, the second-phase sample also included 410 subjects from the deeply phenotyped reference (DPR) group, which is a random sample of subjects with measurements on a common set of phenotypes.

Because the original BP values were not available for those subjects without the sequence

data, we considered the average BP residuals as the outcome of interest in the analysis. We included log-transformed BMI, race, age, age-squared, and cohort indicators as covariates. Although BMI and race are not correlated with the BP residuals, they are potentially correlated with single-nucleotide polymorphism (SNP) genotypes and thus may provide information on SNP genotypes for those subjects without the sequence data. When implementing the SMLE method, we let $\boldsymbol{Z}$ include log-transformed BMI and race and $\boldsymbol{W}$ include the other covariates. In the sieve approximation, we used the histogram basis because $\boldsymbol{Z}$ contains only one continuous component (i.e., log-transformed BMI). We partitioned the domain of BMI using separate evenly-spaced quantiles for the European Americans (EAs) and African Americans (AAs). In genome-wide association studies, a well-behaved quantile-quantile (QQ) plot and a close-to-one genomic control λ, which is defined as the ratio between the observed median of the test statistics and the median of the χ^2_1 distribution, would imply good model fitting and proper type I error control. We used the QQ plot and genomic control λ to select the number of regions; this resulted in three regions for the EAs and one region for the AAs (Figure 4.10).

We restricted our analysis to the 31,009 SNPs with minor allele frequencies (MAFs) greater than 10%. We chose the additive genetic model, under which the genetic variable codes the number of minor alleles that an subject carries at a variant site. Figure 4.11 shows the QQ plots for the SMLE and MLE_0 methods. Because the second-phase selection is solely determined by the outcome of interest, the MLE_0 method is valid. The SMLE method produces more significant results than the MLE_0 method. Table 4.27 lists the top 10 SNPs for the SMLE method. The genetic effect estimates are similar between the two methods. Correlations between log-transformed BMI and the SNP genotypes are weak. When the SNP genotypes are weakly correlated with race, the standard error estimates of the SMLE method are comparable to those of the MLE_0 method; when the SNP genotypes are strongly correlated with race, the standard error estimates of the SMLE method are much smaller

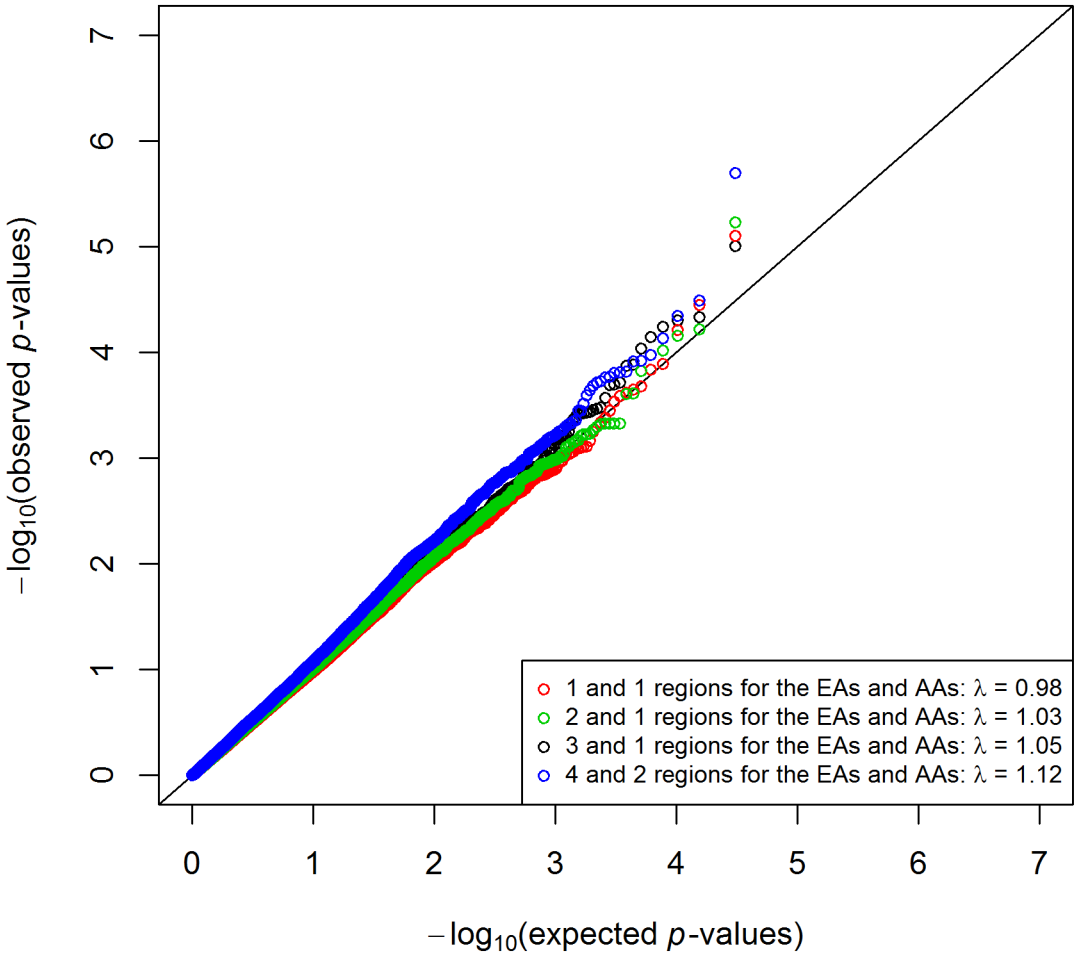

Figure 4.10: Quantile-quantile plots for the analysis of the BP study in the NHLBI ESP using the SMLE method with different numbers of sieve regions.

than those of the MLE_0 method. These results are consistent with the theoretical and simulation results.

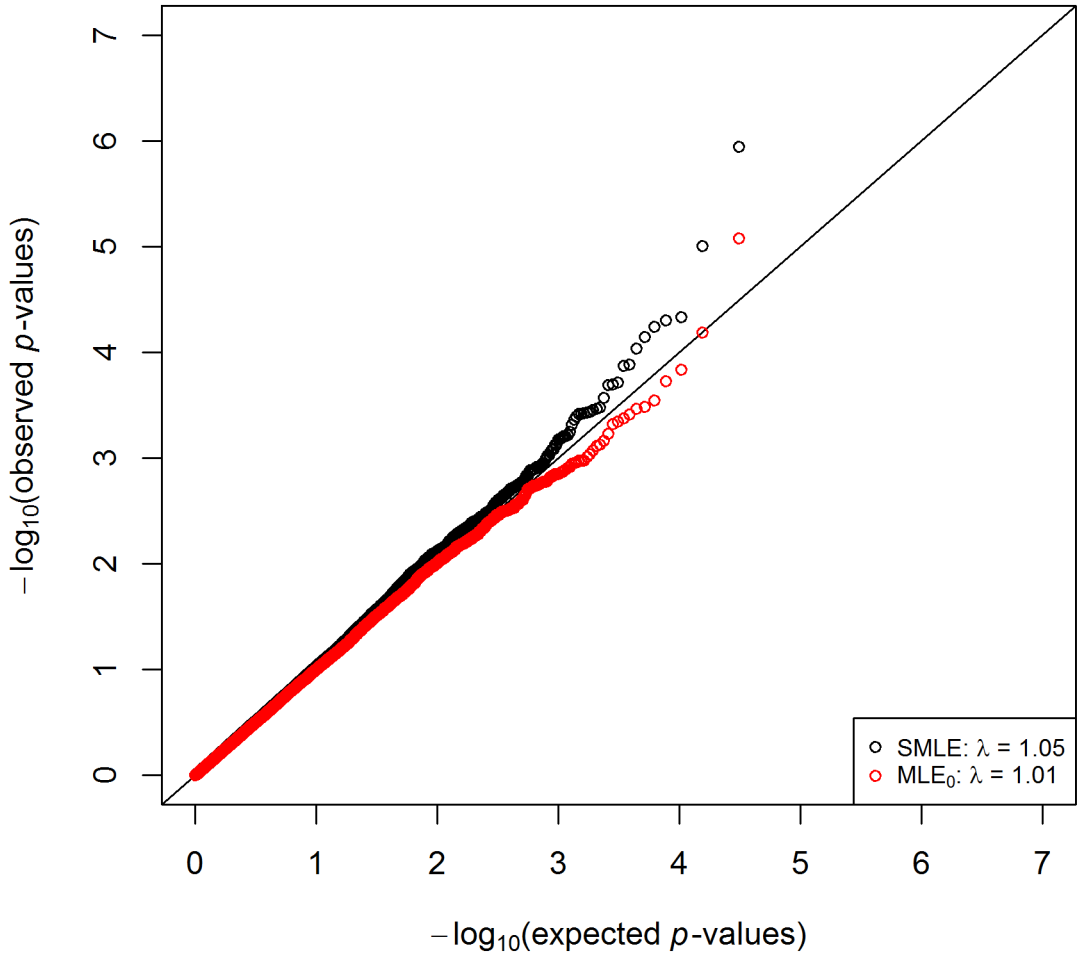

Figure 4.11: Quantile-quantile plots for the analysis of the BP study in the NHLBI ESP using the SMLE and MLE_0 methods.

4.4.2 LDL Study

We next considered the LDL study in the NHLBI ESP. The first phase was comprised of 49,904 subjects from the aforementioned seven cohorts. In the second phase, 604 subjects with extremely large or small values of the residuals from the linear regression of log-transformed LDL on age, gender, race, and lipid medication and 923 subjects from the DPR group were selected by the NHLBI ESP investigators for sequencing. We considered

Table 4.27: Top 10 SNPs in the Analysis of the BP Study in the NHLBI ESP

		Correlation		SMLE			MLE_0		
SNP	MAF	log (BMI)	Race	Est	SE	p-value	Est	SE	p-value
19:001061910	0.12	0.13	0.68	3.37E-01	6.92E-02	1.14E-06	5.06E-01	2.21E-01	2.20E-02
18:044595809	0.43	0.00	0.11	2.04E-01	4.63E-02	9.93E-06	2.05E-01	4.61E-02	8.37E-06
18:051904644	0.25	0.00	−0.27	2.29E-01	5.63E-02	4.65E-05	2.14E-01	5.64E-02	1.47E-04
08:017478527	0.14	0.08	0.40	3.18E-01	7.84E-02	4.98E-05	3.21E-01	9.10E-02	4.19E-04
20:033874784	0.11	0.10	0.64	3.33E-01	8.28E-02	5.73E-05	4.18E-01	1.47E-01	4.46E-03
18:051904641	0.25	−0.01	−0.27	2.23E-01	5.62E-02	7.13E-05	2.09E-01	5.59E-02	1.86E-04
07:101713590	0.18	0.01	−0.02	2.58E-01	6.59E-02	9.24E-05	2.17E-01	6.63E-02	1.06E-03
09:019087196	0.12	0.06	0.43	3.80E-01	9.94E-02	1.30E-04	4.22E-01	1.18E-01	3.43E-04
18:044585955	0.38	−0.08	−0.05	1.84E-01	4.81E-02	1.34E-04	1.89E-01	4.73E-02	6.54E-05
19:007166388	0.28	0.09	0.39	2.01E-01	5.39E-02	1.93E-04	1.79E-01	5.55E-02	1.29E-03

NOTE: SNP name is in the "chromosome:position" format, where the positions are based on the human reference sequence (UCSC Genome Browser, hg19). Est and SE stand for the genetic effect estimate and standard error, respectively. Correlation pertains to the SNP and the covariate.

log-transformed LDL as the outcome of interest and included log-transformed BMI, race, age, age-squared, gender, and cohort as covariates. As in Section 4.4.1, when implementing the SMLE method, we let $\boldsymbol{Z}$ include log-transformed BMI and race and $\boldsymbol{W}$ include the other covariates. In the sieve approximation, we used the histogram basis and partitioned the domain of BMI using separate evenly-spaced quantiles for the EAs and AAs. We used the QQ plot and genomic control λ to select the number of regions; this resulted in one region for both EAs and AAs (Figure 4.12).

We restricted our analysis to the 26,431 SNPs with MAFs greater than 15%. We chose the additive genetic model. Figure 4.13 shows the QQ plots using the SMLE and MLE_0 methods. The observed p-values of the SMLE method agree very well with the global null hypothesis of no association, except at the extreme right tail. By contrast, the observed p-values of the MLE_0 method deviate substantially from the null distribution, reflecting excessive false-positive results. This is because the second-phase selection is determined by both the outcome of interest and the inexpensive covariates. Incidentally, the PSE method of Chatterjee and Chen (2007) could not be applied here because it does not allow the second-phase selection to depend on continuous covariates.

