View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Columbia University Academic Commons

A New Estimating Equation Based Approach for Secondary Trait Analyses in Genetic Case-control Studies

Xiaoyu Song

A thesis submitted in partial fulfillment for the

degree of Doctor of Public Health

in the

Mailman School of Public Health

Columbia University

Supervisor: Ying Wei and Iuliana Ionita-Laza

2015

@ 2015

Xiaoyu Song

All Rights Reserved

Abstract

A New Estimating Equation Based Approach for Secondary Trait Analyses in Genetic Case-control Studies

Xiaoyu Song

Background/Aims: Case-control designs are commonly employed in genetic association studies. In addition to the primary trait of interest, data on additional secondary traits, related to the primary trait, are often collected. Traditional association analyses between genetic variants and secondary traits can be biased in such cases, and several methods have been proposed to address this issue, including the inverse-probability-of-sampling-weighted (IPW) approach and semi-parametric maximum likelihood (SPML) approach.

Methods: Here, we propose a set of new estimating equation based approach that combines observed and counter-factual outcomes to provide unbiased estimation of genetic associations with secondary traits. We extend the estimating equation framework to both generalized linear models (GLM) and non-parametric regressions, and compare it with the existing approaches.

Results: We demonstrate analytically and numerically that our proposed approach provides robust and fairly efficient unbiased estimation in all simulations we consider. Unlike existing methods, it is less sensitive to the sampling scheme and underlying disease model specification. In addition, we illustrate our new approach using two real data examples. The first one is to analyze the binary secondary trait diabetes under GLM framework using a stroke case-control study. The second one is

to analyze the continuous secondary trait serum IgE levels under linear and quantile regression models using an asthma case-control study.

Conclusion: The proposed new estimating equation approach is able to accommodate a wide range of regressions, and it outperforms the existing approaches in some scenarios we consider.

Key words: secondary trait analysis; estimating equations; case-control studies; GWAS.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express the deepest appreciation to many people who made it possible for me to finish the thesis and complete the doctoral education.

First and foremost, I am very grateful to my thesis adviser Professor Ying Wei, who is the associate professor at Columbia University, Department of Biostatistics. I first had the chance to talk with Ying in 2010 when I worked on Rakai Youth Project. Ying provided me thoughtful and timely statistical guidance with lots of patience. This wonderful experience intrigued me to start my dissertation project with her, and Ying pleasantly surprised me with her profound knowledge on statistics and her kindness to students. Without her generous help, this thesis would not have been possible to be accomplished.

Then, I would like to express my deep gratitude to my co-advisor Iuliana Ionita-Laza, who is the assistant professor at Columbia University, Department of Biostatistics. Iuliana broadened my horizon on the keys issues in statistical genetics with many thought-provoking discussions. I thank her for her sage guidance, insightful criticisms, and patient encouragement.

I am also very much thankful to my oral/thesis committee Professor Ying Kuen K. Chueng, Mengling Liu, Istik Pe'er, and Shuang Wang. They all provided valuable suggestions to make the thesis thoroughgoing. I would like to thank Professor John Santelii and Sanyukta Mathur who supervised my research assisantship in the Department of Population and Family Health in the past three years. I am very grateful to the faculty, staff and fellow colleagues in the Biostatistics Department at Columbia. The comprehensive coursework, seminars and events offered by them have been key to my understanding of statistical research.

Last but not least, I owe my special thanks to my family. To my parents Shifeng Song and Yumei Chai, who cultivated my spirits to pursuit of scientific truth since I was a child. To my husband Dawei Liu, who provided his unconditional support for me to chase my dream. To my seven-month baby Chengyao Kyle Liu, who fulfilled my life with enjoys and surprises.

To my family

Chapter 1

Introduction

No two humans are genetically identical. This is true even for monozygotic twins who develop from the same zygote. The differences between individuals come from a variety of aspects, such as single nucleotide polymorphisms (SNPs), structural variation and epigenetics, to name a few. These genetic variations can affect how humans develop diseases and respond to pathogens, chemicals, drugs, vaccines and other agents. Understanding the variations in human genetics is important to detect, prevent and treat the diseases that are caused by genetic abnormalities and mutations. It is also especially critical for the development of personalized medicine that tailors health care for each individual patient.

One of the most commonly occurred variations throughout a person's DNA is the SNP that each person has on average roughly 10 millions SNPs across whole genome. A SNP is a single nucleotide (A, T, C or G) mutation at a specific locus of the DNA sequence between paired chromosomes or members of a biological species. Since humans are diploid organisms, the SNPs have two alleles (where the rare allele frequency is >1%) at each genetic locus, with one allele inherited from each parent. A single SNP may cause a Mendelian disease that follows a simple pattern of inheritance known as the Mendel's laws [Mendel, 1865]. Examples include sickle-cell anemia, Tay-Sachs disease, cystic fibrosis and xeroderma pigmentosa. Most of the Mendelian diseases have been well studied in the literature and the remaining challenges for current studies are the complex diseases, in which the SNPs do not usually function individually, but rather work in coordination with other

SNPs and the environment factors to manifest a disease condition.

An early statistical method to identify disease genes is linkage analysis. Linkage analysis uses marker data on individuals in families/pedigrees, and studies patterns of co-inheritance of the markers and the diseases throughout the pedigree. Although linkage analysis has successfully explained a lot of the Mendelian diseases, such as Huntington's disease and cystic fibrosis, it has been less successful for complex traits. One major reason is that complex diseases often have a large number of SNPs with small or medium effect sizes, and thus researchers need to collect a large number of families with several affected generations. If the disease is rare or having late-onset with a high mortality, finding families with more than one affected generation will be unpractical. Therefore, linkage studies are less helpful for complex traits, where multiple genes work together with small effect size in disease causation.

Unlike linkage studies, Genome-wide association studies (GWAS) allow researchers to identify the associations between the genetic markers and the complex diseases using unrelated individuals. GWAS first proposed by Risch et al. [1996] genotype each subject a dense set of pre-determined SNPs across the genome, and test for the disease-marker association at all SNPs. To carry out a GWAS, researchers use two groups of participants: people with the disease of interest and similar people without the disease. Each person gives a blood or buccal swab sample of DNA, from which millions of genetic variants are genotyped. If one type of the variant (one allele) is more or less frequent in people with the disease than other, then the SNP is said to be "associated" with the disease. The associated SNPs serve as powerful pointers to the region of the human genome where the disease-causing problem resides. However, the associated SNPs themselves may not directly cause the disease. They may just be "tagging along" with the actual causal variants. Researchers often need to take additional steps to identify the exact genetic change involved in the disease after GWAS. For example, researchers could sequence DNA base pairs in that particular region of the genome and conduct additional analysis.

GWAS have successful identified many genetic variations that contribute to a number of diseases [Visscher et al., 2012], such as type 2 diabetes, Parkinson's disease, heart

disorders, obesity, Crohn's disease and prostate cancer. For example, in 2005, three independent studies [Edwards et al., 2005; Haines et al., 2005; Klein et al., 2005] found that age-related macular degeneration, a common form of blindness, is associated with variation in the gene for complement factor H that regulates inflammation. Few previously thought that inflammation might contribute significantly to this type of blindness.

GWAS often use case-control design, and it offers tremendous savings in time and expense compared with a prospective design. Even so, case-control design remains costly, and therefore GWAS often collect rich information on additional traits to further improve the efficiency. The additional traits are mostly important factors associated with the primary diseases, including biomarkers, characterizations of the disease and anthropometric parameters. For example, in a chronic obstructive pulmonary disease study [Regan et al., 2010], the researchers also collected additional respiratory diseases such as asthma, emphysema and bronchitis.

In addition to the primary analysis, which focuses the association between the SNPs and case-control status, researchers are also interested in taking full advantage of the existing data and analyzing genetic associations with the additional traits. The analysis of the association between the SNPs and additional traits using existing case-control data is known as "secondary analysis" in the literature. The secondary analysis enables us to investigate the association between the common variants and secondary traits. Some of them help discovering the genetic pathways of the primary diseases, while others extend to different interest areas. For example, in Lettre et al. [2008], researchers analyzed height as the secondary trait using six GWA case-control studies focusing on diabetes, cardiovascular diseases and cancers. They identified ten SNPs and two previously reported SNPs strongly associated with height. These 12 SNPs together accounted for approximately 2% of the population variation in height, and encompassed both strong biological candidates and unexpected genes. They also highlighted several pathways (let-7 targets, chromatin remodeling proteins and Hedgehog signaling) as important regulators of human stature. No prior GWAS focusing on height has the power to detect these associations, and it showed the great value of secondary analysis in genetic studies.

Although conducting the secondary analysis is appealing, it is not straightforward to obtain an unbiased estimation of the association. In a simple case-control design, the cases are oversampled to improve the efficiency, and therefore the selected subjects are no longer representative of the general population. In particular, the subjects are ascertained by combining two randomly selected groups, the group of individuals with specified primary disease and the group without. When secondary traits are positively associated with the case-control status, subjects with the large secondary trait values are also oversampled; when negatively associated, subjects with the large secondary trait values are undersampled. As a result, the SNP-secondary trait association in the cases may differ from the controls. Ignoring the data structure and analyzing the SNP-secondary trait association using this case-control sample directly would lead to substantive biases. This statistical problem is further illustrated in the motivating examples in Section 1.1.

The existing methods can be broadly divided as three groups. First, one can use traditional methods in terms of direct regressions. The analysis can be done among case sample only, control sample only, combined case-control sample, or combined case-control sample adjusting the case-control status as a covariate. None of these traditional methods could estimate the association in the general population consistently. Second, one can correct the bias by using weights inverse to the probability of selection. This weighted approach has been long proposed and widely used in survey methods for its simplicity, but there are concerns over its efficiency. Third, one can explicitly account for the sampling scheme by modeling the retrospective likelihood function conditioning on the case-control status. A number of the articles based on the likelihood idea is available in the literature, and the semi-parametric maximum likelihood (SPML) approach proposed by Lin and Zeng [2009] is the most recognized for its large improvement in estimating efficiency. However, the method heavily relies on certain assumption on disease prevalence and is not robust against mis-specifications that are common in the GWAS studies. In addition, it introduces profile likelihood function in its estimation process to get rid of the high-dimensional nuisance parameters, which is computational intensive.

While GWAS mainly use parametric regressions at this stage for its simplicity, it is desir-

able to introduce non-parametric regressions to this field. In particular, we are interested to apply quantile regression [Koenker and Bassett Jr, 1978] as a way to systematically examine how the SNPs influence the location, scale, and shape of the entire trait distribution. Quantile regression allows the association between the risk factors and outcomes differ in different quantiles of the distribution, and therefore is especially useful when the risk factors are associated with the variances or the extreme values of the outcome. Quantile regression as well as other non-parametric regressions does not have parametric likelihood functions, and therefore could be applied with the likelihood based approaches such as SPML for the secondary analysis.

1.1 Motivating examples

In this section, we describe two real GWAS that motivate our research. One comes from the Risk Assessment of Cerebrovascular Events Study and the other is from the New York University Bellevue Asthma Registry.

1.1.1 Risk Assessment of Cerebrovascular Events (RACE) Study

Our first motivating example is a case-control GWAS, Risk Assessment of Cerebrovascular Events (RACE) Study [Cornelis et al., 2010], from dbGap as part of the Gene Environment Association Studies initiative funded by the trans-NIH Genes, Environment, and Health Initiative. This study included 1,220 cases with young onset stroke (stroke before age 60 years) in Pakistan and 1,273 controls from Pakistan Risk of Myocardial Infarction Study. For each study subject, the study also collects covariate information, including age, gender, ethnicity, diabetes, cardiovascular disease, myocardial infarction and tobacco usage. The study genotyped 657,366 genetic variants in the whole genome, including SNPs rs6712932 and rs1990760 that we are interested to investigate for their associations with diabetes.

Two previous studies have identified that SNPs rs6712932 and rs1990760 are associated with diabetes in white ethnicity European descents. In details, SNP rs6712932-G is reported to be a protective factor for type-2 diabetes with odds ratio (OR)=0.66 (CI: 0.54

- 0.79) [Salonen et al., 2007] in all white from eastern Finland, Israel, Germany and England, and SNP rs1990760-G is reported to be a protective factor for type-1 diabetes with OR=0.85 (CI: 0.81 - 0.90) [Todd et al., 2007] in self-reported white ethnicity in Great Britain. It is desirable to verify the association between the pre-reported SNPs and diabetes in different populations, as it would answer whether the associations are due to heterogeneity of the populations or disease mechanisms. We would like to re-investigate the associations in Pakistan population using the existing stroke case-control data.

Both types of diabetes are known to be risk factors for stroke [Peters et al., 2014; Sundquist and Li, 2006]. In this dataset, we only have information on whether a subject has diabetes without further specification on the type of diabetes. Applying simple logistic regression to the data, we estimate the OR for having young onset stroke associated with diabetes is 3.18 (p-value$<$ 0.0001). In addition, both SNPs are associated with primary disease (young onset stroke) with marginal per-minor-allele OR as 1.14 (p-value=0.024) and 0.87 (p-value=0.015), respectively. After adjusting for secondary trait (diabetes), the associations between these SNPs and stroke remain significant (rs6712932: OR=1.16; p- value=0.017. rs1990760: OR=0.84; p-value=0.003). This is a situation where the commonly-used estimation methods may provide biased estimation for the association between these two SNPs and diabetes. We could observe from Table 1.1 that the estimates from cases and controls are different. Although we are unable to observe the true coefficient in the population, we would expect it is closer to the ones among controls than the ones among cases, since the most of the subjects in the population are healthy people. Combining the case-control samples and regressing with or without adjusting for disease status gave us estimates that are close to the cases, so we believe they are likely to be biased.

	rs6712932				rs1990760			
	Est $\hat{\beta}_1$	OR	SE	p-value	Est $\hat{\beta}_1$	OR	SE	p-value
Case	-0.100	0.90	0.092	0.2800	-0.112	0.89	0.088	0.2024
Control	-0.125	0.88	0.132	0.3441	-0.247	0.78	0.118	0.0368
CC	-0.060	0.94	0.074	0.4166	-0.126	0.88	0.068	0.0648
Adj CC	-0.108	0.90	0.076	0.1530	-0.159	0.85	0.071	0.0246

Table 1.1: The association between SNPs and diabetes in a young onset stroke case-control sample. "Case" stands for logistic regression among case sample only. "Control" stands for logistic regression among contol sample only. "CC" stands for unadjusted logistic regression using both case and control samples. "Adj CC" stands for logistic regression using both case and control samples adjusting for primary disease status.

1.1.2 New York University Bellevue Asthma Study

Another motivating example is an association study of the Thymic stromal lymphopoietin (TSLP) gene and asthma from the New York University Bellevue Asthma Registry (NYUBAR) [Liu et al., 2011]. Asthma is a common chronic inflammatory disease of the airways characterized by variable and recurring symptoms, reversible airflow obstruction and bronchospasm. Asthma is thought to be caused by a combination of genetic and environmental factors and is usually diagnosed based on the pattern of symptoms, response to therapy over time and spirometry. TSLP gene, viewed as a "master switch" of allergic inflammation at the epithelial cell and dendritic cell interface, is upregulated in asthma. In their primary analysis, ten tag-SNPs in the TSLP gene were analyzed for association with asthma using 387 clinically diagnosed asthmatic cases and 212 healthy controls. One SNP (rs1898671) showed nominally significant association with asthma (OR = 1.50; 95% CI: 1.09 - 2.05, p = 0.01) after adjusting for age, BMI, income, education and population stratification.

In this study, we are interested to understand the mechanical pathways of TSLP gene in affecting the occurrence of asthma. Asthma is almost surely to have allergic basis that it is very likely to be associated with some type of Immunoglobulin E (IgE) related reaction. The IgE is a class of antibody that mediates the immune responses in the pathogenesis of allergic asthma [Burrows et al., 1989]. It binds to allergens and triggers the release of substances from mast cells that can cause inflammation. In addition to asthma, IgE is also associated with other allergic diseases, such as allergic rhinitis, peanut allergy, latex sensitivity, atopic dermatitis, chronic urticaria and allergic bronchopulmonary aspergillosis [Morjaria and Polosa, 2009]. To further understand the genetic basis of asthma, we would like to investigate the association between TSLP gene with serum IgE level to uncover the mediation pathways. It is also helpful to understand the impact of TLSP gene on other allergic diseases. Figure 1.1 shows the distribution of log serum IgE levels by case-control status. According to the figure, the log serum IgE levels are approximately normally distributed among cases and controls, and therefore we can apply least square regression to analyze the mean genetic association with serum IgE levels. In addition, since high not

the average serum IgE level is an indicator of allergic diseases, we also want to consider quantile regression for the genetic association with upper quantiles of IgE in the analysis.

Based on Figure 1.1, asthma is associated with serum IgE levels that cases are more likely to have high serum IgE levels than healthy controls. This association is clinical and statistical significant that on average the OR of having asthma with one unit increase in log serum IgE level is 1.41 (p-value<0.0001). When the secondary trait (serum IgE level) is associated with primary disease (asthma), the direct analysis using the case-control data may be biased due to its oversampled cases from the population. Table 1.2 illustrates this problem by summarizing the association between the ten-tag SNPs in TSLP gene and serum IgE level separately in cases and controls. For example, we observe the genetic associations of SNP rs10035870 with log IgE level among cases and controls are very different. Combining cases and controls with an arbitrary proportion invoke substantive biases. Therefore, there is a need to utilize novel statistical methods to adjust for the biases, and this novel method should be able to facilitate quantile regression that does not based on likelihood functions in the analysis.

		Mean		$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.85$	
SNPs	Sample	Est	P-val	Est	P-val	Est	P-val	Est	P-val
rs2289276	Case	0.0	0.913	0.0	0.760	0.1	0.472	0.1	0.756
	Control	0.0	0.866	0.1	0.505	0.2	0.318	-0.2	0.472
rs1898671	Case	-0.2	0.052	-0.2	0.182	-0.2	0.085	-0.2	0.187
	Control	-0.2	0.219	-0.5	0.065	0.0	0.982	-0.2	0.363
rs11466741	Case	-0.1	0.557	0.0	0.764	0.1	0.500	0.0	0.942
	Control	0.2	0.212	0.3	0.088	0.4	0.068	0.2	0.322
rs11466743	Case	0.3	0.473	0.0	0.979	0.1	0.841	0.7	0.646
	Control	-0.5	0.275	0.0	0.904	-0.8	0.064	-1.0	0.003
rs2289277	Case	0.0	0.789	-0.1	0.525	0.1	0.387	0.1	0.507
	Control	0.1	0.515	0.1	0.638	0.3	0.105	0.2	0.254
rs2289278	Case	0.3	0.107	0.2	0.490	0.4	0.066	0.2	0.197
	Control	-0.3	0.294	-0.2	0.470	-0.5	0.275	-0.4	0.444
rs11241090	Case	0.4	0.125	0.4	0.107	0.1	0.779	0.6	0.339
	Control	0.3	0.355	0.3	0.416	-0.1	0.842	0.6	0.489
rs10035870	Case	-0.1	0.579	0.0	0.987	-0.1	0.657	-0.1	0.668
	Control	0.9	0.011	0.9	0.207	0.9	0.000	0.5	0.130
rs11466749	Case	0.2	0.414	-0.1	0.777	0.4	0.223	0.6	0.079
	Control	0.0	0.958	-0.1	0.760	0.1	0.827	0.3	0.478
rs11466750	Case	0.2	0.132	0.1	0.541	0.4	0.114	0.6	0.040
	Control	0.0	0.818	-0.1	0.695	-0.2	0.625	0.2	0.718

Table 1.2: The association of ten-tag TSLP SNPs and log serum IgE levels in an asthmatic case-control sample. "Mean" stands for linear regression. "τ" stands for quantile regression at the τth quantile. "Case" stands for regressions among case sample only. "Control" stands for regressions among contol sample only.

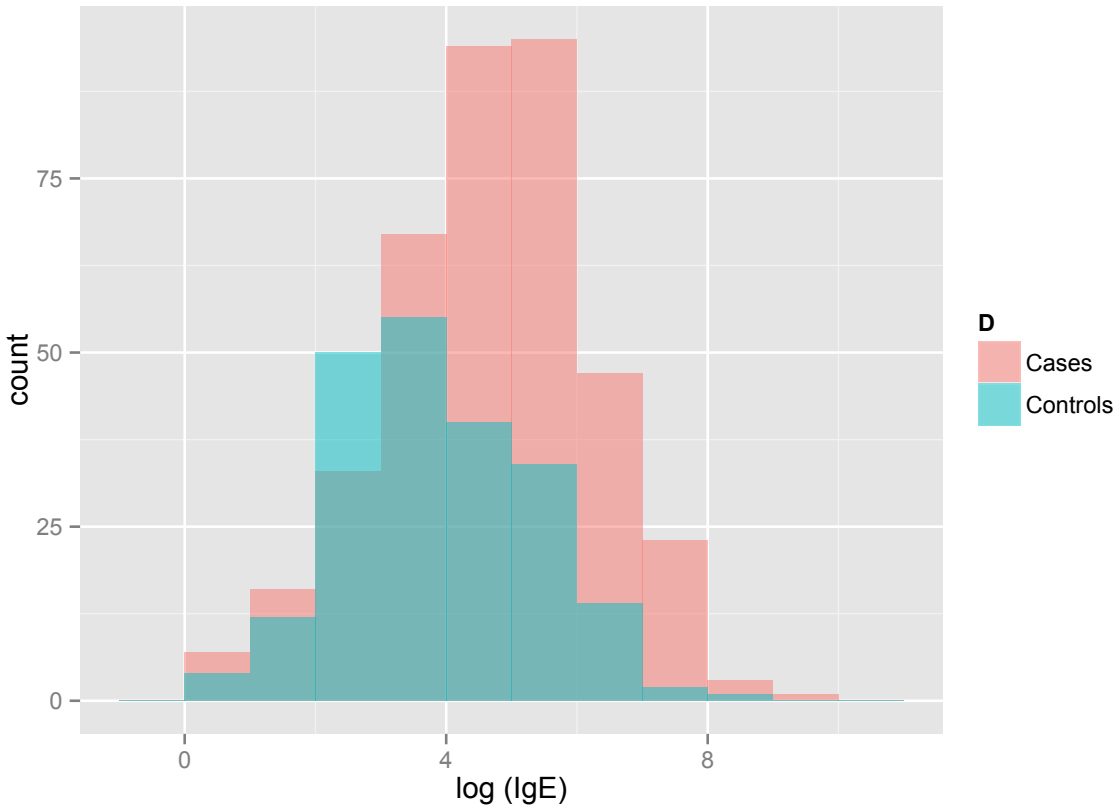


Figure 1.1: The log serum IgE levels in Asmthatic cases and controls

1.2 Our contribution

In summary, the aforementioned situations have two problems. First, although an extensive literature have been found on the secondary analysis, there is no approach that is robust to most of the situations in genetic studies and fairly efficient in identifying the SNPs. Second, when the secondary traits are continuous, the researchers mainly transform the outcome to approximately normal distribution and use linear regressions. It ignores the potential non-linear associations and the associations in other quantities than means.