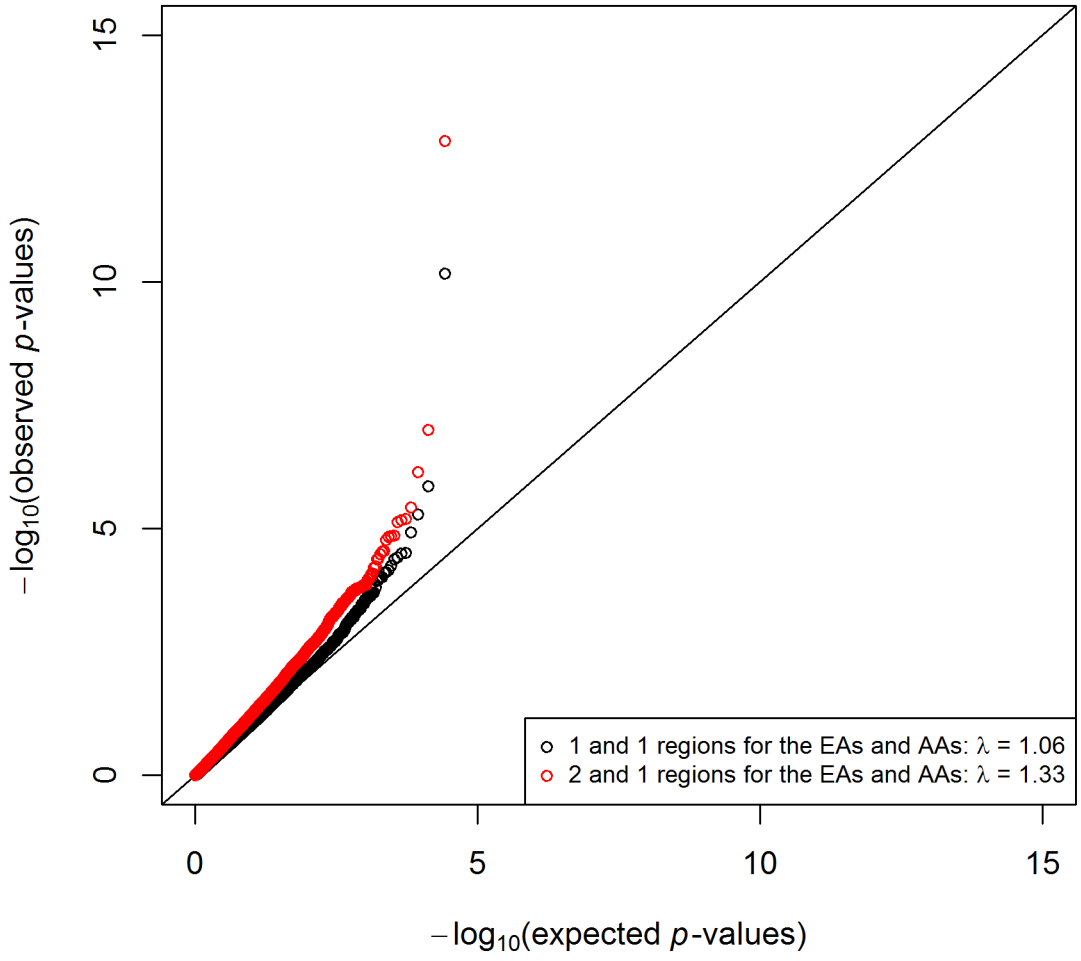

Figure 4.12: Quantile-quantile plots for the analysis of the LDL study in the NHLBI ESP using the SMLE method with different numbers of sieve regions.

4.5 Discussion

We have developed efficient semiparametric inference procedures for general two-phase designs. The likelihood function of interest is not tractable because it involves the conditional density function of expensive covariates given continuous inexpensive covariates. We approximate this conditional density function by the method of sieves. We prove the

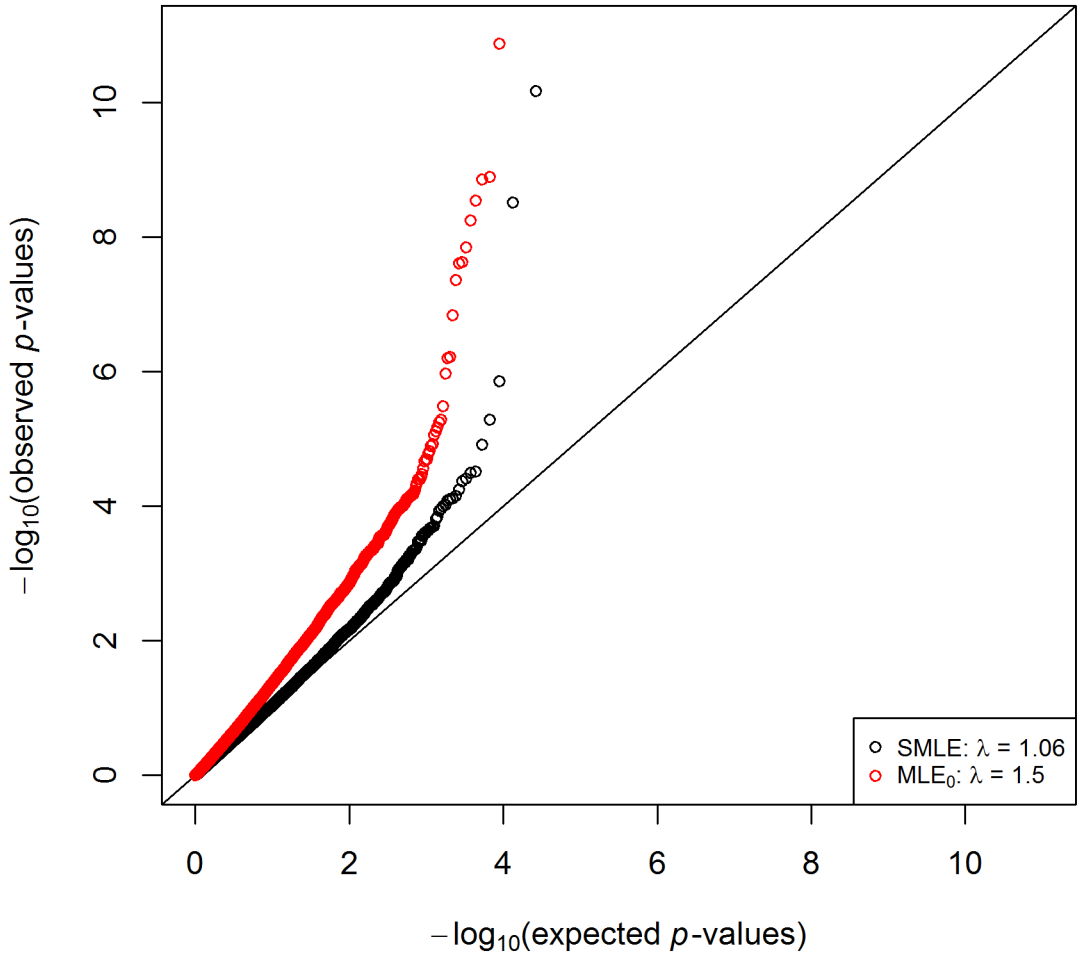

Figure 4.13: Quantile-quantile plots for the analysis of the LDL study in the NHLBI ESP using the SMLE and MLE_0 methods.

asymptotic properties of the proposed estimators through a novel combination of modern empirical process theory and sieve approximation theory. Our framework does not require every study subject to have a positive selection probability in the second phase and thus covers a wide spectrum of two-phase designs. We provide easily-verifiable conditions on model identifiability that rely only on subjects with complete data.

The proposed EM algorithm is numerically stable and computationally efficient. The M-step only involves maximizing the log-likelihood of weighted regression, and the calculations of $\widehat{q}_{ik}$ and $\widehat{\psi}_{kji}$ in the E-step, as well as $\widehat{p}_{kj}$, have explicit formulas. If Z is a scalar, then one can use the histogram basis, such that the algorithm becomes extremely simple. In our analysis of the BP and LDL studies in the NHLBI ESP, it took $\sim$10 seconds on an IBM HS21 machine to perform one association analysis. An R package that implements the proposed method is available on our website.

Lin et al. (2013) analyzed the LDL study in the NHLBI ESP using the MLE_0 method. To avoid the dependence of the second-phase selection on the inexpensive covariates, they used the residuals instead of the original LDL values as the outcome of interest, even though the LDL values were available for all subjects. This workaround is not desirable because the resulting genetic effect estimates are difficult to interpret and not comparable with estimates from studies that use the original LDL values.

In our sieve approximation to $P(\boldsymbol{X}|\boldsymbol{Z})$, the number of interior knots b_n in the domain of $\boldsymbol{Z}$ can be chosen in a data-adaptive manner. One possible approach for choosing b_n is through cross-validation. For any fixed b_n, we use part of the data as the test set and the remainder as the validation set. We evaluate expression (4.24) in the validation set using estimates obtained from the test set. The optimal number of interior knots b_n is the value that maximizes the average cross-validation likelihood. Alternative approaches can also be used to choose b_n. As demonstrated in Section 4, one can use the QQ plot and genomic control λ to choose the appropriate b_n in genetic association studies.

We have assumed that the second-phase selection depends on a single outcome. If the selection depends on multiple outcomes in one study, then one should consider all of them simultaneously in a multivariate regression model in order to obtain valid inference. Recently, Tao et al. (2015) extended the MLE_0 approach to multivariate outcome-dependent sampling

without inexpensive covariates. We can extend our SMLE approach to multivariate outcome-dependent sampling with inexpensive covariates. We simply replace $P_{\boldsymbol{\theta}}(Y|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W})$ in expression (4.24) by the conditional density function $P_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W})$ of the multivariate outcome $\boldsymbol{Y}$ given covariates. If $\boldsymbol{Y}$ contains missing components, then we need to modify the EM algorithm in Section 2.2 by first calculating the conditional expectations of the missing components given the observed data in the E-step and then replacing the missing components with their conditional expectations in the M-step. We expect Theorems 1 and 2 to continue to hold.

In both the simulation studies and NHLBI ESP applications, the outcome of interest is always used in the second-phase sampling process. In practice, investigators may be interested in a secondary outcome that is not used for sampling but is correlated with the primary outcome used for sampling. In light of the above discussion on multivariate outcome-dependent sampling, it is straightforward to analyze the secondary outcome by assuming a bivariate regression model for the primary and secondary outcomes.

Although we have focused on the parametric regression model $P_{\boldsymbol{\theta}}(Y|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W})$, our approach can be extended to semiparametric regression models, particularly those with censored time-to-event outcomes. When $\boldsymbol{\theta}$ contains both Euclidean and infinite-dimensional components, the maximizization of expression (4.26) in the EM algorithm is more involved, and the proof of Theorem 1 requires modification. Recently, Zeng and Lin (2014) considered efficient estimation of semiparametric transformation models for two-phase cohort studies with censored data. They employed a kernel smoothing approach to approximate $P(\boldsymbol{X}|\boldsymbol{Z})$ when $\boldsymbol{Z}$ contains continuous components. Both kernel smoothing and sieve approximation are powerful nonparametric tools for density estimation. We adopted the sieve approximation approach because it is computationally more efficient, especially when the dimension of $\boldsymbol{Z}$ is one.

This paper is focused on the inference procedures rather than the design aspects of

two-phase studies. An important topic of investigation is the optimal study design when the primary interest is to estimate $\boldsymbol{\beta}$. When the outcome is continuous and there is no inexpensive covariate, Lin et al. (2013) showed that the efficient information for estimating β using the MLE_0 method is approximately $\text{Var}(Y|R=1)\text{Var}(X|R=1)/\sigma^4$ (assuming that X is a scalar). This implies that the study design is more efficient if it selects subjects with more extreme values of Y. For general two-phase studies with (possibly multivariate) continuous outcomes of interest, it is unclear what the best sampling strategy is. Because our likelihood framework applies to any two-phase design, the variance estimators for the SMLE method can be used to evaluate the efficiencies of different designs.

4.6 Proofs of Theorems

Proof of Theorem 4.1. Because $\widehat{\boldsymbol{\theta}}$ is bounded and $\widehat{F}(\boldsymbol{x},\boldsymbol{z})$ is a distribution function with bounded support, it follows from Helly's selection theorem that, for any subsequence of $\widehat{\boldsymbol{\theta}}$ and $\widehat{F}(\boldsymbol{x},\boldsymbol{z})$, there exists a further subsequence, still denoted as $\widehat{\boldsymbol{\theta}}$ and $\widehat{F}(\boldsymbol{x},\boldsymbol{z})$, such that $\widehat{\boldsymbol{\theta}}$ converges almost surely to some vector $\boldsymbol{\theta}^*$ and $\widehat{F}(\boldsymbol{x},\boldsymbol{z})$ converges weakly to some function $F^*(\boldsymbol{x},\boldsymbol{z})$. Theorem 1 will hold if we can show that $\boldsymbol{\theta}^*=\boldsymbol{\theta}_0$ and $F^*=F_0$.