To address the aforementioned questions, we proposed a new estimating equation based approach for the analysis of secondary traits in the genetic case-control studies. Our contributions are two-fold. First, the proposed approach balances the robustness and efficiency for the secondary analysis. In particular, it has very similar point estimates to the most robust approach in the literature with smaller standard errors. Second, we generalize the secondary analysis to the quantile regression, which has great potentials to deepen and expand the existing knowledge on traditional secondary analysis, and therefore discover additional candidate SNPs for the complex diseases.

1.3 Structure of this dissertation

The rest of the thesis is organized as follows. In Chapter 2, we first review the existing approaches for the secondary analysis in genetic case-control studies. It includes discussing the conditions that traditional methods are able to work, and describing two most popular novel methods. One is widely used for its robustness, and the other is for its efficiency. In Chapter 3, we proposed a new estimating equation based approach that provides a general framework for a wide range of regressions in secondary analysis. An estimation algorithm for the model parameters is described. The bootstrap method for confidence interval and hypothesis tests is proposed. In Chapter 4, a series of simulation studies is conducted to evaluate the performance of the proposed estimation equations in generalized linear models (GLM) and quantile regression in finite sample sizes. Its performance is compared with proper existing methods. The Type I error and the model robustness under mis-

specification of proposed approach is investigated. In Chapter 5, we applied the proposed estimating equations with existing approaches to the Risk Assessment of Cerebrovascular Events Study and the New York University Bellevue Asthma Study mentioned earlier. In Chapter 6, we summarize the important findings in previous chapters and discuss some future directions of the study.

Chapter 2

Review of secondary analysis in genetic case-control studies

The secondary analysis of the genetic case-control studies is an important topic which has received considerable attention in recent years. The existing methods can be broadly divided as three groups.

First, one can use traditional methods in terms of standard regressions to analyze the marker-secondary trait associations. The analysis can be done among cases only, controls only, entire sample ignoring the case-control status, entire sample using case-control status as a covariate. None of these traditional methods is able to provide unbiased estimation of marker-secondary trait associations, because cases and controls are selected at different rates from their respective subpopulations. The case-control sample does not constitute a random sample of the general population. As a result, the population association between a SNP and a secondary trait can be distorted in the case-control sample.

Second, one can correct the bias using weighting schemes originally developed from sampling schemes [Jiang et al., 2006; Monsees et al., 2009; Richardson et al., 2007; Scott and Wild, 2002]. The inverse-probability-of-sampling-weighted (IPW) regression, also know as survey-weighted approach, uses weights inversely proportional to the sampling fractions to the analysis of the secondary traits. This approach provides robust estimation for the marker-secondary trait associations but lacks of the efficiency. Technically, this

approach requires knowledge of the case-control sampling fractions, so it is proposed in a case-control study nested within a big cohort study. However, in reality, the sampling scheme is often not clear, and researchers sometimes use the disease prevalence as an approximation of the sampling scheme.

Third, one can explicitly account for the case-control sampling scheme via maximizing the retrospective likelihood function conditioning on the sampling scheme [He et al., 2011; Jiang et al., 2006; Lee et al., 1997; Lin and Zeng, 2009; Scott and Wild, 2001]. The semi-parametric maximum likelihood (SPML) method proposed by Lin and Zeng [2009] is the most widely recognized approach using this idea. This method made linear logit assumption for disease probability relating secondary trait and SNPs, and estimated the coefficients by maximizing the retrospective likelihood function conditionally on sampling scheme. This approach largely improves the efficiency of the estimations from IPW, but when the model assumptions are violated, the resulting estimates could be biased. For this paper, we mainly review for the SPML approach on behalf of the retrospective likelihood based methods in the following section.

Other methods based on similar idea of likelihood functions are not as widely applied in the data analysis as SPML method for different reasons. For example, the bias correction method [Wang and Shete, 2011, 2012] only deals with binary secondary traits and is unable to adjust for covarites. The adaptive weighted approach [Li and Gail, 2012] has difficulties to deal with additive genetic models. Therefore,

In addition to the SPML approach we will review later, there are other approaches available in the literature based retrospective likelihood functions. They are not as widely applied in the data analysis as SPML method for different reasons. For example, Li and Gail [2012] proposed a adaptive weighted approach to weighted sum two estimates to improve the robustness of the SPML approach. The first estimate is from SPML method, and the second one follows the same structure of SPML approach but revises $P(D|X,Y)$ model to add the $X-Y$ interaction. It is designed to put more weight on SPML estimators when there is no interaction between X and Y in predicting the primary disease to improve efficiency, and put less weight on it when there is an interaction effect to improve the ro-

bustness. However, simulations in Wang and Shete [2012] showed that this approach no longer provides unbiased estimation for $X - Y$ association, it losses most of the efficiency in the SPML approach, and finally it has difficulties to hand additive genetic models. Wang and Shete [2011] applied a method of moments approach to produce bias-corrected odds ratio estimates for binary secondary traits using prevalence estimates for the primary and secondary traits from the literature. Later, they modified their method to add in an interactive effect of the $X - Y$ on the primary disease risk [Wang and Shete, 2012]. The original version demonstrated the same efficiency as SPML, and same as SPML, it does not withstand mis-specified $P(D|X,Y)$ model assumptions. The modified version improves the robustness but also losses the efficiency of the original version. To make it worse, this method requires external information on secondary trait prevalence, it has difficulties to hand covariates, which largely narrows its applications. He et al. [2011] proposed a Gaussian copula-based approach that models the joint distribution in terms of the marginals for the primary and secondary phenotypes and uses the multivariate normal distribution to build in correlation between the phenotypes. Their method can handle multiple correlated secondary phenotypes, but they did not improve the efficiency of Lin and Zeng [2009]'s SPML method. Because of these reasons, we do not review theses methods in details in this paper. In case of interest, one can review their original articles [He et al., 2011; Li and Gail, 2012; Wang and Shete, 2011, 2012].

For the notation in the literature review, we let X denote the genotype score for an SNP of interest, Y denote the secondary phenotype, and D={0, 1} denote the primary case-control status. In a case-control dataset, the data at a single variant consists of n_1 cases $\{x_i, y_i, d_i = 1\}$, $i = 1, 2, ..., n_1$, and n_0 controls $\{x_i, y_i, d_i = 0\}$, $i = n_1 + 1, n_1 + 2, ..., n_1 + n_0$. We denote $n = n_1 + n_0$ as the total sample size.

2.1 Traditional approaches

While the real interest is in $P(Y|X)$ in the general population, the traditional methods described below are attractive because they can be performed using long-established standard

software, and because they require less model building than more efficient and theoretically justified approaches. Therefore, it is of interest to determine the situations in which they might be expected to work adequately.

Four types of traditional methods have been conducted to assess the effects of SNPs on secondary traits using data from case-control association studies: (1) cases only; (2) controls only; (3) combined sample of cases and controls; (4) joint analysis of cases and controls adjusted for the disease status. Methods (1) and (2) are restricted to controls and cases, respectively. Method (3) ignores the sampling scheme and analyzes cases and controls together. Method (4) analyzes cases and controls together and includes the disease status as a covariate in the model.

If the secondary phenotype is not related to the case-control status, or more precisely, D is independent of Y given X, then all four methods are valid. If the SNP is not associated with the case-control status, or more precisely, D is independent of X given Y, then all four methods yield correct estimates from the logistic regression for dichotomous traits except for the intercept, but the least-squares estimates for quantitative traits produced by the four methods are biased unless [Nagelkerke et al., 1995]. When the disease is rare, all standard methods except (1) are approximately valid. However, how rare must it be for a "rare disease" is unclear to the researchers. It is also clear that method (1) and (2) are inefficient because they involve discarding the part of the data. Another problematic situation for method (2) is where an exposure is very rare for controls but rather more common amongst cases.

In GWAS, most of the secondary traits collected are strongly correlated to the disease risk to improve efficiency, and therefore any SNPs that are associated with the case-control status will tend to be detected as being associated with secondary traits by standard methods even when the latter associations do not exist. It is true that the majority of tested SNPs in genome are not associated with disease risk, so a standard prospective regression model provides valid tests of and nearly unbiased estimates of marker-secondary trait association. However, when the associations truly exist, all four methods may produce estimates that are biased toward the null and thus reduce statistical power. These biases have been

demonstrated in previous researches in Jiang et al. [2006]; Lin and Zeng [2009]; Monsees et al. [2009]; Richardson et al. [2007]. We also illustrated this problem in motivating examples in Section 1.1.1 and 1.1.2.

2.2 Inverse-probability-of-sampling-weighted (IPW) approach

The inverse-probability-of-sampling-weighted (IPW) approach is widely used in the field of secondary analysis for its simplicity. It takes the contributions to the score equations for fitting a model to prospective data and weight them inversely to their probabilities of selection. For this particular problem, we solve

$$\sum_{i=0}^{1} w_i \sum_{j:D_j=i} S(\boldsymbol{\beta}; Y_j|\mathbf{X}_j) = \mathbf{0}$$

where $S(\boldsymbol{\beta}; Y_j|\mathbf{X}_j)$ is the score function of Y given X, and w_i is a weight inverse to their probabilities of selection. In a random sample of cases and controls, the w_i is $1/P(D = i)$, and in more complicated sampling designs or post-stratification, w_i is a consistent estimator thereof. Asymptotic variance estimation from linearization involves a sandwich estimator of the type common to estimating equation methods. Several packages in SAS and R are available for the implementation of IPW estimation with appropriate variance correction via the weight statement in PROC GENMOD and weights option in the geeglm() function.

The IPW approach provides unbiased estimates of genotype-secondary trait association even when both the genotype and secondary trait are independently associated with primary disease. It can accommodate various types of phenotypes (e.g., binary, continuous and ordinal) and different models for SNPs (additive, dominant or recessive). It can easily accommodate covariates, including population substructure, an important confounding in genetic studies. However, the IPW approach is sensitive to sampling schemes. The resulting estimates may not be efficient, especially under complex sampling schemes where the sampling variables are unrelated to the disease status.

2.3 Semi-parametric maximum likelihood (SPML) approach

In the SPML approach by Lin and Zeng [2009], they use a generalized linear model to formulate the effects of X on Y, and write the conditional density of Y given X as $P(Y|X)$. If Y is a quantitative trait, they use the linear regression model, and if Y is a dichotomous trait, they use the logistic regression model, under which

$$P(Y=1|X)=\frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}.$$

In addition, they made the assumption that $P(D|X,Y)$ follows the logistic regression model as follows

$$P(D=1|X,Y)=\frac{\exp(\gamma_0+\gamma_1 X+\gamma_2 Y)}{1+\exp(\gamma_0+\gamma_1 X+\gamma_2 Y)}.$$

Because the sampling is conditional on the case-control status, the likelihood function takes the retrospective form that

$$\begin{aligned}
&\prod_{i=1}^{n} P(X_i,Y_i \mid D_i)\\
=&\prod_{i=1}^{n}\{\frac{P(D_i=1\mid X_i,Y_i)P(Y_i\mid X_i)P(X_i)}{P(D_i=1)}\}^{D_i}\{\frac{P(D_i=0\mid X_i,Y_i)P(Y_i\mid X_i)P(X_i)}{P(D_i=0)}\}^{1-D_i}\\
=&\prod_{i=1}^{n}\{\frac{P(Y_i|X_i)p(X_i)\exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_i)/(1+\exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_i))}{P(D_i=1)}\}^{D_i}\\
&\times\{\frac{P(Y_i|X_i)p(X_i)/(1+\exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_i))}{P(D_i=0)}\}^{1-D_i}
\end{aligned}$$

Writ $p_i = P(X_i)$, and $\sum_{i=1}^{n} p_i = 1$. The p_i is a nuisance parameter that they treat non-parametrically and it has potentially high dimensions. They use the profile likelihood approach to profile out this parameter to improve the computational efficiency. Let the disease prevalence to be known as ξ, maximizing the retrospective likelihood function is equivalent as maximizing the following function

$$L=\prod_{i=1}^{n}\left\{P(Y_i|X_i)p_i\frac{\exp(D_i(\gamma_0+\gamma_1 X_i+\gamma_2 Y_i))}{1+\exp(\gamma_0+\gamma_1 X_i+\gamma_2 Y_i)}\right\}$$

subject to two constraints that (1) $\sum_{i=1}^{n} p_i = 1$ and (2)

$$\xi = \sum_{i=1}^{n} p_i \int_y P(y|X_i) \frac{\exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i)}{1 + \exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i)} dy.$$

By Using Largrange multiplier λ, they see that the estimate for p_i satisfies

$$\frac{\partial \log L}{\partial p_i} = \frac{1}{p_i} - \lambda_1 \int_y P_\theta(y|X_i) \frac{\exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i))}{1 + \exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i)} dy - \lambda_2 = 0.$$

Multiplying the above equation by p_i and summing over i, we see $\lambda_1 \xi + \lambda_2 = n$. Therefore, the above equation is equivalent to

$$p_i = \{\lambda_1 \int_y P_\theta(y|X_i) \frac{\exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i))}{(1 + \exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i))} dy - (n - \lambda_1 \xi)\}^{-1},$$

where λ_1 satisfies $\sum_{i=1}^{n} p_i = 1$. Thus, the profile log-likelihood function for β, γ_0, γ_1 and γ_2 is

$$\begin{aligned} l = logL = \sum_{i}^{n} \{ & \log P(Y_i|X_i) + D_i(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i) - \log(1 + \exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i)) \\ & - \log(\lambda_1 \int_y P_\theta(y|X_i) \frac{\exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i))}{(1 + \exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i))} dy - (n - \lambda_1 \xi))\}, \end{aligned}$$

where λ_1 is determined by the equation

$$\sum_{i=1}^{n} \{\lambda_1 \int_y P_\theta(y|X_i) \frac{\exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i))}{(1 + \exp(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i))} dy - (n - \lambda_1 \xi)\}^{-1} = 1.$$

They maximize the profile log-likelihood function by the Newton-Raphson algorithm or optimization algorithms. Likelihood-based statistics (i.e., Wald, score and likelihood-ratio statistics) can be used to make inference about the parameter of main interest β_1.

There are a number of attractive features for the SPML approach. First, when the model is correctly specified, the SPML approach is by far the most efficient method in estimating the association between Y and X in the general population. In addition, this approach is

applicable to both binary and continuous phenotypes, and handle covariates in a flexible manner. The major disadvantage is that when the disease model is misspecified, the estimation for the association between Y and X is largely biased. One minor disadvantage is that it is very computationally intensive, especially when considering the continuous covaraites **Z**. The probability distribution of the continuous covariates will enter the likelihood function as a high-dimenstional nuisance parameter, and the the profile-likelihood approach to eliminate of such nuisance parameters is difficult. A computer program SPREG is available online http://dlin.web.unc.edu/software/spreg-2/ to perform logistic and linear regression analysis of secondary trait data in case-control association studies. In application, however, it sometimes fails to generate an estimate due to algorithm problems in their software (further investigated in Section 5.2).

2.4 Quantile regression

Quantile regression first proposed by Koenker and Bassett Jr [1978] is a type of non-parametric regressions estimating functional relationship between variables for all portions of a probability distribution. While least square regression estimates the conditional mean of the response variable given certain values of the predictor variables, quantile regression aims at estimating either the conditional median or other quantiles of the response variable. For any real-valued random variable Y with cumulative distribution function $F(y) = P(Y \leq y)$, the τth **quantile** of Y is defined as the inverse function

$$Q_\tau(Y) = F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\},$$

where $0 < \tau < 1$. The median $Q_{1/2}(Y)$ plays the central role. The loss of quantiles is defined by the piecewise linear function

$$\rho_\tau(u) = (\tau - I(u < 0))u$$

for some $\tau \in (0,1)$ as illustrated in Figure 2.4. A specific quantile can be found by minimizing the expected loss $E\rho_\tau(Y - \widehat{y})$.

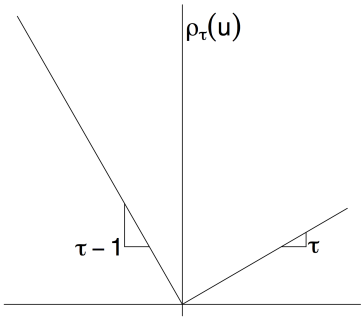


Figure 2.1: Quantile regression ρ function.

For a random sample $\{y_1, ..., y_n\}$ of Y, the problem of finding the τth **sample quantile** α_τ may be formulated as the solution of this optimization problem

$$\min_{\alpha \in \mathscr{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha_\tau).$$

Linear quantile regression

Quantiles efficiently describe marginal distribution and minimize asymmetric linear loss. This leads to the more general methods of estimating models of *conditional* quantile functions. Least squared regressions offer a template for this development. Knowing that the sample mean solves the problem

$$\min_{\mu \in \mathscr{R}} \sum_{i=1}^{n} (y_i - \mu)^2$$

suggest that, if we are willing to express the *conditional* mean of y given x as $\mu(x) = x^T\boldsymbol{\beta}$, then $\boldsymbol{\beta}$ may be estimated by solving

$$\min_{\beta \in \mathscr{R}^p} \sum_{i=1}^{n} (y_i - x_i^T\boldsymbol{\beta})^2.$$

Similarly, since the τ sample quantile, $\widehat{\alpha}_\tau$, solves

$$\min_{\alpha \in \mathscr{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha_\tau)$$

we are led to specifying the τth *conditional* quantile function as $Q_\tau(\tau|x) = x^T\boldsymbol{\beta}_\tau$, and to consider $\widehat{\boldsymbol{\beta}}_\tau$ solving

$$\min_{\beta\in\mathcal{R}^p}\sum_{i=1}^{n}\rho_\tau(y_i - x_i^T\boldsymbol{\beta}_\tau)$$

for any quantile $\tau \in (0,1)$. A specific regression quantile $\widehat{\boldsymbol{\beta}}_\tau$ can be found by minimizing the expected loss of $(Y - X^T\boldsymbol{\beta}_\tau)$ with respect to $\boldsymbol{\beta}_\tau$. We minimize the expected loss function by taking the first derivative of it, which generates the estimating equation for the quantile regression. We let $S_\tau(\mathbf{X}, Y, \boldsymbol{\beta}) = [\tau - I\{Y \leq \mathbf{X}^T\boldsymbol{\beta}\}]\mathbf{X}$ be the set of quantile regression estimating functions. The estimated $\widehat{\boldsymbol{\beta}}_\tau$ minimizes the absolute value of the estimating function that

$$\widehat{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta}} E_Y[\|S_\tau(\mathbf{X}, Y, \boldsymbol{\beta}_\tau)\| \mid \mathbf{X}] = 0.$$

Quantile regression does not have a parametric likelihood function, and its optimization is achieved through linear programs, as it can be written as linear function subject to linear constraints. Statistical software packages are available to conduct quantile analysis such as quantreg in R and proc quantreg in SAS.

Quantile regression results offer a much richer, more focused view of the applications than could be achieved by looking exclusively at conditional mean models. First of all, it offers a systematic strategy for examining how covariates influence the location, scale, and shape of the entire response distribution. In addition, its estimates are more robust against outliers in the response measurements in comparison with least squared regressions. Quantile regression is widely used in biostatistics field for various reasons. One, the vulnerable or high risk group to certain disease often consists of subjects with high or low values for their quantitative traits. For example, people with high body mass index (BMI) are predisposed to diabetes, cancers and many other disorders [Hjartåker et al., 2008]. Therefore, instead of examining risk factors for the mean of BMI, it is practically meaningful to investigate the risk factors for the upper quantiles of BMI, which are directly associated with high risk for many disease. Second, many studies also observe that the covariate effect

varies across quantile levels. For example, Yang et al. [2012] found that an important genotype FTO is not only associated with the mean of BMI [Frayling et al., 2007] but also with the variance, suggesting that the FTO genotype influences the entire distribution of BMI and impacts differently at various quantiles. As a result, examining covariate effect at certain or multiple quantiles provides a more comprehensive view of association between genetic markers and traits. For these reasons, quantile-based analyses have great potential to deepen and expand the existing knowledge from traditional secondary analysis.

2.5 Major deficiencies of existing approaches

Despite all of the efforts in secondary analysis, a number of deficiencies remain. First, the performance of the existing methods depends heavily on knowing either the correct sampling scheme or the $P(D|X,Y)$ in the population. However, most case-control studies are not nested within a larger cohort with clear selection probabilities, and also the underlying disease model is often unknown. Therefore, there is a need to propose new methods that are valid and robust in analyzing secondary phenotypes in case-control association studies with limited information on the sampling scheme and underlying disease model. The proposed methods should be generally applicable to a wide range of genetic models (dominant, recessive and additive), adjust for covariates easily, handle multiple types of regressions, relax the common conditions of the disease prevalence models, and be computationally simple and easy to implement.