Because $\widehat{p}_{kj}$ maximizes expression (4.24), differentiating expression (4.24) with respect to p_{kj} yields

$$\sum_{i=1}^{n} R_i \frac{I(\boldsymbol{X}_i=\boldsymbol{x}_k)B_j^q(\boldsymbol{Z}_i)}{p_{kj}} + \sum_{i=1}^{n}(1-R_i)\frac{P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x}_k,\boldsymbol{Z}_i,\boldsymbol{W}_i)B_j^q(\boldsymbol{Z}_i)}{\sum_{j'=1}^{s_n}\sum_{k'=1}^{m}P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x}_{k'},\boldsymbol{Z}_i,\boldsymbol{W}_i)B_{j'}^q(\boldsymbol{Z}_i)p_{k'j'}} = \widehat{\mu}_j, \tag{4.27}$$

where $\widehat{\mu}_j$ is the Lagrange multiplier for the constraint that $\sum_{k=1}^{m}\widehat{p}_{kj}=1$. By multiplying both sides of equation (4.27) with p_{kj} and then summing over k, we have

$$\widehat{\mu}_j = \sum_{i=1}^{n} R_i B_j^q(\boldsymbol{Z}_i) + \sum_{i=1}^{n}(1-R_i)\frac{\sum_{k'=1}^{m}P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x}_{k'},\boldsymbol{Z}_i,\boldsymbol{W}_i)B_j^q(\boldsymbol{Z}_i)p_{k'j}}{\sum_{j'=1}^{s_n}\sum_{k'=1}^{m}P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x}_{k'},\boldsymbol{Z}_i,\boldsymbol{W}_i)B_{j'}^q(\boldsymbol{Z}_i)p_{k'j'}}.$$

Consequently,

$$\widehat{p}_{kj} = \frac{\sum_{i=1}^{n} R_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{Z}_i)}{\widehat{\mu}_j - \sum_{i=1}^{n}(1-R_i)\frac{P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x}_k,\boldsymbol{Z}_i,\boldsymbol{W}_i)B_j^q(\boldsymbol{Z}_i)}{\sum_{j'=1}^{s_n}\sum_{k'=1}^{m} P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x}_{k'},\boldsymbol{Z}_i,\boldsymbol{W}_i)B_{j'}^q(\boldsymbol{Z}_i)\widehat{p}_{k'j'}}},$$

and

$$\widehat{F}(\boldsymbol{x},\boldsymbol{z}) = n^{-1}\sum_{k=1}^{m}\sum_{i=1}^{n} I(\boldsymbol{x}_k \le \boldsymbol{x}, \boldsymbol{Z}_i \le \boldsymbol{z}) \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i)\widehat{p}_{kj}.$$

It follows that

$$\begin{aligned}
&\widehat{P}(\boldsymbol{X} = \boldsymbol{x}_k|\boldsymbol{z}) = \sum_{j=1}^{s_n} B_j^q(\boldsymbol{z})\widehat{p}_{kj} \\
=&\sum_{j=1}^{s_n} B_j^q(\boldsymbol{z}) \frac{\sum_{i=1}^{n} R_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{Z}_i)}{\sum_{i=1}^{n}\left\{R_i + (1-R_i)\frac{\sum_{k'=1}^{m} P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x}_{k'},\boldsymbol{Z}_i,\boldsymbol{W}_i)\widehat{p}_{k'j} - P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x}_k,\boldsymbol{Z}_i,\boldsymbol{W}_i)}{\sum_{j'=1}^{s_n}\sum_{k'=1}^{m} P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x}_{k'},\boldsymbol{Z}_i,\boldsymbol{W}_i)B_{j'}^q(\boldsymbol{Z}_i)\widehat{p}_{k'j'}}\right\} B_j^q(\boldsymbol{Z}_i)}.
\end{aligned}$$

Because the B-spline basis functions have local support, we have $|B_j^q(\widetilde{\boldsymbol{z}}) - B_j^q(\boldsymbol{z})I(\|\widetilde{\boldsymbol{z}} - \boldsymbol{z}\| \le \xi_n)| \lesssim \xi_n$ for nonzero $B_j^q(\widetilde{\boldsymbol{z}})$ and $B_j^q(\boldsymbol{z})$, $j = 1, \ldots, s_n$, where $\xi_n = (b_n + 1)^{-1}$, and "$\lesssim$" means less than or equal to up to a constant. Thus, the distribution function $\widehat{F}(\boldsymbol{x},\boldsymbol{z})$ is asymptotically equivalent to

$$n^{-1}\sum_{k=1}^{m}\sum_{i=1}^{n} I(\boldsymbol{x}_k \le \boldsymbol{x}, \boldsymbol{Z}_i \le \boldsymbol{z}) \frac{\sum_{j=1}^{s_n}\sum_{i'=1}^{n} R_{i'} I(\boldsymbol{X}_{i'} = \boldsymbol{x}_k, \|\boldsymbol{Z}_{i'} - \boldsymbol{Z}_i\| \le \xi_n) B_j^q(\boldsymbol{Z}_i)}{g_{1n}(\boldsymbol{x}_k, \boldsymbol{Z}_i; \widehat{\boldsymbol{\theta}}, \widehat{F})},$$

where

$$\begin{aligned}
g_{1n}(\boldsymbol{x},\boldsymbol{z};\widehat{\boldsymbol{\theta}},\widehat{F}) =& \sum_{j=1}^{s_n}\sum_{i=1}^{n}\left\{1 - (1-R_i)\frac{P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x},\boldsymbol{Z}_i,\boldsymbol{W}_i)}{\sum_{j'=1}^{s_n}\sum_{k'=1}^{m} P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x}_{k'},\boldsymbol{Z}_i,\boldsymbol{W}_i)B_{j'}^q(\boldsymbol{Z}_i)\widehat{p}_{k'j'}}\right\} \\
&\times I(\|\boldsymbol{Z}_i - \boldsymbol{z}\| \le \xi_n) B_j^q(\boldsymbol{z}).
\end{aligned}$$

We wish to show that $(ns_n)^{-1} g_{1n}(\boldsymbol{x},\boldsymbol{z};\widehat{\boldsymbol{\theta}},\widehat{F})$ is bounded away from zero for sufficiently

large n. Because

$$
\begin{aligned}
& n^{-1}\sum_{j'=1}^{s_n}\sum_{k=1}^{m} P_{\widehat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_k,\boldsymbol{z},\boldsymbol{w})B_{j'}^{q}(\boldsymbol{z})p_{kj'} \\
&= \int_{\widetilde{\boldsymbol{x}}} P_{\widehat{\boldsymbol{\theta}}}(y|\widetilde{\boldsymbol{x}},\boldsymbol{z},\boldsymbol{w})\widehat{F}(d\widetilde{\boldsymbol{x}},\boldsymbol{z}) \to \int_{\widetilde{\boldsymbol{x}}} P_{\boldsymbol{\theta}^*}(y|\widetilde{\boldsymbol{x}},\boldsymbol{z},\boldsymbol{w})F^*(d\widetilde{\boldsymbol{x}},\boldsymbol{z})
\end{aligned}
$$

uniformly in $(y,\boldsymbol{z},\boldsymbol{w})$, $(ns_n)^{-1}g_{1n}(\boldsymbol{x},\boldsymbol{z};\widehat{\boldsymbol{\theta}},\widehat{F})$ converges to $g_1(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^*,F^*)$ for $(\boldsymbol{x},\boldsymbol{z})$ in the support of $(\boldsymbol{X},\boldsymbol{Z})$, where

$$
g_1(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^*,F^*) = \mathrm{E}\left[\left\{1-(1-R)\frac{P_{\boldsymbol{\theta}^*}(Y|\boldsymbol{x},\boldsymbol{Z},\boldsymbol{W})\int_{\widetilde{\boldsymbol{x}}}F^*(d\widetilde{\boldsymbol{x}},\boldsymbol{Z})}{\int_{\widetilde{\boldsymbol{x}}}P_{\boldsymbol{\theta}^*}(Y|\widetilde{\boldsymbol{x}},\boldsymbol{Z},\boldsymbol{W})F^*(d\widetilde{\boldsymbol{x}},\boldsymbol{Z})}\right\} f_{\boldsymbol{z}}(\boldsymbol{Z})\middle|\boldsymbol{Z}=\boldsymbol{z}\right] \geq 0,
$$

and $f_{\boldsymbol{z}}(\cdot)$ is the density function of $\boldsymbol{Z}$. Consequently, $\sum_{k=1}^{m}\widehat{P}(\boldsymbol{X}=\boldsymbol{x}_k|\boldsymbol{z})$ converges to

$$
\int \frac{\mathrm{E}\left\{Rf_{\boldsymbol{z}}(\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\right\}}{g_1(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^*,F^*)}d\boldsymbol{x} = 1.
$$

If $g_1(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^*,F^*)$ is not bounded away from zero, then there exists $\boldsymbol{x}_0 \in \mathbb{D}_{\boldsymbol{x}}$, where $\mathbb{D}_{\boldsymbol{x}}$ is the support of $\boldsymbol{X}$, such that $g_1(\boldsymbol{x}_0,\boldsymbol{z};\boldsymbol{\theta}^*,F^*)=0$. Because $g_1(\boldsymbol{x}_0,\boldsymbol{z};\boldsymbol{\theta}^*,F^*)$ is a smooth function of the continuous components of $\boldsymbol{x}$, there exists a positive constant δ such that for any $\epsilon>0$,

$$
\begin{aligned}
1 &\geq \int \frac{\mathrm{E}\left\{Rf_{\boldsymbol{z}}(\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\right\}}{g_1(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^*,F^*)+\epsilon}d\boldsymbol{x} \geq \int_{\|\boldsymbol{x}-\boldsymbol{x}_0\|\leq\delta} \frac{\mathrm{E}\left\{Rf_{\boldsymbol{z}}(\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\right\}}{|g_1(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^*,F^*)|+\epsilon}d\boldsymbol{x} \\
&\gtrsim \int_{\|\boldsymbol{x}-\boldsymbol{x}_0\|\leq\delta} \frac{\mathrm{E}\left\{Rf_{\boldsymbol{z}}(\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\right\}}{\|\boldsymbol{x}-\boldsymbol{x}_0\|+\epsilon}d\boldsymbol{x}, \qquad (4.28)
\end{aligned}
$$

where "$\gtrsim$" means greater than or equal to up to a constant. Because $\int_{\|\boldsymbol{x}-\boldsymbol{x}_0\|\leq\delta}(1/\|\boldsymbol{x}-\boldsymbol{x}_0\|)d\boldsymbol{x}$ is infinite, the last integration in expression (4.28) also goes to ∞ when $\epsilon\to 0$, which is a contradiction. Thus, $g_1(\boldsymbol{x}_0,\boldsymbol{z};\boldsymbol{\theta}^*,F^*)$ is bounded away from zero for $(\boldsymbol{x},\boldsymbol{z})$ in the support of $(\boldsymbol{X},\boldsymbol{Z})$. The same conclusion holds for $(ns_n)^{-1}g_{1n}(\boldsymbol{x},\boldsymbol{z};\widehat{\boldsymbol{\theta}},\widehat{F})$ when n is sufficiently large.

The final step is to prove that $\boldsymbol{\theta}^*=\boldsymbol{\theta}_0$ and $F^*=F_0$ through the Kullback-Leibler

inequality. Let

$$\widetilde{p}_{kj} = \frac{\sum_{i=1}^{n} R_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{Z}_i)/P(R_i = 1|Y_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)}{\sum_{i=1}^{n} R_i B_j^q(\boldsymbol{Z}_i)/P(R_i = 1|Y_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)},$$

and let $\widetilde{F}(\boldsymbol{x}, \boldsymbol{z}) = n^{-1} \sum_{k=1}^{m} \sum_{i=1}^{n} I(\boldsymbol{x}_k \leq \boldsymbol{x}, \boldsymbol{Z}_i \leq \boldsymbol{z}) \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i) \widetilde{p}_{kj}$. By the approximation theory of B-splines (Schumaker 1981), $\widetilde{F}(\boldsymbol{x}, \boldsymbol{z}) \to F_0(\boldsymbol{x}, \boldsymbol{z})$ uniformly. Furthermore, it follows from the definitions of $\widehat{F}$ and $\widetilde{F}$ that $\widehat{F}$ is absolutely continuous with respect to $\widetilde{F}$. Thus, $d\widehat{F}/d\widetilde{F}$ converges uniformly to dF^*/dF_0. By Condition (C.3), F^* is continuously differentiable with respect to $\boldsymbol{x}$ and $\boldsymbol{z}$.