In addition, extra challenges arise from secondary analysis in non-paramtetric regressions with no likelihood functions, such as quantile regression. The likelihood functions based approaches mentioned above focus on covariate effects on the mean of secondary traits, which is only one measure of the central tendency of the outcome, and often require parametric distribution assumption on the secondary traits. As we have noticed the attractive features of quantile regression in genetic studies, there is a great need to propose new methods that could extend quantile regression techniques to estimate the conditional quantiles of the secondary traits in genetic case-control studies.

Chapter 3

New estimating equation approach

3.1 Notations and settings

Let X denote a genetic variant of interest, Y denote a secondary phenotype, $\boldsymbol{Z}$ denote the vector of covariates we want to adjust for, and D={0, 1} denote the primary disease status. The aim of the secondary trait analysis is to estimate the genetic effect of X onto Y in a general population. A commonly used model can be written as

$$g(Y) = \beta_0 + X\beta_1 + \boldsymbol{Z}^T\boldsymbol{\beta_2}, \tag{3.1}$$

where $g(\cdot)$ is a link function, and β_1 is the coefficient of primary interest. Depending on the choices of the link function g, Model (3.1) covers a wide range of regressions. If g is an identity link for continuous outcome that $g(Y) = E[Y \mid X, \mathbf{Z}]$, then Model (3.1) is a mean regression; if g is a logit link for binary outcome that $g(Y)$ =logit $P(E[Y] = 1 \mid X, \mathbf{Z})$, then Equation (3.1) is a logistic regression; if g is a quantile function for quantitative outcome at the τth quantile that $g_\tau(Y) = Q_Y(\tau \mid X, \mathbf{Z})$, then Equation (3.1) is in the form of quantile regression.

In a case-control dataset, the data consists of n_1 cases $\{x_i, \boldsymbol{z}_i, y_i, d_i = 1\}$, $i = 1, 2, ..., n_1$, and n_0 controls $\{x_i, \boldsymbol{z}_i, y_i, d_i = 0\}$, $i = n_1+1, n_1+2, ..., n_1+n_0$. We denote $n = n_1+n_0$ as the total sample size. When both the genotype X and the secondary phenotype Y are associated with the primary disease D, the association between X and Y often differs between the

cases and controls. Consequently, directly regressing Y against X using a case-control sample yields biased estimation of β_1. In this thesis, we propose a new estimating equation based approach to estimate the Model (3.1) for secondary traits from case-control samples. The new approach utilizes the entire case-control sample, and yields consistent estimation of β_1 in the general population.

3.2 New estimating equations for the secondary phenotypes in genetic case-control studies

Constructing estimating equations is a common estimation method. Here we define $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \boldsymbol{\beta}_2^*)$ are the true coefficients in the general population. Then the key of is to find an estimating function $S(X, Y, \mathbf{Z}, \boldsymbol{\beta})$ that for any randomly selected subjects from the general population, the following equations hold at the true $\boldsymbol{\beta}^*$,

$$E_Y[S(X, Y, \mathbf{Z}, \boldsymbol{\beta}^*) \mid X, \mathbf{Z}] = 0.$$

In linear least square regressions, the estimating function for the regression coefficient β_1 is

$$S(X, Y, \mathbf{Z}, \beta_1) = \sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i - \mathbf{Z}_i^T \boldsymbol{\beta}_2),$$

which is the first derivative of least square loss function with respect to β_1. In the likelihood based regressions, the estimating function $S(X, Y, \mathbf{Z}, \boldsymbol{\beta})$ can be constructed as the first derivative of log-likelihood function, which is also known as Fisher's score function. Specifically, let $L(\boldsymbol{\beta}; X, Y, \mathbf{Z})$ denote the likelihood function, and then $S(X, Y, \mathbf{Z}, \boldsymbol{\beta}) = \frac{\partial log L(\boldsymbol{\beta}; X, Y, \mathbf{Z})}{\partial \boldsymbol{\beta}}$. For example, in logistic regression, the score function with respect to β_1 is

$$S(X, Y, \mathbf{Z}, \beta_1) = \sum_{i=1}^{n} X_i(Y_i - \frac{\exp(\beta_0 - \beta_1 X_i - \mathbf{Z}_i^T \boldsymbol{\beta}_2)}{1 + \exp(\beta_0 - \beta_1 X_i - \mathbf{Z}_i^T \boldsymbol{\beta}_2)});$$

and in Poisson regression,

$$S(X,Y,\mathbf{Z},\beta_1)=\sum_{i=1}^{n}X_i(Y_i-\exp(\beta_0-\beta_1X_i-\mathbf{Z}_i^T\boldsymbol{\beta}_2)).$$

In regressions with no parametric likelihood functions, $S(X,Y,\mathbf{Z},\boldsymbol{\beta})$ is an estimating function that minimizes the corresponding loss function. For example, in quantile regression, we define the loss function by a piecewise linear function at sample τth quantile as $\rho_\tau(u)=(\tau-I(u<0))u$. The expected loss function of quantile regression is not differentiable at $\tau=0$ and 1. However, one can construct piecewise first derivative as follows

$$S_\tau(X,Y,\mathbf{Z},\beta_{1,\tau})=\sum_{i=1}^{n}X_i[\tau-I\{Y_i\le\beta_{0,\tau}+\beta_{1,\tau}X_i+Z_i^T\boldsymbol{\beta}_{2,\tau}\}].$$

We know the equation $E_Y[S(X,Y,\mathbf{Z},\boldsymbol{\beta}^*)\mid X,\mathbf{Z}]=0$ holds at true $\boldsymbol{\beta}^*$ in the general population. As we do not have a representative sample of the general population, the solving the equation directly using case-control sample is biased. we can, however, expand the equation conditional on the disease status D as follows

$$\begin{aligned}&E_Y[S(X,Y,\mathbf{Z},\boldsymbol{\beta}^*)\mid X,\mathbf{Z}]\\ =&E_Y[S(X,Y,\mathbf{Z},\boldsymbol{\beta}^*)\mid X,\mathbf{Z},D=0]P(D=0\mid X,\mathbf{Z})+E_Y[S(X,Y,\mathbf{Z},\boldsymbol{\beta}^*)\mid X,\mathbf{Z},D=1]P(D=1\mid X,\mathbf{Z})\\ =&0.\end{aligned}\tag{3.2}$$

This expansion provides the basis of constructing the proposed estimating equations.

Let's define $\tilde{y}$ as the counter-factual secondary phenotype under alternative disease status. Specifically, for each subject in the case group, we define $\tilde{y}_i$, $i=1,...,n_1$, as his or her phenotype if he or she is actually a control. And for each subject on the control group, we define $\tilde{y}_i$, $i=n_1+1,...,n$, as his or her phenotype if he or she is actually a case. If we are able to observe both y_i and $\tilde{y}_i$'s, we can then construct the unbiased estimation equations following the expanded estimating equation (3.2). The sample estimation equations can

be written as follows:

$$\mathscr{S}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} [S(x_i, y_i, \boldsymbol{z}_i, \boldsymbol{\beta}) p(d_i \mid x_i, \boldsymbol{z}_i) + S(x_i, \tilde{y}_i, \boldsymbol{z}_i, \boldsymbol{\beta}) p(1 - d_i \mid x_i, \boldsymbol{z}_i)] = 0, \qquad (3.3)$$

where $p(d_i|x_i, \mathbf{z}_i)$ is the probability of being the observed disease status given $(x_i, \mathbf{z}_i)$, and $p(1-d_i|x_i, \mathbf{z}_i)$ is the probability of being counter-factual disease status. One can show that for each summand of Equation (3.3), its conditional expectation given $(x_i, \mathbf{z}_i, d_i)$ is zero at the true $\boldsymbol{\beta}^*$, and thus constitutes an unbiased estimating equation. Following classical theories for M- and Z- estimations (Theorems 5.7 and 5.9 in Van der Vaart [2000]), solving Equation (3.3), $\mathscr{S}_n(\boldsymbol{\beta}) = 0$, leads to the consistent estimation of $\boldsymbol{\beta}$ under certain regulation conditions. The idea of counter-factual outcomes is widely used in causal inference, but in this application we use counter-factual outcomes to estimate the gene-secondary trait association rather than making inferences on causality. Although the estimating equations involve $P(D|X, \mathbf{Z})$, we are not assuming the disease probability only relates to $(X, \mathbf{Z})$. In reality, the disease risk can relate to Y or other auxiliary variables $\mathbf{W}$ as well, and $p(D|X, \mathbf{Z})$ in Equation (3.3) can be viewed as the marginal probability given $(X, \mathbf{Z})$, i.e. $p(D|X, \mathbf{Z}) = \int_{y,\mathbf{w}} p(D|X, \mathbf{Z}, y, \mathbf{w}) dF_{(y,\mathbf{w})}(y, \mathbf{w})$, where $F_{(y,w)}$ is the joint distribution of $(y, \mathbf{w})$.

Solving Equation (3) directly is unfeasible since we are unable to observe the counter-factual secondary outcomes. To get around this difficulty, we propose two approaches. We first propose a model-based simulation approach to simulate the pseudo counter-factual observations, and assemble the estimating equations accordingly. In the second approach, we replace $S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$ by its conditional expectation over $\tilde{y}$. In the next two sections, we elaborate on the two approaches under the assumption that the probability $p(d_i|x_i, \boldsymbol{z}_i)$ is known. An algorithm to estimate $p(d_i|x_i, \boldsymbol{z}_i)$ from the case-control sample is provided in Section 4.5.

3.3 Estimation Approach A: generating pseudo counter-factual observations

Under model (3.1), the linear association between Y and $(X, \boldsymbol{Z})$ holds among both cases and controls. The regression coefficients, however, could vary between them. Hence, we propose to fit Model (3.1) separately for cases and controls, and use the resulting stratified models to simulate pseudo counter-factual outcomes. We define $\boldsymbol{\beta}_d^*$ as the coefficient functions given disease status $D = d$ such that

$$\boldsymbol{\beta}_d^* = \arg\min_{\boldsymbol{\beta}} E_Y[\|S(X, Y, \mathbf{Z}, \boldsymbol{\beta})\| \mid X, \mathbf{Z}, D = d]. \tag{3.4}$$

As $\boldsymbol{\beta}_d^* = (\boldsymbol{\beta}_{d0}^*, \boldsymbol{\beta}_{d1}^*, \boldsymbol{\beta}_{d2}^{*T})$ is a vector of the true coefficients conditional on disease status d, we could define the disease status stratified model $g_1(Y) = g(Y; \boldsymbol{\beta}_1^*) = \beta_{10}^* + X\beta_{11}^* + \mathbf{Z}^T\boldsymbol{\beta}_{12}^*$ as the conditional function for y given $(x, \mathbf{z})$ among cases, and $g_0(Y) = g(Y; \boldsymbol{\beta}_0^*) = \beta_{00}^* + X\beta_{01}^* + \mathbf{Z}^T\boldsymbol{\beta}_{02}^*$ as that among controls. We consider two scenarios to illustrate this idea, one is Generalized Linear Models (GLM) with parametric likelihood functions, and the other is quantile regression with no parametric form.

Simulating counter-factual outcomes in GLM

When the regressions are based on likelihood functions, one can generate the counter-factual outcomes from the stratified estimated model of the alternative disease statue. For example, in logistic regression, we often assume logit link function. Therefore, for each case y_i, we generate its counter-factual outcome from the estimated control model, i.e. $\widehat{\widetilde{y}}_i$ is a random draw from a Bernoulli distribution with success probability $\exp\{g_0(\widehat{y}_i)\}/[1 + \exp\{g_0(\widehat{y}_i)\}]$. Likewise, for each control, we generate its counter-factual pseudo outcome from the estimated case model, where $\widehat{\widetilde{y}}_i$ is a Bernoulli random variable with success probability $\exp\{g_1(\widehat{y}_i)\}/[1 + \exp\{g_1(\widehat{y}_i)\}]$. Plugging the pseudo counter-factual outcomes into the estimating equations (3.3), we could solve for β.

$$\widehat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \left\| \sum_{i=1}^{n} [S(x_i, y_i, \boldsymbol{z}_i, \boldsymbol{\beta}) p(d_i \mid x_i, \boldsymbol{z}_i) + S(x_i, \widehat{y}_i, \boldsymbol{z}_i, \boldsymbol{\beta}) p(1 - d_i \mid x_i, \boldsymbol{z}_i)] \right\| \tag{3.5}$$

The optimization can be viewed as a weighted regression, where one has $2n$ observations, and weights are $p(d_i|x_i, \mathbf{z}_i)$ for the actual outcomes, and $p(1-d_i|x_i, \mathbf{z}_i)$ for the pseudo outcomes. Similar ideas can be applied to other link functions, such as mean regression with identity link and Poisson regression with log link.

Simulating counter-factual outcomes in Quantile Regression

When the regressions do not have full parametric likelihood functions, one need to consider the main model (3.1) nonparametrically or semi-parametrically across the entire distribution of Y. For example, in quantile regresssion, which does not assume any parametric distribution in Y, we need expand the main model (3.1) to the entire quantile process in order to simulate counter-factual phenotypes. This joint modeling approach has been explored in recent work, including Wei et al. [2006], to approximate the conditional quantile function without assuming a parametric likelihood. Specifically, we assume that the linear quantile model holds for an quantile level $\tau \in (0, 1)$. Under this assumption, we define $\boldsymbol{\beta}^*(\tau \mid d)_{d=0,1}$ as the quantile coefficient functions given disease status D=d such that

$$\boldsymbol{\beta}^*(\tau \mid d)_{d=0,1} = \arg\min_{\beta} E_Y \left[\|S_\tau(X, Y, \mathbf{Z}, \boldsymbol{\beta})\| \mid X, \mathbf{Z}, D = d \right], \tag{3.6}$$

for any $\tau \in (0, 1)$. We let $g_{1,\tau}(Y) = \beta_0^*(\tau \mid 1) + X\beta_1^*(\tau \mid 1) + \mathbf{Z}^T\boldsymbol{\beta}_2^*(\tau \mid 1)$ define the conditional quantile function of Y given $(X, \mathbf{Z})$ among cases, and $g_{0,\tau}(Y) = \beta_0^*(\tau \mid 0) + X\beta_1^*(\tau \mid 0) + \mathbf{Z}^T\boldsymbol{\beta}_2^*(\tau \mid 0)$ define that among controls. In what follows, we outline an estimation algorithm to estimate $\boldsymbol{\beta}^*(\tau \mid d)$ from the data, and simulate counter-factual outcomes accordingly. Let $0 < \tau_1 < \tau_2 < \cdots < \tau_k < 1$ be a set of k_n evenly spaced quantile levels.

1. We denote $\widehat{\boldsymbol{\beta}}(\tau_k \mid d)$, $d = 1/0$ as the estimated quantile coefficients for $\boldsymbol{\beta}(\tau_k \mid d)$ in Equation (3.6) within cases and controls, respectively.

2. To approximate the coefficient process $\boldsymbol{\beta}^*(\tau \mid d)$, we define $\widehat{\boldsymbol{\beta}}(\tau \mid d)$ be a piecewise linear functions on [0,1] that concatenates the estimates $\widehat{\boldsymbol{\beta}}(\tau_k \mid d)$ for $0 < \tau_1 < \tau_2 < \cdots < \tau_{k_n} < 1$ and is subject to the constraint of $\widehat{\beta}'(0 \mid d) = \widehat{\beta}'(1 \mid d) = \mathbf{0}$.

3. For the ith subject, $i = 1, \ldots, n$, we simulate its pseudo outcome $\widetilde{y}_i$ by $\widehat{\widetilde{y}}_i = \widehat{\beta}_0(u_i \mid 1-d_i) + \widehat{\beta}_1(u_i \mid 1-d_i)X + \mathbf{Z}^T\widehat{\boldsymbol{\beta}}_2(u_i \mid 1-d_i)$, where u_i is a random draw from Uniform $(0,1)$ distribution.

The simulated $\widehat{\widetilde{y}}_i$'s follows the model-estimated conditional distribution of y_i given $(x_i, \mathbf{z}_i)$ and d_i. Under certain mild conditions as outlined in Wei et al. [2006], $\widehat{\boldsymbol{\beta}}(\tau \mid 1)$ and $\widehat{\boldsymbol{\beta}}(\tau \mid 0)$ uniformly converge to the underlying true ones over the interval $[1/(k_n+1), k_n/(k_n+1)]$ as n_1 and n_2 go to the infinity. Hence, with a reasonably large sample sizes, the simulated $\widehat{\widetilde{y}}_i$ approximates the counter-factual outcome $\widetilde{y}_i$ well.

No matter the regression have likelihood functions or not, we are able to generated simulated $\widehat{\widetilde{y}}_i$. With $\widehat{\widetilde{y}}_i$, we construct the sampling estimating equations as

$$\sum_{i=1}^{n}[S(x_i, y_i, \boldsymbol{z}_i, \boldsymbol{\beta})p(d_i \mid x_i, \boldsymbol{z}_i) + S(x_i, \widehat{\widetilde{y}}_i, \boldsymbol{z}_i, \boldsymbol{\beta})p(1-d_i \mid x_i, \boldsymbol{z}_i)] = 0. \tag{3.7}$$

Simulating pseudo outcomes is subject to sampling uncertainty, and brings extra variability into parameter estimation. To further stabilize the variance, we suggest to repeat the above simulation procedures T time, and use their average as final estimation. Let $\widehat{\beta}_n^{(t)}$ as the estimated coefficients from the t-th replicate, we then use the average of $\widehat{\beta}_n^{(t)}$ as the final estimate of the coefficients. i.e.

$$\widehat{\boldsymbol{\beta}}_n = T^{-1}\sum_{t=1}^{T}\widehat{\boldsymbol{\beta}}_n^{(t)}.$$

Similar to the multiple imputation technique that is commonly used to handle missing data, the variance of $\widehat{\boldsymbol{\beta}}_n$ is fairly stable with a small number of T between 5 and 10. We will demonstrate the effect of different T in the section of simulations. In the rest of paper, we call $\widehat{\boldsymbol{\beta}}_n$ the SICO estimate since it uses SImulated Counter-factual Outcomes.

3.4 Estimation Approach B: estimating $S(X, Y, \mathbf{Z}, \beta)$ by its conditional expectation

An alternative approach to circumvent the difficulty of unobserved $\tilde{y}_i$ is to take the conditional expectation of $S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$ over $\tilde{y}_i$. One can easily show that the following estimating equations lead to unbiased estimators as well:

$$S_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} [S(x_i, y_i, \mathbf{z}_i, \boldsymbol{\beta}) p(d_i \mid x_i, \mathbf{z}_i) + E_{\tilde{y}_i}\{S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta}) | x_i, \mathbf{z}_i\} p(1 - d_i \mid x_i, \mathbf{z}_i)] = 0. \tag{3.8}$$

When $S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$ is linear in $\tilde{y}_i$

In a special case that the estimating function $S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$ is linear in $\tilde{y}_i$, this approach is particularly appealing since one can simply replace $\tilde{y}_i$ by its conditional mean. In this case, the estimation equations (3.8) are equivalent to

$$S_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} [S(x_i, y_i, \boldsymbol{z}_i, \boldsymbol{\beta}) p(d_i \mid x_i, \boldsymbol{z}_i) + S\{x_i, E(\tilde{y}_i | x_i, \mathbf{z}_i), \mathbf{z}_i, \boldsymbol{\beta}\} p(1 - d_i \mid x_i, \boldsymbol{z}_i)] = 0. \tag{3.9}$$

The conditional mean $E(\tilde{y}_i | x_i, z_i)$ can be easily estimated from stratified least square regression. Specifically, one can regress y_i against x_i and $\mathbf{z}_i$ separately among cases and controls, and estimate $E(\tilde{y}_i | x_i, \mathbf{z}_i)$ by the predicted value under alternative disease status model. This way, the estimate can be obtained using one-step optimization, by solving the following equations

$$\widehat{S}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[S(x_i, y_i, \mathbf{z}_i, \boldsymbol{\beta}) p(d_i | x_i, \mathbf{z}_i) + S\{x_i, \widehat{E}(\tilde{y}_i | x_i, \mathbf{z}_i), \mathbf{z}_i, \boldsymbol{\beta}\} p(1 - d_i | x_i, \mathbf{z}_i) \right] = 0, \tag{3.10}$$

where $\widehat{E}(\tilde{y}_i | x_i, \mathbf{z}_i)$ is the predicted outcome given x_i and $\mathbf{z}_i$ under the alternative disease status. We can define $\widetilde{\boldsymbol{\beta}}_n$ as resulting estimate from conditional expectation of $S(X, \widetilde{Y}, \mathbf{Z}, \boldsymbol{\beta})$. Then

$$\widetilde{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \left\| \sum_{i=1}^{n} \left[S(x_i, y_i, \mathbf{z}_i, \boldsymbol{\beta}) p(d_i|x_i, \mathbf{z}_i) + S\{x_i, \widehat{E}(\tilde{y}_i|x_i, \mathbf{z}_i), \mathbf{z}_i, \boldsymbol{\beta}\} p(1-d_i|x_i, \mathbf{z}_i) \right] \right\|. \tag{3.11}$$

Similar as in the first approach, this optimization is equivalent to weighted linear regression where the weights are $p(d_i|x_i, z_i)$ for the actual outcomes, and are $p(1-d_i|x_i, z_i)$ for the $\widehat{E}(\tilde{y}_i|x_i, \mathbf{z}_i)$.