By the definitions of $\widehat{\boldsymbol{\theta}}$ and $\{\widehat{p}_{kj}\}$, we have $n^{-1} l_n(\widehat{\boldsymbol{\theta}}, \{\widehat{p}_{kj}\}) \geq n^{-1} l_n(\boldsymbol{\theta}_0, \{\widetilde{p}_{kj}\})$, i.e.,

$$\begin{aligned} &- n^{-1} \sum_{i=1}^{n} R_i \log \frac{P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)}{P_{\boldsymbol{\theta}_0}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i)} - n^{-1} \sum_{i=1}^{n} R_i \sum_{k=1}^{m} I(\boldsymbol{X}_i = \boldsymbol{x}_k) \sum_{j=1}^{s_n} B_j^q(\boldsymbol{Z}_i) \log \frac{\widehat{p}_{kj}}{\widetilde{p}_{kj}} \\ &- n^{-1} \sum_{i=1}^{n} (1 - R_i) \log \frac{\int P_{\widehat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{x}, \boldsymbol{Z}_i, \boldsymbol{W}_i) \widehat{F}(d\boldsymbol{x}, \boldsymbol{Z}_i)}{\int P_{\boldsymbol{\theta}_0}(Y_i|\boldsymbol{x}, \boldsymbol{Z}_i, \boldsymbol{W}_i) \widetilde{F}(d\boldsymbol{x}, \boldsymbol{Z}_i)} \leq 0. \end{aligned} \tag{4.29}$$

The first term in expression (4.29) converges to

$$-\mathrm{E}\left\{ R \log \frac{P_{\boldsymbol{\theta}^*}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})}{P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})} \right\}. \tag{4.30}$$

By the approximation theory of B-splines (Schumaker 1981), $\sum_{j=1}^{s_n} B_j^q(\boldsymbol{z}) \log(\widehat{p}_{kj}/\widetilde{p}_{kj})$ is asymptotically equivalent to

$$\log \frac{\sum_{j=1}^{s_n} B_j^q(\boldsymbol{z}) \widehat{p}_{kj}}{\sum_{j=1}^{s_n} B_j^q(\boldsymbol{z}) \widetilde{p}_{kj}} = \log \frac{d\widehat{F}(\boldsymbol{x}, \boldsymbol{z})}{d\widetilde{F}(\boldsymbol{x}, \boldsymbol{z})}\bigg|_{\boldsymbol{x}=\boldsymbol{x}_k}.$$

Thus $\sum_{j=1}^{s_n} B_j^q(\boldsymbol{z}) \log(\widehat{p}_{kj}/\widetilde{p}_{kj})$ converges uniformly to $\log\{dF^*(\boldsymbol{x}, \boldsymbol{z})/dF_0(\boldsymbol{x}, \boldsymbol{z})\}|_{\boldsymbol{x}=\boldsymbol{x}_k}$. As a result, the second term in expression (4.29) converges to

$$-\mathrm{E}\left\{ R \log \frac{dF^*(\boldsymbol{X}, \boldsymbol{Z})}{dF_0(\boldsymbol{X}, \boldsymbol{Z})} \right\}. \tag{4.31}$$

The third term in expression (4.29) converges to

$$-\mathrm{E}\left\{(1-R)\log\frac{\int P_{\boldsymbol{\theta}^*}(Y|\boldsymbol{x},\boldsymbol{Z},\boldsymbol{W})F^*(d\boldsymbol{x},\boldsymbol{Z})}{\int P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{x},\boldsymbol{Z},\boldsymbol{W})F_0(d\boldsymbol{x},\boldsymbol{Z})}\right\}. \tag{4.32}$$

By combining expressions (4.30), (4.31), and (4.32), we conclude that the Kullback-Leibler information of the density indexed by $\boldsymbol{\theta}^*$ and F^* with respect to the true density is nonpositive and thus must be zero. Therefore, the two densities are identical almost surely. For $R = 1$, this implies that $P_{\boldsymbol{\theta}^*}(Y|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W})F^*(\boldsymbol{X},\boldsymbol{Z}) = P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W})F_0(\boldsymbol{X},\boldsymbol{Z})$. It follows from Condition (C.2) that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $F^* = F_0$. Thus, Theorem 1 holds. ■

Proof of Theorem 4.2. Let $l_{\boldsymbol{\theta}}$ denote the score function for $\boldsymbol{\theta}_0$ and $l_F(h)$ denote the score function along the submodel $\{1+\epsilon h(\boldsymbol{x},\boldsymbol{z})\}dF_0(\boldsymbol{x},\boldsymbol{z})$ based on one complete observation $(Y,\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W})$, where $h \in L_2(\mathcal{P})$, $\mathcal{P}$ is the probability measure indexed by $(\boldsymbol{\theta}_0, F_0)$, and $\mathrm{E}\{h(\boldsymbol{X},\boldsymbol{Z})\} = 0$. We have $l_{\boldsymbol{\theta}} = \partial \log P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W})/\partial\boldsymbol{\theta}$ and $l_F(h) = h$. For two-phase studies, the score operators are $l^o_{\boldsymbol{\theta}} = Rl_{\boldsymbol{\theta}} + (1-R)\mathrm{E}(l_{\boldsymbol{\theta}}|Y,\boldsymbol{Z},\boldsymbol{W})$ and $l^o_F = Rl_F + (1-R)\mathrm{E}(l_F|Y,\boldsymbol{Z},\boldsymbol{W})$. The information operator is

$$\begin{bmatrix} l^{o*}_{\boldsymbol{\theta}} l^o_{\boldsymbol{\theta}} & l^{o*}_{\boldsymbol{\theta}} l^o_F \\ l^{o*}_F l^o_{\boldsymbol{\theta}} & l^{o*}_F l^o_F \end{bmatrix},$$

where $l^{o*}_{\boldsymbol{\theta}}$ and l^{o*}_F are the adjoint operators of $l^o_{\boldsymbol{\theta}}$ and l^o_F, respectively. We calculate the information operator as

$$\begin{aligned} l^{o*}_{\boldsymbol{\theta}} l^o_{\boldsymbol{\theta}} &= \mathrm{E}\left\{Rl_{\boldsymbol{\theta}}^{\otimes 2} + (1-R)\mathrm{E}(l_{\boldsymbol{\theta}}|Y,\boldsymbol{Z},\boldsymbol{W})^{\otimes 2}\right\}, \\ l^{o*}_{\boldsymbol{\theta}} l^o_F(h) &= l^{o*}_F l^o_{\boldsymbol{\theta}}(h)^{\mathrm{T}} = \mathrm{E}\left[\mathrm{E}\left\{Rl_{\boldsymbol{\theta}} + (1-R)\mathrm{E}(l_{\boldsymbol{\theta}}|Y,\boldsymbol{Z},\boldsymbol{W})|\boldsymbol{X},\boldsymbol{Z}\right\}h(\boldsymbol{X},\boldsymbol{Z})\right], \text{ and} \\ l^{o*}_F l^o_F(h) &= \mathrm{E}(R|\boldsymbol{X},\boldsymbol{Z})h(\boldsymbol{X},\boldsymbol{Z}) + \mathrm{E}\left\{(1-R)\mathrm{E}(h(\boldsymbol{X},\boldsymbol{Z})|Y,\boldsymbol{Z},\boldsymbol{W})|\boldsymbol{X},\boldsymbol{Z}\right\}. \end{aligned}$$

This information operator is the sum of an invertible operator and a compact operator from the space $\mathbb{M} \equiv \mathbb{R}^d \times BV(\mathbb{D}_{\boldsymbol{x},\boldsymbol{z}})$ to itself, where d is the dimension of $\boldsymbol{\theta}$, and $BV(\mathbb{D}_{\boldsymbol{x},\boldsymbol{z}})$ is the

space of functions with bounded total variation in the support of $(\boldsymbol{X}, \boldsymbol{Z})$. By Theorem 4.7 of Rudin (1973), the information operator is invertible if it is one to one, or equivalently, the Fisher information along any nontrivial submodel is nonzero.

Suppose that the Fisher information is zero along some submodel $[\boldsymbol{\theta}_0 + \epsilon \boldsymbol{v}, dF_0(\boldsymbol{x}, \boldsymbol{z})\{1 + \epsilon h(\boldsymbol{x}, \boldsymbol{z})\}]$. Then, the score function along this submodel, i.e., ${l_{\boldsymbol{\theta}}^o}^{\mathrm{T}} \boldsymbol{v} + l_F^o(h)$, is zero. We set $R = 1$ to obtain $l_{\boldsymbol{\theta}}^{\mathrm{T}} \boldsymbol{v} + l_F(h) = 0$ for any $(Y, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) \in \mathcal{C}$. Specifically, for any $(y_i, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}_i) \in \mathcal{C}$, $i = 1, 2$, we have

$$\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_0}(y_1 | \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}_1) \right\}^{\mathrm{T}} \boldsymbol{v} + h(\boldsymbol{x}, \boldsymbol{z}) = \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_0}(y_2 | \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}_2) \right\}^{\mathrm{T}} \boldsymbol{v} + h(\boldsymbol{x}, \boldsymbol{z}),$$

which can be rewritten as a linear equation on $\boldsymbol{v}$, i.e.,

$$\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_0}(y_1 | \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}_1) - \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_0}(y_2 | \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}_2) \right\}^{\mathrm{T}} \boldsymbol{v} = 0.$$

By Condition (C.2), $\boldsymbol{v} = 0$ and $h = 0$ with probability one. Thus, the information operator is invertible. Consequently, there exists a function h such that ${l_F^o}^* l_F^o(h) = {l_F^o}^* l_{\boldsymbol{\theta}}^o$, i.e.,

$$\begin{aligned} &\mathrm{E}(R | \boldsymbol{X}, \boldsymbol{Z}) h + \mathrm{E}\left\{ (1 - R) \mathrm{E}(h | Y, \boldsymbol{Z}, \boldsymbol{W}) | \boldsymbol{X}, \boldsymbol{Z} \right\} \\ =&\mathrm{E}\left\{ R l_{\boldsymbol{\theta}} + (1 - R) \mathrm{E}(l_{\boldsymbol{\theta}} | Y, \boldsymbol{Z}, \boldsymbol{W}) | \boldsymbol{X}, \boldsymbol{Z} \right\}. \end{aligned} \tag{4.33}$$

This means that the least favorable direction for $\boldsymbol{\theta}_0$ exists. In addition, by using the arguments in the proof of Theorem 3.4 of Zeng (2005), we can show that h is q-times continuously differentiable.

Because $(\widehat{\boldsymbol{\theta}}, \widehat{F})$ maximizes expression (4.24), the derivatives of the log-likelihood function with respect to ϵ along the submodel $(\widehat{\boldsymbol{\theta}} + \epsilon \boldsymbol{v}, d\widehat{F})$ for any $\boldsymbol{v}$ and the submodel $\{\widehat{\boldsymbol{\theta}}, d\widehat{F}(1 + \epsilon h_n)\}$ must be zero, where h_n is the projection of h onto the tangent space of the sieve space. By the approximation theory of B-splines (Schumaker 1981), we have $\|h_n - h\|_{L_2} \lesssim$

$s_n^{-q/d_{\boldsymbol{z}}}$. Therefore, $(\widehat{\boldsymbol{\theta}}, \widehat{F})$ is the solution to the functional $\Psi_n(\boldsymbol{\theta}, F) = 0$, where $\Psi_n(\boldsymbol{\theta}, F) = \Psi_{1n}(\boldsymbol{\theta}, F) - \Psi_{2n}(\boldsymbol{\theta}, F)$,

$$\begin{aligned}
\Psi_{1n}(\boldsymbol{\theta}, F) = & \mathcal{P}_n \left\{ R \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) \right\} \\
& + \mathcal{P}_n \left\{ (1-R) \int \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) g_2(Y, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{x}; \boldsymbol{\theta}, F) F(d\boldsymbol{x}, \boldsymbol{Z}) \right\}, \\
\Psi_{2n}(\boldsymbol{\theta}, F) = & \mathcal{P}_n \{ R h_n(\boldsymbol{X}, \boldsymbol{Z}) \} \\
& + \mathcal{P}_n \left\{ (1-R) \int g_2(Y, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{x}; \boldsymbol{\theta}, F) h_n(\boldsymbol{x}, \boldsymbol{Z}) F(d\boldsymbol{x}, \boldsymbol{Z}) \right\},
\end{aligned}$$

$\mathcal{P}_n$ is the empirical measure of the sample, and

$$g_2(Y, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{x}; \boldsymbol{\theta}, F) = \frac{P_{\boldsymbol{\theta}}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W})}{\int P_{\boldsymbol{\theta}}(Y|\widetilde{\boldsymbol{x}}, \boldsymbol{Z}, \boldsymbol{W}) F(d\widetilde{\boldsymbol{x}}, \boldsymbol{Z})}.$$