When $S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$ is not linear in $\tilde{y}_i$

When the estimating function $S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$ is not linear in $\tilde{y}_i$, we are unable to pass the expectation into the estimating function. In a simple scenario where we have sufficient number of cases and controls given each value of $(x_i, \mathbf{z}_i)$, we could estimate the expectation terms by

$$\begin{aligned}
\widehat{E}_{\tilde{y}_i}[S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})|x_i, \mathbf{z}_i] &= \frac{\sum_{j=n_1+1}^{n_1+n_2} I(x_j = x_i) I(\mathbf{z}_j = \mathbf{z}_i) S(x_j, y_j, \mathbf{z}_j, \boldsymbol{\beta})}{\sum_{j=n_1+1}^{n_1+n_2} I(x_j = x_i) I(\mathbf{z}_j = \mathbf{z}_i)}, i = 1, ..., n_1; \\
\widehat{E}_{\tilde{y}_i}[S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})|x_i, \mathbf{z}_i] &= \frac{\sum_{j=1}^{n_1} I(x_j = x_i) I(\mathbf{z}_j = \mathbf{z}_i) S(x_j, y_j, \mathbf{z}_j, \boldsymbol{\beta})}{\sum_{j=1}^{n_1} I(x_j = x_i) I(\mathbf{z}_j = \mathbf{z}_i)}, i = n_1 + 1, ..., n
\end{aligned} \tag{3.12}$$

where $I(\cdot)$ is an indicator function. These are essentially the sample means of the estimating function with the same $(x_i, \mathbf{z}_i)$ but alternative diseases status. Following the law of large numbers, both estimates converge to the true expectations with $\sqrt{n}$ rate. Such applications can be found in single loci analysis in genetic studies [Kraft, 2007]. In more general scenarios, especially when $\mathbf{Z}$ includes continuous variables, the indicator function no longer produces valid estimates, since we may have very few observations at a given value of $\mathbf{Z}$. We propose to replace it by some suitable kernel function $K_h(\cdot)$ with bandwidth h, and approximate the expectation by

$$\begin{aligned}\widehat{E}_{\tilde{y}_i}[S(x_i,\tilde{y}_i,\mathbf{z}_i,\boldsymbol{\beta})|x_i,\mathbf{z}_i] &= \frac{\sum_{j=n_1+1}^{n_1+n_2} I(x_j=x_i)K_h(\|\mathbf{z}_j=\mathbf{z}_i\|)S(x_j,y_j,\mathbf{z}_j,\boldsymbol{\beta})}{\sum_{j=n_1+1}^{n_1+n_2} I(x_j=x_i)K_h(\|\mathbf{z}_j=\mathbf{z}_i\|)}, i=1,...,n_1;\\ \widehat{E}_{\tilde{y}_i}[S(x_i,\tilde{y}_i,\mathbf{z}_i,\boldsymbol{\beta})|x_i,\mathbf{z}_i] &= \frac{\sum_{j=1}^{n_1} I(x_j=x_i)K_h(\|\mathbf{z}_j=\mathbf{z}_i\|)S(x_j,y_j,\mathbf{z}_j,\boldsymbol{\beta})}{\sum_{j=1}^{n_1} I(x_j=x_i)K_h(\|\mathbf{z}_j=\mathbf{z}_i\|)}, i=n_1+1,...,n\end{aligned} \tag{3.13}$$

With the estimated $\widehat{E}_{\tilde{y}_i}[S(x_i,\tilde{y}_i,\mathbf{z}_i,\boldsymbol{\beta}) \mid x_i,\mathbf{z}_i]$, we can assemble the working estimating equations

$$\widehat{S}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n}[S(x_i,y_i,\boldsymbol{z}_i,\boldsymbol{\beta})p(d_i \mid x_i,\boldsymbol{z}_i) + \widehat{E}_{\tilde{y}_i}\{S(x_i,\tilde{y}_i,\boldsymbol{z}_i,\boldsymbol{\beta}) \mid x_i,\mathbf{z}_i\}p(1-d_i \mid x_i,\boldsymbol{z}_i)] = 0. \tag{3.14}$$

The $\widetilde{\boldsymbol{\beta}}_n$ is the solution to this equation. It is equivalently as

$$\widetilde{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \left\| \sum_{i=1}^{n}[S(x_i,y_i,\boldsymbol{z}_i,\boldsymbol{\beta})p(d_i \mid x_i,\boldsymbol{z}_i) + \widehat{E}_{\tilde{y}_i}\{S(x_i,\tilde{y}_i,\boldsymbol{z}_i,\boldsymbol{\beta}) \mid x_i,\mathbf{z}_i\}p(1-d_i \mid x_i,\boldsymbol{z}_i)] \right\| \tag{3.15}$$

Note that the estimates in (3.12) - (3.13) are linear functions of the original regression estimating functions. Hence one could reorganize the estimating functions (3.14) as

$$\widehat{S}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n_1} w_i S(x_i,y_i,\mathbf{z}_i,\boldsymbol{\beta}) + \sum_{j=n_1+1}^{n} w_j S(x_j,y_j,\mathbf{z}_j,\boldsymbol{\beta})$$

where

$$w_i = p(d_i=1|x_i,\mathbf{z}_i) + \sum_{j=n_1+1}^{n_1+n_2} \frac{I(x_j-x_i)K_h(\|\mathbf{z}_j-\mathbf{z}_i\|)p(d_j=1|x_j,\mathbf{z}_j)}{\sum_{i=1}^{n_1} I(x_j-x_i)K_h(z_j=z_i)}$$

and

$$w_j = p(d_j=0|x_j,\mathbf{z}_j) + \sum_{i=1}^{n_1} \frac{I(x_i-x_j)K_h(\|\mathbf{z}_i-\mathbf{z}_j\|)p(d_i=0|x_i,\mathbf{z}_i)}{\sum_{j=n_1+1}^{n} K_h(\|x_i-x_j\|)}.$$

Since the weights w_i are not functions of $\boldsymbol{\beta}$, solving the working estimating equations is equivalent to a weighted regression and is computationally straightforward.

Finally, to choose an optimal bandwidth or a kernel function in (3.13), we propose to use K-fold cross-validation. Specifically, we randomly partition the data into K subsets and

denote $\widetilde{\beta}^{(-\ell)}(h)$ as the estimated coefficients using bandwidth h without the the ℓth subset of data, $\ell = 1, \ldots, K$. The optimal bandwidth is defined as

$$h^{opt} = \arg\min_{h} \sum_{\ell=1}^{K} \left[\sum_{i \in C_\ell} w_i \mathscr{L}\{x_i, y_i, \mathbf{z}_i, \widetilde{\beta}^{(-\ell)}(h)\} + \sum_{j \in \Gamma_\ell} w_j \mathscr{L}\{x_j, y_j, \mathbf{z}_j, \widetilde{\beta}^{(-\ell)}(h)\} \right],$$

where C_ℓ is the index set for the ℓ-th case subset, Γ_ℓ is the index set for the ℓ-th control subset, and $\mathscr{L}(x, y, \mathbf{z}, \boldsymbol{\beta})$ is the loss function. For example, in least square mean regression, $\mathscr{L}(x, y, \mathbf{z}, \boldsymbol{\beta}) = (y - \beta_0 - x\beta_1 - z^T \boldsymbol{\beta}_2)^2$; in quantile regression, $\mathscr{L}_\tau(x, y, \mathbf{z}, \boldsymbol{\beta}) = (y - \beta_0 - x\beta_1 - z^T \boldsymbol{\beta}_2)\{\tau - I(y - \beta_0 - x\beta_1 - z^T \boldsymbol{\beta}_2 < 0)\}$. Essentially, we choose the bandwidth that minimizes the weighted cross-validated regression loss functions.

We call $\widetilde{\boldsymbol{\beta}}_n$ in Approach B as the CE estimates, since the Conditional Expectation is used to estimate the estimating function. Both SICO and CE estimates are consistent and asymptotic normal. One can refer Wei et al. [2015] for the large sample properties of the proposed estimators. When the dimension of $(x, \mathbf{z})$ increases or when covariate space is sparse, the kernel smoothing in the approach B could be difficult due to the curse of dimensionality. Approach A avoids the smoothness, and hence is readily applicable for any dimension of $(x, \mathbf{z})$. However, it makes a stronger assumption of the linear model. For quantile regression, the linear model needs to be hold for the entire quantile process. This assumption could be relaxed by using more general models such as semiparametric partly linear models.

3.5 Estimation of $p(d_i|x_i, \mathbf{z}_i)$

In the aforementioned two estimation algorithms, we assumed that the conditional disease probability $p(d_i|x_i, \mathbf{z}_i)$ is known. In practice, it needs to be estimated. To estimate $p(d_i|x_i, \mathbf{z}_i)$, we could use the model in primary analysis or assume a logistic model as follows:

$$P(D = 1|X, \mathbf{Z}) = \exp(\gamma_0 + X\gamma_1 + \mathbf{Z}^T \boldsymbol{\gamma}_2)/\{1 + \exp(\gamma_0 + X\gamma_1 + \mathbf{Z}^T \boldsymbol{\gamma}_2)\}. \tag{3.16}$$

Note that model (3.16) is a working model to approximate the distribution of disease given $(X, \mathbf{Z})$ and may differ from the true disease model because the secondary outcome Y may also affect disease risk. In our simulation study in later section, we consider three disease models that are based on Y.

Further note that the intercept γ_0 cannot be consistently estimated directly form the case-control data, and needs to be calibrated to yield valid estimation of $p(d_i|x_i, \mathbf{z}_i)$ [Prentice and Pyke, 1979]. Assuming that the overall disease prevalence in the general population, denoted by P_0, is known, we can estimate γ_0 by solving the following equation

$$P_0 = \int_{X,\mathbf{Z}} \exp(\gamma_0 + X\widehat{\gamma}_1 + \mathbf{Z}^T\widehat{\boldsymbol{\gamma}}_2)/\{1 + \exp(\gamma_0 + X\widehat{\gamma}_1 + \mathbf{Z}^T\widehat{\boldsymbol{\gamma}}_2)\} dF_{X\mathbf{Z}}, \tag{3.17}$$

where $F_{X\mathbf{Z}}$ is the joint distribution of X and $\mathbf{Z}$, and $\widehat{\gamma}_1$ and $\widehat{\boldsymbol{\gamma}}_2$ are the estimated γ_1 and $\boldsymbol{\gamma}_2$ from logistic regression. When the joint distribution $F_{X\mathbf{Z}}$ is difficult to obtain, we propose to approximate γ_0 by solving its sample version

$$\widehat{\gamma}_0 = \arg\min_{\gamma_0} \left(P_0 - n^{-1} \sum_{i=1}^{n} \exp(\gamma_0 + x_i\widehat{\gamma}_1 + \mathbf{z}_i^T\widehat{\boldsymbol{\gamma}}_2)/\{1 + \exp(\gamma_0 + x_i\widehat{\gamma}_1 + \mathbf{z}_i^T\widehat{\boldsymbol{\gamma}}_2)\} \right)^2 \tag{3.18}$$

Both Equation (3.17) and (3.18) are univariate optimization. Therefore, obtaining $\widehat{\gamma}_0$ from either equation is computationally easy. The estimate of the conditional disease probability $p(d_i|x_i, \mathbf{z}_i)$ can be written as $\widehat{p}(d_i|x_i, \mathbf{z}_i) = \exp(\widehat{\gamma}_0 + \widehat{\gamma}_1 X + \mathbf{Z}^T\widehat{\gamma}_2)/\{1 + \exp(\widehat{\gamma}_0 + \widehat{\gamma}_1 X + \mathbf{Z}^T\widehat{\gamma}_2)\}$. When the working model or disease prevalence P_0 is mis-specified, the resulting $\widehat{\gamma}_0$ could be slightly biased. The simulation studies in Section 4.5 show the results on the estimation of the coefficients when the prevalence the working model or P_0 is mis-specified.