Let $\Psi(\boldsymbol{\theta}, F)$ be the same as $\Psi_n(\boldsymbol{\theta}, F)$ except that $\mathcal{P}_n$ is replaced by $\mathcal{P}$. Clearly, $\widehat{\boldsymbol{\theta}}$ satisfies the following equation:

$$n^{1/2} \left\{ \Psi_n(\widehat{\boldsymbol{\theta}}, \widehat{F}) - \Psi(\widehat{\boldsymbol{\theta}}, \widehat{F}) \right\} = -n^{1/2} \Psi(\widehat{\boldsymbol{\theta}}, \widehat{F}). \tag{4.34}$$

We wish to use Theorem 2.11.22 of van der Vaart and Wellner (1996) to show that

$$n^{1/2} \left\{ \Psi_n(\widehat{\boldsymbol{\theta}}, \widehat{F}) - \Psi(\widehat{\boldsymbol{\theta}}, \widehat{F}) \right\} = n^{1/2} (\mathcal{P}_n - \mathcal{P}) \{ l^o_{\boldsymbol{\theta}} - l^o_F(h_n) \} + o_p(1). \tag{4.35}$$

Note that the left-hand side of equation (4.35) is an empirical process of the following two classes of functions indexed by $(\widehat{\boldsymbol{\theta}}, \widehat{F})$:

$$\begin{aligned}
\mathcal{F}_{1n} = & \left\{ R \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) + (1-R) \int \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) \right. \\
& \left. \times g_2(Y, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{x}; \boldsymbol{\theta}, F) F(d\boldsymbol{x}, \boldsymbol{Z}) : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| + \|F - F_0\| \leq \epsilon_0 \right\};
\end{aligned}$$

$$\mathcal{F}_{2n} = \Big\{ Rh_n(\boldsymbol{X}, \boldsymbol{Z}) + (1-R) \int g_2(Y, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{x}; \boldsymbol{\theta}, F) h_n(\boldsymbol{x}, \boldsymbol{Z}) F(d\boldsymbol{x}, \boldsymbol{Z}) : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| + \|F - F_0\| \leq \epsilon_0 \Big\},$$

where $\|F - F_0\|$ is the supreme norm in $\mathbb{D}_{\boldsymbol{x},\boldsymbol{z}}$. By Theorem 1 and the approximation theory of B-splines (Schumaker 1981), it is straightforward to verify that

$$\begin{aligned}
& R \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\widehat{\boldsymbol{\theta}}}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) + (1-R) \int \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\widehat{\boldsymbol{\theta}}}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) \\
& \quad \times g_2(Y, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{x}; \widehat{\boldsymbol{\theta}}, \widehat{F}) \widehat{F}(d\boldsymbol{x}, \boldsymbol{Z}) \\
\to & R \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) + (1-R) \int \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) \\
& \quad \times \frac{P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) F_0(d\boldsymbol{x}, \boldsymbol{Z})}{\int P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) F_0(d\boldsymbol{x}, \boldsymbol{Z})} \\
= & R l_{\boldsymbol{\theta}} + (1-R) \mathrm{E}\{l_{\boldsymbol{\theta}} | Y, \boldsymbol{Z}, \boldsymbol{W}\} = l^o_{\boldsymbol{\theta}},
\end{aligned}$$

and

$$\begin{aligned}
& Rh_n(\boldsymbol{X}, \boldsymbol{Z}) + (1-R) \int g_2(Y, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{x}; \widehat{\boldsymbol{\theta}}, \widehat{F}) h_n(\boldsymbol{x}, \boldsymbol{Z}) \widehat{F}(d\boldsymbol{x}, \boldsymbol{Z}) \\
\to & Rh(\boldsymbol{X}, \boldsymbol{Z}) + (1-R) \frac{\int h(\boldsymbol{x}, \boldsymbol{Z}) P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) F_0(d\boldsymbol{x}, \boldsymbol{Z})}{\int P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) F_0(d\boldsymbol{x}, \boldsymbol{Z})} \\
= & Rh(\boldsymbol{X}, \boldsymbol{Z}) + (1-R) \mathrm{E}\left\{h(\boldsymbol{X}, \boldsymbol{Z}) | Y, \boldsymbol{Z}, \boldsymbol{W}\right\} = l^o_F(h)
\end{aligned}$$

uniformly in $(Y, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$.

Clearly, all functions in the classes $\mathcal{F}_{1n}$ and $\mathcal{F}_{2n}$ are uniformly bounded. We wish to verify the conditions in Theorem 2.11.22 of van der Vaart and Wellner (1996). We first show that the classes of functions $\mathcal{F}_{1n}$ and $\mathcal{F}_{2n}$ satisfy the uniform entropy condition. Pick any two functions from $\mathcal{F}_{1n}$, say f_1 and f_2, which are indexed by $(\boldsymbol{\theta}_1, F_1)$ and $(\boldsymbol{\theta}_2, F_2)$, respectively.

The difference between the two functions is bounded from above by

$$
\begin{aligned}
&\left|\frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_1}(Y|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W}) - \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_2}(Y|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W})\right| \\
&+\left|\int \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_1}(Y|\boldsymbol{x},\boldsymbol{Z},\boldsymbol{W}) g_2(Y,\boldsymbol{Z},\boldsymbol{W},\boldsymbol{x};\boldsymbol{\theta}_1,F_1)(F_1-F_2)(d\boldsymbol{x},\boldsymbol{Z})\right| \\
&+\left|\int \left\{\frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_1}(Y|\boldsymbol{x},\boldsymbol{Z},\boldsymbol{W}) - \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_2}(Y|\boldsymbol{x},\boldsymbol{Z},\boldsymbol{W})\right\}\right. \\
&\quad \left.\times\, g_2(Y,\boldsymbol{Z},\boldsymbol{W},\boldsymbol{x};\boldsymbol{\theta}_1,F_1)F_2(d\boldsymbol{x},\boldsymbol{Z})\right| \\
&+\left|\int \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_2}(Y|\boldsymbol{x},\boldsymbol{Z},\boldsymbol{W})\Big\{g_2(Y,\boldsymbol{Z},\boldsymbol{W},\boldsymbol{x};\boldsymbol{\theta}_1,F_1)\right. \\
&\quad \left.- g_2(Y,\boldsymbol{Z},\boldsymbol{W},\boldsymbol{x};\boldsymbol{\theta}_2,F_2)\Big\}F_2(d\boldsymbol{x},\boldsymbol{Z})\right| \\
&\equiv (i)+(ii)+(iii)+(iv).
\end{aligned}
$$

By the mean-value theorem, $(i) \lesssim \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$. Because the denominator in the expression of $g_2(Y,\boldsymbol{Z},\boldsymbol{W},\boldsymbol{x};\boldsymbol{\theta},F)$ is bounded away from zero, we obtain that

$$
(ii) \lesssim \int |F_1(\boldsymbol{x},\boldsymbol{Z}) - F_2(\boldsymbol{x},\boldsymbol{Z})|\, d\boldsymbol{x} \lesssim \int |F_1(\boldsymbol{x},\boldsymbol{z}) - F_2(\boldsymbol{x},\boldsymbol{z})|\, d\boldsymbol{x}d\boldsymbol{z}.
$$

By the mean-value theorem,

$$
(iii) \lesssim \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \int g_2(Y,\boldsymbol{Z},\boldsymbol{W},\boldsymbol{x};\boldsymbol{\theta}_1,F_1)F_2(d\boldsymbol{x},\boldsymbol{Z}) \lesssim \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.
$$

Likewise,

$$
(iv) \lesssim \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \int |F_1(\boldsymbol{x},\boldsymbol{z}) - F_2(\boldsymbol{x},\boldsymbol{z})| d\boldsymbol{x}d\boldsymbol{z}.
$$

Combining the above inequalities for (i), (ii), (iii), and (iv), we have

$$
|f_1 - f_2| \lesssim \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \int |F_1(\boldsymbol{x},\boldsymbol{z}) - F_2(\boldsymbol{x},\boldsymbol{z})| d\boldsymbol{x}d\boldsymbol{z}.
$$

Thus, the Cauchy-Schwartz inequality implies that, for any finite measure $\mathcal{Q}$,

$$\begin{aligned}\|f_1 - f_2\|_{L_2(\mathcal{Q})} &\lesssim \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \left\{ \int |F_1(\boldsymbol{x}, \boldsymbol{z}) - F_2(\boldsymbol{x}, \boldsymbol{z})|^2 d\boldsymbol{x} d\boldsymbol{z} \right\}^{1/2} \\ &= \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \|F_1(\boldsymbol{X}, \boldsymbol{Z}) - F_2(\boldsymbol{X}, \boldsymbol{Z})\|_{L_2(\widetilde{\mathcal{Q}})}, \end{aligned} \tag{4.36}$$

where $\widetilde{\mathcal{Q}}$ is the uniform measure on $\mathbb{D}_{\boldsymbol{x},\boldsymbol{z}}$. We conclude that

$$\begin{aligned} N\{\epsilon, \mathcal{F}_{1n}, L_2(\mathcal{Q})) \lesssim & N(\epsilon/2, (\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \epsilon_0), |\cdot|\} \\ & \times N(\epsilon/2, (F : \|F - F_0\|_\infty < \epsilon_0), L_2(\widetilde{\mathcal{Q}})\}, \end{aligned} \tag{4.37}$$

where $N(\cdot, \cdot, \cdot)$ denotes the covering number. On the right-hand side of (4.37), the first covering number is $O(1/\epsilon^d)$. The second covering number is $O[\exp\{\epsilon^{-2V/(V+2)}\}]$, where V is some positive index. To see the latter result, we observe that $(F : \|F - F_0\|_\infty < \epsilon)$ is in the symmetric convex hull of a Vapnik-Chervonenkis class $[I\{\boldsymbol{a} < (\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}} \leq \boldsymbol{b}\}\colon \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{d_{\boldsymbol{x}}+d_{\boldsymbol{z}}}]$, where $d_{\boldsymbol{x}}$ denotes the dimension of $\boldsymbol{X}$. The result follows from Theorem 2.6.9 of van der Vaart and Wellner (1996). Therefore, expression (4.37) implies that $\mathcal{F}_{1n}$ satisfies the uniform entropy condition in Theorem 2.11.22 of van der Vaart and Wellner (1996). By similar arguments and the fact that $\|h_n\|_{L_2} \lesssim \|h\|_{L_2}$, we can show that $\mathcal{F}_{2n}$ also satisfies the uniform entropy condition.

If we replace measure $\mathcal{Q}$ by $\mathcal{P}$, then expression (4.36) implies that the functions in $\mathcal{F}_{1n}$ and $\mathcal{F}_{2n}$ are Lipschitz continuous with respect to $(\boldsymbol{\theta}, F)$ in the metric defined as

$$\rho\{(\boldsymbol{\theta}_1, F_1), (\boldsymbol{\theta}_2, F_2)\} = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \|F_1 - F_2\|_{L_2(\mathcal{P})}.$$

As a result, condition (2.11.21) in Theorem 2.11.22 of van der Vaart and Wellner (1996) holds. In addition, the total boundedness of the index set (θ, F) holds due to the precompactness of $(\boldsymbol{\theta}, F)$ under the uniform metric. We have now verified all of the conditions in

Theorem 2.11.22 of van der Vaart and Wellner (1996). Thus, equation (4.35) follows from that theorem.

By combining equations (4.34) and (4.35), we have

$$-n^{1/2}\left\{\Psi_1(\widehat{\boldsymbol{\theta}},\widehat{F})-\Psi_2(\widehat{\boldsymbol{\theta}},\widehat{F})\right\}=n^{1/2}(\mathcal{P}_n-\mathcal{P})\{l^o_{\boldsymbol{\theta}}-l^o_F(h_n)\}+o_p(1), \tag{4.38}$$

where $\Psi_1(\boldsymbol{\theta},F)$ and $\Psi_2(\boldsymbol{\theta},F)$ are the same as $\Psi_{1n}(\boldsymbol{\theta},F)$ and $\Psi_{2n}(\boldsymbol{\theta},F)$, respectively, except that $\mathcal{P}_n$ is replaced by $\mathcal{P}$. The left-hand side of equation (4.38) can be linearized around $(\boldsymbol{\theta}_0,F_0)$. Specifically,

$$\begin{aligned}\Psi_1(\widehat{\boldsymbol{\theta}},\widehat{F})=\Psi_1(\boldsymbol{\theta}_0,F_0)&+\mathcal{P}\{R\frac{\partial^2}{\partial\boldsymbol{\theta}^{\mathrm{T}}\partial\boldsymbol{\theta}}\log P_{\boldsymbol{\theta}^*}(Y|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\}\\
&+\mathcal{P}\left[(1-R)\int\frac{\partial}{\partial\boldsymbol{\theta}}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log P_{\boldsymbol{\theta}^*}(Y|\boldsymbol{x},\boldsymbol{Z},\boldsymbol{W})g_2(Y,\boldsymbol{Z},\boldsymbol{W},\boldsymbol{x};\boldsymbol{\theta}^*,F^*)\right\}\right.\\
&\left.\times(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\widehat{F}(d\boldsymbol{x},\boldsymbol{Z})\right]\\
&+\mathcal{P}\left[(1-R)\int\frac{\partial}{\partial\boldsymbol{\theta}}\log P_{\boldsymbol{\theta}^*}(Y|\boldsymbol{x},\boldsymbol{Z},\boldsymbol{W})\left\{\frac{\partial}{\partial F}g_2(Y,\boldsymbol{Z},\boldsymbol{W},\boldsymbol{x};\boldsymbol{\theta}^*,F^*)\right.\right.\\
&\left.\left.\times F^*(d\boldsymbol{x},\boldsymbol{Z})\right\}(\widehat{F}-F_0)\right],\end{aligned}$$

where $\partial/\partial F$ denotes the pathwise derivative, and $(\boldsymbol{\theta}^*,F^*)$ lies between $(\widehat{\boldsymbol{\theta}},\widehat{F})$ and $(\boldsymbol{\theta}_0,F_0)$. Similar expansions can be obtained for $\Psi_2(\widehat{\boldsymbol{\theta}},\widehat{F})$. By the approximation theory of B-splines (Schumaker 1981), we can show that the left-hand side of (4.38) equals

$$\begin{aligned}&-n^{1/2}\left\{1+o_p(1)\right\}\mathrm{E}\left\{l^o_{\boldsymbol{\theta}\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+l^o_{\boldsymbol{\theta}F}(\widehat{F}-F_0)-l^o_{F\boldsymbol{\theta}}(h_n)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)-l^o_{FF}(h_n,\widehat{F}-F_0)\right\}\\
&-n^{1/2}\left\{\Psi_1(\boldsymbol{\theta}_0,F_0)-\Psi_2(\boldsymbol{\theta}_0,F_0)\right\},\end{aligned} \tag{4.39}$$

where $l^o_{\boldsymbol{\theta}\boldsymbol{\theta}}$ is the derivative of $l^o_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\theta}$, $l^o_{\boldsymbol{\theta}F}(h)$ is the derivative of $l^o_{\boldsymbol{\theta}}$ with respect to F along the direction h, $l^o_{F\boldsymbol{\theta}}(h)$ is the derivative of $l^o_F(h)$ with respect to $\boldsymbol{\theta}$, and $l^o_{FF}(h_1,h_2)$ is the derivative of $l^o_F(h_1)$ with respect to F along the direction h_2.

Because we have chosen h to be the least favorable direction for $\boldsymbol{\theta}_0$ and $\|h_n - h\|_{L_2} \lesssim s_n^{-q/d_{\boldsymbol{z}}}$, we have $\mathrm{E}\{l^o_{FF}(h_n, \widehat{F} - F_0)\} = \mathrm{E}\{l^o_{\boldsymbol{\theta}F}(\widehat{F} - F_0)\} + O(s_n^{-q/d_{\boldsymbol{z}}})$ and $\mathrm{E}\{l^o_{F\boldsymbol{\theta}}(h_n)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} = \mathrm{E}\{l^o_{F\boldsymbol{\theta}}(h)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} + O(s_n^{-q/d_{\boldsymbol{z}}})$. Thus, by Condition (C.5), the first term in expression (4.39) is $n^{1/2}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + O(n^{1/2}s_n^{-q/d_{\boldsymbol{z}}}) = n^{1/2}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o(1)$, where $\boldsymbol{\Sigma} = -\mathrm{E}\{l^o_{\boldsymbol{\theta}\boldsymbol{\theta}} - l^o_{F\boldsymbol{\theta}}(h)\}$, which is an invertible matrix due to the invertibility of the information operator for $(\boldsymbol{\theta}_0, F_0)$. Because $\mathcal{P}\{R\partial \log P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})/\partial\boldsymbol{\theta}\} = 0$ and

$$\mathcal{P}\left\{(1-R)\int \frac{\partial}{\partial\boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W}) \frac{P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W})F_0(d\boldsymbol{x}, \boldsymbol{Z})}{\int P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{W})F_0(d\boldsymbol{x}, \boldsymbol{Z})}\right\} = 0,$$

the last term in (4.39) equals zero. It follows from equation (4.38) that

$$n^{1/2}\{1 + o_p(1)\}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1) = n^{1/2}(\mathcal{P}_n - \mathcal{P})\left\{l^o_{\boldsymbol{\theta}} - l^o_F(h)\right\}.$$

Thus, we have established the asymptotic normality in Theorem 4.2. Because $\boldsymbol{\Sigma}^{-1}\{l^o_{\boldsymbol{\theta}} - l^o_F(h)\}$ is the efficient influence function for $\boldsymbol{\theta}_0$, its limiting covariance matrix attains the semiparametric efficiency bound. ■

For a given $\boldsymbol{\theta}$, we define $\widehat{F}_{\boldsymbol{\theta}}$ as the joint distribution function of $(\boldsymbol{X}, \boldsymbol{Z})$ that maximizes $l_n(\boldsymbol{\theta}, \{p_{kj}\})$. By the arguments in the proof of Theorem 4.1, we can show that for any $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ in probability, the estimator $\widehat{F}_{\widehat{\boldsymbol{\theta}}} \to F_0$ uniformly. Furthermore, given the existence of the least favorable directions, we can construct the least favorable model. These two facts imply that the profile likelihood theory in Murphy and van der Vaart (2000) holds for our approach. Thus, the inverse of the negative Hessian matrix of the profile likelihood function is a consistent estimator for the limiting covariance matrix of $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

CHAPTER 5: OPTIMAL TWO-PHASE DESIGNS AND FUTURE RESEARCH

5.1 Optimal Two-Phase Designs

5.1.1 Introduction

Previous research on two-phase studies has largely focused on the inference procedures rather than the design aspects. An important topic of investigation is the optimal study design when the primary interest is to estimate the regression coefficients of the expensive covariates. If a discrete outcome is of primary interest and there are no inexpensive covariates, then it is well known that the case-control design with an equal number of cases and controls is the most efficient one. When a continuous outcome is of primary interest, two types of designs are commonly used in practice. The first design selects subjects from the two tails of the outcome distribution in the second phase. We call this design the outcome-dependent sampling (ODS) design. If the inexpensive covariates are also available in the first phase, then we can fit a marginal model relating the outcome to inexpensive covariates and use the residuals from the fitted model to select subjects. We call this design the residual-dependent sampling (RDS) design. Both the ODS and RDS designs have been adopted in the NHLBI ESP. Specifically, the ODS design has been adopted in the BMI study and the RDS design has been adopted in the BP and LDL studies.

Although the design issue is important, little research has been done to study the optimal design. Recently, Lin et al. (2013) showed that the ODS design is optimal when there is no inexpensive covariates. For general two-phase studies, it is unclear what the best sampling strategy is. In this project, we study optimal designs for parametric and semiparametric

regression problems under general two-phase studies. We derive the efficient information for estimating the regression coefficients of the expensive covariates under local alternatives. We compare efficiencies between the optimal and existing designs in extensive simulation studies.

5.1.2 Methods

Let Y be the outcome(s) of interest, X denote the expensive covariate, and Z denote the vector of inexpensive covariates. The data (Y, X, Z) are assumed to be generated from the joint distribution $P_{\theta,\phi}(Y|X,Z)\eta(X,Z)$, where $P_{\theta,\phi}(Y|X,Z) = f\{Y|\mu(X,Z);\phi\}$ is a regression model indexed by $\theta = (\alpha, \beta, \gamma^{\mathrm{T}})^{\mathrm{T}}$ and ϕ, (α, β, γ) are the regression coefficients in the linear predictor $\mu(X,Z) = \alpha + \beta X + \gamma^T Z$, ϕ is a (possibly infinite-dimensional) nuisance parameter, and $\eta(X,Z)$ is the joint distribution of X and Z.

If (Y, X, Z) is observed for all n subjects in the study, then it is standard to base inferences about θ on the conditional distribution of Y given (X, Z), such that the likelihood is

$$\prod_{i=1}^{n} P_{\theta,\phi}(Y_i|X_i, Z_i).$$

Under the two-phase design, however, only (Y, Z) is measured for all n subjects in the first phase, and X is measured for a sub-sample of size n_2 in the second phase. Let R indicate, by the values 1 versus 0, whether the subject is selected for the measurement of X in the second phase. A key assumption for the two-phase design is that the distribution of R depends on (Y, X, Z) only through the first-phase data (Y, Z). Under this assumption, the data on X are missing at random, such that the sampling indicators $(R_1, \ldots, R_n)$ can be omitted from the likelihood function when estimating θ. Thus, the observed-data likelihood and log-likelihood take the forms

$$L(\theta, \phi, \eta) = \prod_{i=1}^{n} \{P_{\theta,\phi}(Y_i|X_i, Z_i)\eta(X_i, Z_i)\}^{R_i} \left\{ \int P_{\theta,\phi}(Y_i|x, Z_i)\eta(x, Z_i)dx \right\}^{1-R_i},$$

and

$$l(\theta,\phi,\eta) = \sum_{i=1}^{n} R_i \log P_{\theta,\phi}(Y_i|X_i,Z_i) + \sum_{i=1}^{n} R_i \log \eta(X_i,Z_i) + \sum_{i=1}^{n}(1-R_i)\log \int P_{\theta,\phi}(Y_i|x,Z_i)\eta(x,Z_i)dx,$$

respectively. Our main interest lies in the inference of β.

Let $f_\mu\{Y|\mu(X,Z),\phi\} \equiv \partial \log P_{\theta,\phi}(Y|X,Z)/\partial\mu$, and $f_\phi(h_1)$ denote the score for ϕ along the submodel $t \to \phi_t(h_1)$ for one complete observation (Y,X,Z), where $h_1 \in L_2(\mathcal{P})$, $\mathcal{P}$ is the probability measure indexed by (θ,ϕ,η), and $\phi_0(h_1) = \phi$. Let $U_\theta \equiv (U_\alpha, U_\beta, U_\gamma^{\mathrm{T}})^{\mathrm{T}}$ denote the score for θ, $U_\phi(h_1)$ denote the score for ϕ along the submodel $\phi_t(h_1)$, and $U_\eta(h_2)$ denote the score for η along the submodel $\{1+th_2(x,z)\}\eta(x,z)$ under the two-phase design, where $h_2 \in L_2^0(\eta)$. We have

$$\begin{aligned}
U_\alpha =& Rf_\mu\{Y|\mu(X,Z),\phi\} + (1-R)\mathrm{E}\left[f_\mu\{Y|\mu(X,Z),\phi\}|Y,Z\right],\\
U_\beta =& Rf_\mu\{Y|\mu(X,Z),\phi\}X + (1-R)\mathrm{E}\left[f_\mu\{Y|\mu(X,Z),\phi\}X|Y,Z\right],\\
U_\gamma =& Rf_\mu\{Y|\mu(X,Z),\phi\}Z + (1-R)\mathrm{E}\left[f_\mu\{Y|\mu(X,Z),\phi\}Z|Y,Z\right],\\
U_\phi(h_1) =& Rf_\phi(h_1) + (1-R)\mathrm{E}\left\{f_\phi(h_1)|Y,Z\right\},\\
U_\eta(h_2) =& Rh_2(X,Z) + (1-R)\mathrm{E}\left[h_2(X,Z)|Y,Z\right].
\end{aligned}$$

The information operator is

$$\begin{bmatrix} U_\theta^* U_\theta & U_\theta^* U_\phi & U_\theta^* U_\eta \\ U_\phi^* U_\theta & U_\phi^* U_\phi & U_\phi^* U_\eta \\ U_\eta^* U_\theta & U_\eta^* U_\phi & U_\eta^* U_\eta \end{bmatrix},$$

where $U_\theta^* = (U_\alpha^*, U_\beta^*, U_\gamma^*)$, and U_α^*, U_β^*, U_γ^*, U_ϕ^*, and U_η^* are the adjoint operators of U_α, U_β, U_γ, U_ϕ, and U_η, respectively. Under the null hypothesis $\beta = 0$, we have $\mu(X,Z) = \alpha + \gamma^{\mathrm{T}} Z \equiv$

$\mu(Z)$. Thus, $f_\mu\{Y|\mu(Z),\phi\}$ and $f_\phi(h_1)$ do not depend on X. Consequently, we calculate the information operator as

$$
\begin{aligned}
U_\alpha^* U_\alpha =& \mathrm{E}\left(R f_\mu^2\{Y|\mu(Z),\phi\} + (1-R)\mathrm{E}\left[f_\mu\{Y|\mu(Z),\phi\}|Y,Z\right]^2\right) = \mathrm{E}\left[f_\mu^2\{Y|\mu(Z),\phi\}\right], \\
U_\gamma^* U_\gamma =& \mathrm{E}\left[f_\mu^2\{Y|\mu(Z),\phi\} Z Z^{\mathrm{T}}\right], \\
U_\alpha^* U_\gamma =& \mathrm{E}\left[f_\mu^2\{Y|\mu(Z),\phi\} Z^{\mathrm{T}}\right], \\
U_\alpha^* U_\beta =& \mathrm{E}\left[R f_\mu^2\{Y|\mu(Z),\phi\} X\right] \\
& + \mathrm{E}\left((1-R)\mathrm{E}\left[f_\mu\{Y|\mu(Z),\phi\}|Y,Z\right]\mathrm{E}\left[f_\mu\{Y|\mu(Z),\phi\}X|Y,Z\right]\right) \\
=& \mathrm{E}\left[f_\mu^2\{Y|\mu(Z),\phi\}\mathrm{E}\left(X|Z\right)\right], \\
U_\beta^* U_\gamma =& \mathrm{E}\left[f_\mu^2\{Y|\mu(Z),\phi\}\mathrm{E}\left(X|Z\right) Z\right], \\
U_\beta^* U_\beta =& \mathrm{E}\left(R f_\mu^2\{Y|\mu(Z),\phi\} X^2 + (1-R)\mathrm{E}\left[f_\mu\{Y|\mu(Z),\phi\}X|Y,Z\right]^2\right) \\
=& \mathrm{E}\left[R f_\mu^2\{Y|\mu(Z),\phi\} X^2 + (1-R) f_\mu^2\{Y|\mu(Z),\phi\}\mathrm{E}\left(X|Z\right)^2\right] \\
=& \mathrm{E}\left[f_\mu^2\{Y|\mu(Z),\phi\}\mathrm{E}\left(X|Z\right)^2\right] + \mathrm{E}\left[R f_\mu^2\{Y|\mu(Z),\phi\}\mathrm{Var}\left(X|Z\right)\right], \\
U_\alpha^* U_\phi(h_1) =& \mathrm{E}\left\{f_\mu\{Y|\mu(Z),\phi\} f_\phi(h_1)\right\}, \\
U_\gamma^* U_\phi(h_1) =& \mathrm{E}\left\{f_\mu\{Y|\mu(Z),\phi\} Z f_\phi(h_1)\right\}, \\
U_\beta^* U_\phi(h_1) =& \mathrm{E}\left\{f_\mu\{Y|\mu(Z),\phi\}\mathrm{E}(X|Z) f_\phi(h_1)\right\}, \\
U_\phi^* U_\phi(h_1) =& f_\phi^* f_\phi(h_1), \\
U_\alpha^* U_\eta(h_2) =& \mathrm{E}\left\{\mathrm{E}\left(R f_\mu\{Y|\mu(Z),\phi\} + (1-R)\mathrm{E}\left[f_\mu\{Y|\mu(Z),\phi\}|Y,Z\right]|X,Z\right) h_2(X,Z)\right\} \\
=& \mathrm{E}\left(\mathrm{E}\left[f_\mu\{Y|\mu(Z),\phi\}|Z\right] h_2(X,Z)\right) = 0, \\
U_\gamma^* U_\eta(h_2) =& 0, \\
U_\phi^* U_\eta(h_2) =& 0, \\
U_\beta^* U_\eta(h_2) =& \mathrm{E}\left\{\mathrm{E}\left(R f_\mu\{Y|\mu(Z),\phi\} X + (1-R)\mathrm{E}\left[f_\mu\{Y|\mu(Z),\phi\}X|Y,Z\right]|X,Z\right) h_2(X,Z)\right\} \\
=& \mathrm{E}\left(\mathrm{E}\left[R f_\mu\{Y|\mu(Z),\phi\}|Z\right]\left\{X - \mathrm{E}(X|Z)\right\} h_2(X,Z)\right), \\
U_\eta^* U_\eta(h_2) =& \mathrm{E}(R|X,Z) h_2(X,Z) + \mathrm{E}\left[(1-R)\mathrm{E}\left\{h_2(X,Z)|Y,Z\right\}|X,Z\right]
\end{aligned}
$$

$$
\begin{aligned}
=&\mathrm{E}(R|X,Z)h_2(X,Z) + \mathrm{E}\left\{(1-R)|X,Z\right\}\mathrm{E}\left\{h_2(X,Z)|Z\right\} \\
=&\mathrm{E}(R|Z)\left[h_2(X,Z) - \mathrm{E}\left\{h_2(X,Z)|Z\right\}\right] + \mathrm{E}\left\{h_2(X,Z)|Z\right\},
\end{aligned}
$$

where f_ϕ^* is the adjoint operator of f_ϕ.

To calculate the efficient information of β, denoted by $I^{\beta\beta}$, we observe that the score spaces for (α, γ, ϕ) and η are orthogonal. Therefore, we have

$$
\begin{aligned}
I^{\beta\beta} =& U_\beta^* U_\beta - \langle M_1 U_\beta, U_\beta\rangle - \langle M_2 U_\beta, U_\beta\rangle \\
=& \mathrm{E}\left[f_\mu^2\{Y|\mu(Z),\phi\}\mathrm{E}\left(X|Z\right)^2\right] - \langle M_1 U_\beta, U_\beta\rangle - \langle M_2 U_\beta, U_\beta\rangle \\
&+ \mathrm{E}\left[R f_\mu^2\{Y|\mu(Z),\phi\}\mathrm{Var}\left(X|Z\right)\right],
\end{aligned}
$$

where

$$
M_1 = \begin{bmatrix} U_\alpha & U_\gamma & U_\phi \end{bmatrix}
\begin{bmatrix} U_\alpha^* U_\alpha & U_\alpha^* U_\gamma & U_\alpha^* U_\phi \\ U_\gamma^* U_\alpha & U_\gamma^* U_\gamma & U_\gamma^* U_\phi \\ U_\phi^* U_\alpha & U_\phi^* U_\gamma & U_\phi^* U_\phi \end{bmatrix}^{-1}
\begin{bmatrix} U_\alpha^* \\ U_\gamma^* \\ U_\phi^* \end{bmatrix},
$$

and $M_2 = U_\eta (U_\eta^* U_\eta)^{-1} U_\eta^*$ are the projection operators onto the score spaces of (α, γ, ϕ) and η, respectively. Let $I_0^{\beta\beta} \equiv \mathrm{E}\left[f_\mu^2\{Y|\mu(Z),\phi\}\mathrm{E}\left(X|Z\right)^2\right] - \langle M_1 U_\beta, U_\beta\rangle$. We observe that $I_0^{\beta\beta}$ is the efficient information for β in the regression model $P_{\theta,\phi}(Y|X,Z)$, except that X is replaced by $E(X|Z)$. Because $I_0^{\beta\beta}$ does not depend on R, it is invariant under any type of two-phase design.

Next, we calculate $\langle M_2 U_\beta, U_\beta\rangle$ as follows:

$$
\begin{aligned}
(U_\eta^* U_\eta)^{-1}(h_2) =& \mathrm{E}(R|Z)^{-1}\left[h_2(X,Z) - \mathrm{E}\left\{h_2(X,Z)|Z\right\}\right], \\
(U_\eta^* U_\eta)^{-1} U_\eta^*(h_2) =& h_2(X,Z) - \mathrm{E}\left\{h_2(X,Z)|Z\right\}, \\
U_\eta (U_\eta^* U_\eta)^{-1} U_\eta^*(h_2) =& R\left[h_2(X,Z) - \mathrm{E}\left\{h_2(X,Z)|Z\right\}\right],
\end{aligned}
$$

$$
\begin{aligned}
M_2U_\beta =& R\big(\mathrm{E}\left[Rf_\mu\{Y|\mu(Z),\phi\}|Z\right]X + (1-R)\mathrm{E}\left[f_\mu\{Y|\mu(Z),\phi\}|Z\right]\mathrm{E}(X|Z) \\
& - \mathrm{E}\left[Rf_\mu\{Y|\mu(Z),\phi\}X|Z\right] - \mathrm{E}\left[(1-R)f_\mu\{Y|\mu(Z),\phi\}\mathrm{E}(X|Z)|Z\right]\big) \\
=& R\mathrm{E}\left[Rf_\mu\{Y|\mu(Z),\phi\}|Z\right]X - R\mathrm{E}\left[Rf_\mu\{Y|\mu(Z),\phi\}X|Z\right] \\
& - R\mathrm{E}\left[(1-R)f_\mu\{Y|\mu(Z),\phi\}\mathrm{E}(X|Z)|Z\right] \\
=& R\mathrm{E}\left[Rf_\mu\{Y|\mu(Z),\phi\}|Z\right]\{X - \mathrm{E}(X|Z)\}, \\
\langle M_2U_\beta, U_\beta\rangle =& \mathrm{E}\big\{R\mathrm{E}\left[Rf_\mu\{Y|\mu(Z),\phi\}|Z\right]\{X-\mathrm{E}(X|Z)\} \\
& \times \left(Rf_\mu\{Y|\mu(Z),\phi\}X + (1-R)\mathrm{E}\left[f_\mu\{Y|\mu(Z),\phi\}X|Y,Z\right]\right)\big\} \\
=& \mathrm{E}\big(R\mathrm{E}\left[Rf_\mu\{Y|\mu(Z),\phi\}|Z\right]\{X-\mathrm{E}(X|Z)\}f_\mu\{Y|\mu(Z),\phi\}X\big) \\
=& \mathrm{E}\big(\mathrm{E}\left[Rf_\mu\{Y|\mu(Z),\phi\}|Z\right]^2\mathrm{Var}(X|Z)\big).
\end{aligned}
$$

Consequently,

$$
I^{\beta\beta} = I_0^{\beta\beta} + \mathrm{E}\left(\mathrm{E}\left[Rf_\mu^2\{Y|\mu(Z),\phi\} - \mathrm{E}\left[Rf_\mu\{Y|\mu(Z),\phi\}|Z\right]^2\middle|Z\right]\mathrm{Var}\left(X|Z\right)\right). \tag{5.40}
$$

The design is more efficient if the second term on the right-hand side of equation (5.40) is larger.

5.1.3 Simulation Studies

We conducted simulation studies to compare the efficiencies of different designs when the outcome is continuous. Specifically, we set Z to be a Bernoulli random variable with mean 0.5, and X to be a Bernoulli random variable with mean p_0 or p_1 depending on whether $Z = 0$ or 1. We generated the outcome from the linear model: $Y = \beta X + \gamma Z + \epsilon_1$, where ϵ_1 is a standard normal random variable independent of X and Z. We set $n = 4000$ and considered three types of two-phase designs. The ODS design selects 200 and 200 subjects with extremely large and small values of Y, respectively; the RDS design selects 200 and 200 subjects with extremely large and small values of $Y - \widehat{\gamma}Z$, respectively; and the optimal design

selects 200 and 200 subjects with extremely large or small values of $\sqrt{\text{Var}(X|Z)}(Y - \widehat{\gamma}Z)$, respectively, where $\text{Var}(X|Z = j) = p_j(1 - p_j)$, $j = 0, 1$. For benchmark comparisons, we included a forth design where the first-phase information is ignored and a simple random sample (SRS) of 400 subjects is selected.

The relative efficiencies between each of the three types of two-phase designs and the SRS design are shown in Table 5.28. We can see that all three designs are much more efficient than the SRS design. When Z has no effect on Y, the ODS design is as efficient as the RDS design. When Z has effects on Y, the RDS design is more efficient than the ODS design. When $\text{Var}(X|Z)$ is a constant, the RDS design is as efficient as the optimal design. When $\text{Var}(X|Z)$ depends on Z, the optimal design is substantially more efficient than the RDS design.

Table 5.28: Efficiency Comparisons Between the ODS, RDS, and Optimal Designs

			SE of $\widehat{\beta}$				RE		
p_0 and p_1	β	γ	SRS	ODS	RDS	Optimal	ODS	RDS	Optimal
$p_0 = p_1 = 0.7$	0.0	0.0	0.110	0.053	0.053	0.053	4.25	4.32	4.32
		0.5	0.110	0.059	0.053	0.053	3.47	4.32	4.32
		1.0	0.110	0.078	0.053	0.053	1.99	4.32	4.32
	0.3	0.0	0.108	0.054	0.054	0.054	4.01	3.98	3.98
		0.5	0.108	0.059	0.054	0.054	3.34	3.98	3.98
		1.0	0.108	0.080	0.054	0.054	1.82	3.98	3.98
$p_0 = 0.5$, $p_1 = 0.9$	0.0	0.0	0.125	0.060	0.060	0.056	4.29	4.29	5.05
		0.5	0.125	0.069	0.060	0.056	3.34	4.29	5.05
		1.0	0.125	0.093	0.060	0.056	1.80	4.29	5.05
	0.3	0.0	0.119	0.061	0.061	0.055	3.79	3.81	4.72
		0.5	0.119	0.070	0.061	0.055	2.86	3.81	4.72
		1.0	0.119	0.098	0.061	0.055	1.48	3.81	4.72

NOTE: SE is standard error; RE is the empirical variance of $\widehat{\beta}$ under the two-phase design over that under the SRS design.

5.2 Future Extensions

5.2.1 Efficient Inference Under General Two-Phase Sampling

In many epidemiological studies, the covariates of primary interest involve biochemical or genetic analysis of blood specimens or extraction of detailed exposure histories and thus are prohibitively expensive to measure in large studies. Two-phase designs that concentrate resources on where there is the greatest amount of information are extremely useful in this setting. The NHLBI ESP and CHARGE TSS are two recent examples. It is not hard to envision that many large-scale studies will adopt two-phase designs. I am in the progress of extending my research on efficient semiparametric inference under general two-phase sampling to different types of outcomes, including combinations of continuous and discrete outcomes, longitudinal outcomes, and censored time-to-event outcomes. I will study the theoretical properties and finite sample performance of each of these extensions. Another important direction worth pursuing is the analysis of secondary outcomes that are not used for sampling but are correlated with the primary outcome(s) used for sampling. To popularize our methods and facilitate broad collaborations, I will develop computationally efficient software packages that are capable of handling large datasets, including whole-exome and whole-genome sequencing studies.

5.2.2 Optimal Two-Phase Designs

I will continue my research on optimal two-phase designs. We have derived the efficient information for estimating the regression coefficients of the expensive covariates. We will use this general result to study the optimal design for a number of scenarios, including binary outcomes with inexpensive covariates, multiple continuous outcomes of equal interest, longitudinal outcomes with interest either in the baseline effect or the trend effect, and censored time-to-event outcomes. It would be also of interest to study the optimal design when the interaction effect between expensive and inexpensive covariates is of primary interest.

REFERENCES

Allison, D. B. (1997), "Transmission-Disequilibrium Tests for Quantitative Traits," *American Journal of Human Genetics*, 60, 676–690.

Anderson, J. (1972), "Separate Sample Logistic Discrimination," *Biometrika*, 59, 19–35.

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V., Folsom, A. R., Greenland, P., JacobsJr., D. R., Kronmal, R., Liu, K., Nelson, J. C., OâĂŹLeary, D., Saad, M. F., Shea, S., Szklo, M., and Tracy, R. P. (2002), "Multi-Ethnic Study of Atherosclerosis: Objectives and Design," *American Journal of Epidemiology*, 156, 871–881.

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*, 59, 447–461.

Breslow, N. E., McNeney, B., and Wellner, J. A. (2003), "Large Sample Theory for Semiparametric Regression Models with Two-Phase, Outcome Dependent Sampling," *The Annals of Statistics*, 31, 1110–1139.

Carroll, R. and Wand, M. (1991), "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society, Series B*, 53, 573–585.

Chatterjee, N. and Chen, Y. H. (2007), "A Semiparametric Pseudo-Score Method for Analysis of Two-Phase Studies with Continuous Phase-I Covariates," *Lifetime Data Analysis*, 13, 607–622.

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, 158–168.

Chen, Z., Zheng, G., Ghosh, K., and Li, Z. (2005), "Linkage Disequilibrium Mapping of Quantitative-Trait Loci by Selective Genotyping," *American Journal of Human Genetics*, 77, 661–669.

Dawber, T. R., Meadors, G. F., and Moore Jr, F. E. (1951), "Epidemiological Approaches to Heart Disease: the Framingham Study," *American Journal of Public Health and the Nations Health*, 41, 279–286.

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*, 39, 1–38.

Fried, L. P., Borhani, N. O., Enright, P., Furberg, C. D., Gardin, J. M., Kronmal, R. A., Kuller, L. H., Manolio, T. A., Mittelmark, M. B., Newman, A., O'Leary, D. H., Psaty, B., Rautaharju, P., Tracy, R. P., and Weiler, P. G. (1991), "The Cardiovascular Health Study: Design and Rationale," *Annals of Epidemiology*, 1, 263–276.

Friedman, G. D., Cutter, G. R., Donahue, R. P., Hughes, G. H., Hulley, S. B., Jr., D. R. J., Liu, K., and Savage, P. J. (1988), "CARDIA: Study Design, Recruitment, and Some Characteristics of the Examined Subjects," *Journal of Clinical Epidemiology*, 41, 1105–1116.

Grenander, U. (1981), *Abstract Inference*, New York: Wiley.

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, 663–685.

Hsieh, D. A., Manski, C. F., and McFadden, D. (1985), "Estimation of Response Probabilities From Augmented Retrospective Observations," *Journal of the American Statistical Association*, 80, 651–662.

Hu, Y. J., Lin, D. Y., and Zeng, D. (2010), "A General Framework for Studying Genetic Effects and Gene-Environment Interactions with Missing Data," *Biostatistics*, 11, 583–598.

Huang, B. E. and Lin, D. Y. (2007), "Efficient Association Mapping of Quantitative Trait Loci With Selective Genotyping," *American Journal of Human Genetics*, 80, 567–576.

Kalbfleisch, J. D. and Lawless, J. F. (1988), "Likelihood Analysis of Multi-State Models for Disease Incidence and Mortality," *Statistics in Medicine*, 7, 149–160.

Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999), "Semiparametric Methods for Response-Selective and Missing Data Problems in Regression," *Journal of the Royal Statistical Society, Series B*, 61, 413–438.

Li, B. and Leal, S. M. (2008), "Methods for Detecting Associations With Rare Variants for Common Diseases: Application to Analysis of Sequence Data," *American Journal of Human Genetics*, 83, 311–321.

Lin, D. Y. and Tang, Z. Z. (2011), "A General Framework for Detecting Disease Associations With Rare Variants in Sequencing Studies," *American Journal of Human Genetics*, 89, 354–367.

Lin, D. Y. and Zeng, D. (2006), "Likelihood-Based Inference on Haplotype Effects in Genetic

Association Studies," *Journal of the American Statistical Association*, 101, 89–104.

— (2009), "Proper analysis of secondary phenotype data in case-control association studies," *Genetic Epidemiology*, 33, 256–265.

Lin, D. Y., Zeng, D., and Tang, Z. Z. (2013), "Quantitative Trait Analysis in Sequencing Studies Under Trait-Dependent Sampling," *Proceedings of the National Academy of Sciences of the United States of America*, 110, 12247–12252.

Lin, H., Wang, M., Brody, J. A., Bis, J. C., Dupuis, J., Lumley, T., McKnight, B., Rice, K. M., Sitlani, C. M., Reid, J. G., Bressler, J., Liu, X., Davis, B. C., Johnson, A. D., O'Donnell, C. J., Kovar, C. L., Dinh, H., Wu, Y., Newsham, I., Chen, H., Broka, A., DeStefano, A. L., Gupta, M., Lunetta, K. L., Liu, C.-T., White, C. C., Xing, C., Zhou, Y., Benjamin, E. J., Schnabel, R. B., Heckbert, S. R., Psaty, B. M., Muzny, D. M., Cupples, L. A., Morrison, A. C., and Boerwinkle, E. (2014), "Strategies to design and analyze targeted sequencing data: cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Targeted Sequencing Study." *Circulation. Cardiovascular genetics*, 7, 335–43.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 44, 226–233.

Madsen, B. E. and Browning, S. R. (2009), "A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic," *PLoS Genetics* [online], 5, e1000384. Available at http://www.plosgenetics.o.

Magee, L. (1998), "Improving Survey-Weighted Least Squares Regression," *Journal of the Royal Statistical Society, Series B*, 60, 115–126.

Murphy, S. A. and van der Vaart, A. W. (2000), "On Profile Likelihood," *Journal of the American Statistical Association*, 95, 449–465.

Page, G. P. and Amos, C. I. (1999), "Comparison of Linkage-Disequilibrium Methods for Localization of Genes Influencing Quantitative Traits in Humans," *American Journal of Human Genetics*, 64, 1194–1205.

Pepe, M. and Fleming, T. (1991), "A Nonparametric Method for Dealing with Mismeasured Dovariate Data," *Journal of the American Statistical Association*, 86, 108–113.

Pfeffermann, D. and Sverchkov, M. (1999), "Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data," *Sankhya*, 61, 166–186.

Prentice, R. L. and Pyke, R. (1979), "Logistic Disease Incidence Models and Case-Control Studies," *Biometrika*, 66, 403–411.

Price, A. L., Kryukov, G. V., de Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L. J., and Sunyaev, S. R. (2010), "Pooled Association Tests for Rare Variants in Exon-resequencing Studies," *American Journal of Human Genetics*, 86, 832–838.

Reilly, M. and Pepe, M. S. (1995), "A Mean Score Method for Missing and Auxiliary Covariate Data in Regression Models," *Biometrika*, 82, 299–314.

Robins, J., 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*, 57, 409–424.

Rudin, W. (1973), *Functional Analysis*, New York: McGraw-Hill.

Schumaker, L. (1981), *Spline Functions: Basic Theory*, New York: Wiley-Interscience.

Scott, A. J. and Wild, C. J. (1986), "Fitting Logistic Models Under Case-Control or Choice Based Sampling," *Journal of the Royal Statistical Society, Series B*, 48, 170–182.

— (1997), "Fitting Regression Models to Case-Control Data by Maximum Likelihood," *Biometrika*, 84, 57–71.

Slatkin, M. (1999), "Disequilibrium Mapping of a Quantitative-Trait Locus in an Expanding Population," *American Journal of Human Genetics*, 64, 1765–1773.

Song, R., Zhou, H., and Kosorok, M. R. M. (2009), "On Semiparametric Efficient Inference for Two-Stage Outcome-Dependent Sampling with a Continuous Outcome," *Biometrika*, 96, 221–228.

Tang, Z. Z. and Lin, D. Y. (2013), "MASS: Meta-Analysis of Score Statistics for Sequencing Studies," *Bioinformatics*, 29, 1803–1805.

Tao, R., Zeng, D., Franceschini, N., North, K. E., Boerwinkle, E., and Lin, D. Y. (2015), "Analysis of Sequence Data Under Multivariate Trait-Dependent Sampling," *Journal of the American Statistical Association*, 110, 560–572.

Taylor Jr, H. A., Wilson, J. G., Jones, D. W., Sarpong, D. F., Srinivasan, A., Garrison, R. J., Nelson, C., and Wyatt, S. B. (2005), "Toward Resolution of Cardiovascular Health Disparities in African Americans: Design and Methods of the Jackson Heart Study," *Ethnicity and Disease*, 15, S6–4–S6–17.

The ARIC Investigators (1989), "The Atherosclerosis Risk in Communites (ARIC) Study: Design and Objectives," *American Journal of Epidemiology*, 129, 687–702.

The Women's Health Initiative Study Group (1998), "Design of the Women's Health Initiative Clinical Trial and Observational Study-Examples from the Women's Health Initiative," *Controlled Clinical Trials*, 19, 61–109.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.

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, 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, 119–128.

Whittemore, A. S. (1997), "Multistage Sampling Designs and Estimating Equations," *Journal of the Royal Statistical Society, Series B*, 59, 589–602.

Wild, C. J. (1991), "Fitting Prospective Regression Models to Case-Control data," *Biometrika*, 78, 705–717.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011), "Rare-Variant Association Testing for Sequencing Data With the Sequence Kernel Association Test," *American Journal of Human Genetics*, 89, 82–93.

Zeng, D. (2005), "Likelihood Approach for Marginal Proportional Hazards Regression in the Presence of Dependent Censoring," *The Annals of Statistics*, 33, 501–521.

Zeng, D. and Lin, D. Y. (2014), "Efficient Estimation of Semiparametric Transformation Models for Two-Phase Cohort Studies," *Journal of the American Statistical Association*, 109, 371–383.

Zhao, L. P. and Lipsitz, S. (1992), "Designs and Analysis of Two-Stage Studies," *Statistics in Medicine*, 11, 769–782.

Zhou, H., Weaver, M. A., Qin, J., Longnecker, M. P., and Wang, M. C. (2002), "A Semiparametric Empirical Likelihood Method for Data from an Outcome-Dependent Sampling Scheme with a Continuous Outcome," *Biometrics*, 58, 413–421.