3.6 Bootstrap procedure for the confidence intervals and hypothesis tests

In Sections 3.3 and 3.4, we outlined two estimation algorithms to estimate the parameters in Model (3.1). Although both SICO and CE estimates can be viewed as some form of

weighted regressions, the direct output of Wald test statistics does not apply, because it does not take into consideration of the uncertainty from the estimated $p(d|x, \mathbf{z})$, simulated $\tilde{y}_i$ and the kernel smoothness. In addition, it is the difficult to estimate asymptotic variances by any analytically tractable form [Wei et al., 2015]. Therefore, we propose to use bootstrap method to obtain the bootstrap standard error of our proposed estimates. With the bootstrap standard error, we are able to construct bootstrap confidence intervals and apply Wald test statistics to test the null hypothesis, say $H_0 : \beta_1 = 0$, i.e. whether the genetic variant(s) are associated with the secondary phenotype Y in the general population. The Bootstrap test statistics is written as follows:

$$\frac{(\widehat{\beta}_1 - \beta_1^*)^2}{var(\widehat{\beta}_1)} \sim \chi_1^2 \tag{3.19}$$

We elaborate the bootstrap procedure as follows:

1. Bootstrap cases and controls separately to assemble a bootstrap case-control sample. In details, we randomly select n_1 cases from case sample and n_0 controls from control sample with replacement.

2. For each bootstrap sample, we re-apply the proposed algorithm to obtain bootstrap estimates. For SICO estimator, that includes re-generating pseudo-outcomes $\tilde{y}_i$ and re-estimating $p(d|x_i, \mathbf{z}_i)$. For CE estimator, that includes re-estimating $E_{\tilde{y}_i}[S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})]$ and $p(d|x_i, \mathbf{z}_i)$.

3. We repeat steps 1 and 2 B times, then calculate the bootstrap standard error. We use the bootstrap standard error to construct confidence intervals and bootstrap Chi-square test statistics for inference.

We evaluated the type I error of this bootstrap procedure using simulations in Section 4.4. The bootstrap based inferences are applied to the real data examples in Sections 5.2 and 5.3.

Chapter 4

Simulation studies

4.1 Finite sample performance with GLM

In this section, we present several numerical studies to investigate the finite sample performance of the proposed estimation method under the GLM framework with comparison to comparable existing methods. We simulate the data mimicking the loci-to-loci comparison in GWA case-control studies.

Model settings: Same as before, we denote by $D = \{0, 1\}$ the primary disease status, by $X = \{0, 1, 2\}$ a single SNP (under additive model) with minor allele frequency (MAF) 0.3, and by Z a covariate of interest following a standard normal distribution. The correlation coefficient between X and Z is set to be 0.3. We consider both binary and continuous secondary phenotypes Y. For binary Y, we assume a linear logistic model. We consider both binary secondary and continuous phenotypes Y . For binary Y , we consider a the following logistic model:

$$P(Y = 1|X, Z) = \frac{\exp(-1 + 0.2X + 0.1Z)}{1 + \exp(-1 + 0.2X + 0.1Z)} \tag{4.1}$$

For continuous Y with homoscedastic error, we consider a linear model as follows:

$$Y = 1 + 0.2X + 0.1Z + e \tag{4.2}$$

In Equation (4.1), the prevalence of Y is approximately 30%. In Equation (4.2), the error term follows independent and identically distributed (i.i.d.) normal distribution that $e \sim N(0,1)$. Since the genetic effects are often small in GWAS, we choose $\beta_1^* = 0.2$ as the true coefficient of X in predicting Y.

To model the disease probability $P(D|X,Y,Z)$, we consider three possible settings. Setting 1 (Logistic Setting) assumes the probability of disease follows a linear logistic model with main effects of X and Y. Similar settings were considered in Lin and Zeng [2009] for SPML and Wang and Shete [2011] for bias correction approach. Setting 2 (Interaction Setting) extends the Logistic Setting by including the interactive term between X and Y. Similar settings were considered in Li and Gail [2012]'s paper for adaptive weighted approach and Wang and Shete [2012]'s paper for modified bias correction approach. Finally, Setting 3 (Piecewise Setting) assumes that $P(D|X,Y,Z)$ follows piecewise linear model instead of logistic regression. The detailed mathematical forms of the disease models are given below. The U_1 and U_2 are the 0.25th and 0.75th quantiles of the $0.3X + \log(2)Y + \log(2)Z$.

- Setting 1 (Logistic Setting):

$$P(D=1|X,Y,Z) = \frac{\exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z)}{1+\exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z)}$$

- Setting 2 (Interaction Setting):

$$P(D=1|X,Y,Z) = \frac{\exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z + 0.2XY)}{1+\exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z + 0.2XY)}$$

- Setting 3 (Piecewise Setting):

$$P(D=1|X,Y,Z) = \begin{cases} 0.05, & 0.3X + \log(2)Y + \log(2)Z \le U_1 \\ 0.05 + 0.1\frac{0.3X+\log(2)Y+\log(2)Z-U_1}{U_2-U_1}, & U_1 < 0.3X + \log(2)Y + \log(2)Z \le U_2 \\ 0.15, & 0.3X + \log(2)Y + \log(2)Z > U_2 \end{cases}$$

The prevalence of the primary disease (P_0) is set to be 10%. The intercepts γ_0 in these models are selected to match the overall disease prevalence in the population. For each of

the model settings above, we first generate a large number of observations ($N = 500,000$), which we treat as a general population. From the initial sample, we first randomly draw 500 cases and 500 controls to mimic a small case-control study, and then increase 2000 cases and 2000 controls for a large case-control study. The selection is under simple sampling scheme, which means the selection probability only depends on the disease status. For the logstic regression, we evaluate the finite sample performance of the proposed methods with SICO and CE estimates. For linear regression, because CE estimate is very simple with only one-step optimization, we only evaluate the finite sample performance of CE estimates. We compared the resulting estimates with traditional methods, IPW and SPML approaches. We select IPW and SPML approaches because considering the current available secondary analysis methods, the IPW method is the most simple and robust method, and SPML is the most efficient method when its model assumption is satisfied. For SICO estimate, we vary the T values from 1 to 100 (T is the number of pseudo samples generated) and compare their estimates. In the CE estimation for logistic regressions, we use a kernel function $K_h((x_1,z_1)^T,(x_2,z_2)^T) = I(x_1 = x_2)\exp\{-(z_1 - z_2)^2/h\}$, where h is the bandwidth selected by the 5 fold cross-validation.

Comparison Methods: We estimated the coefficients under the different settings above using the following methods:

(1) regression using cases only,

(2) regression using controls only,

(3) regression using combined case-control sample without adjustment,

(4) regression using case-control sample adjusting for primary disease status ,

(5) IPW,

(6) SPML,

(7) our proposed SICO estimator (in logistic models),

(8) our proposed CE estimator (in both logistic and linear models).

Results and discussions: Tables 4.1 to 4.2 summarize the relative bias, standard error and mean squared error of the estimated coefficients β_1 based on 500 Monte-Carlo replicates from logistic model and linear model, respectively. According to the tables, the traditional methods, including direct regressions applied to case only, control only, the combined case-control sample , and combined case-control sample adjusting for primary disease status, are all biased in all settings we consider. Hence, without appropriate adjustment, traditional methods are easy to provide biased estimation for the $X - Y$ association in genetic case-control data.

Both SICO and CE estimates produce fairly accurate estimates in all the models. In SICO, the estimated coefficients are unbiased even with $T = 1$. The standard errors do decrease slightly as T increases, but they quickly stabilize after $T = 10$. Therefore, we conclude that a relatively small number of imputations is enough to reach the optimal efficiency of this approach and it is computationally efficient. The CE estimates are obtained using one-step optimization in linear model, and kernel smoothing techniques in logistic model. We could observe the SICO estimates perform better than CE estimates in logistic model when CE estimates requires smoothness. Overall, it suggests that the proposed estimating equation approach works well in performing unbiased secondary analysis in case-control studies.

The IPW performs well in correcting the bias in all settings we consider. The calculation of IPW method requires the information on sampling scheme. Under the simple sampling scheme, where the selection of cases and controls solely depends on the disease status, it is equivalent to use the disease prevalence as in the proposed methods. Therefore, it's performance is comparable to the proposed estimates. We will consider additional comparison under complex sampling scheme later in Section 4.3.

The SPML approach provides efficient unbiased estimations when the linear logistic model assumption is satisfied but introduces biases when violated. In details, Under the Logistic Setting, the SPML estimate is of most efficiency of all methods we consider. Under Interaction and Piecewise Settings, when the linear logistic model assumption is violated, the SPML estimates contain considerable bias.

In our algorithm, we assume the working model as $P(D|X,Z) = \exp(\gamma_0+\gamma_1 X+\gamma_2 Z)/(1+\exp(\gamma_0+\gamma_1 X+\gamma_2 Z))$. It is different from generated data that is based on three $P(D|X,Y,Z)$ settings. Although under the mis-specified $P(D|X,Z)$, the proposed estimating equation based approach performances fairly well, which shows the proposed approach is quite robust to the $P(D|X,Z)$ model mis-specification. In Section 4.5, we will consider additional scenarios to test the robustness boundary of $\widehat{P}(D|X,Z)$.

		Logistic			Interaction			Piecewise		
n	Method	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE$\times n$
2000	Case	-13.3	0.067	10.2	66.3	0.064	42.4	-36.5	0.072	21.4
	Control	-10.2	0.085	15.3	-30.1	0.081	20.9	3.5	0.081	13.3
	CC	10.4	0.050	5.9	61.4	0.050	34.4	-12.0	0.056	7.5
	Adj CC	-12.1	0.051	6.2	26.9	0.051	10.6	-18.1	0.056	9.1
	IPW	0.3	0.072	10.5	2.0	0.069	9.5	1.1	0.073	10.5
	SPML	0.5	0.050	5.0	46.5	0.050	21.7	-15.5	0.056	8.3
	SICO (T=1)	-0.4	0.082	13.6	-0.8	0.076	11.7	0.9	0.080	12.9
	SICO (T=10)	-0.5	0.074	10.8	-0.5	0.071	10.0	0.9	0.073	10.8
	SICO (T=100)	-0.5	0.073	10.6	-0.4	0.070	9.8	0.6	0.073	10.6
	CE	-6.0	0.076	11.8	-7.8	0.072	10.9	-0.4	0.073	10.6
500	Case	-19.6	0.145	11.2	54.5	0.138	15.5	-44.7	0.137	13.3
	Control	-14.6	0.156	12.5	-27.0	0.168	15.5	6.7	0.152	11.5
	CC	4.0	0.107	5.8	55.4	0.103	11.4	-15.7	0.101	5.6
	Adj CC	-17.9	0.107	6.4	21.3	0.104	6.3	-20.9	0.102	6.0
	IPW	-4.2	0.134	9.0	2.9	0.141	9.9	2.7	0.134	8.9
	SPML	-5.3	0.107	5.8	40.7	0.102	8.5	-18.6	0.101	5.8
	SICO (T=1)	-5.6	0.152	11.5	-0.8	0.163	13.3	1.8	0.152	11.5
	SICO (T=10)	-4.2	0.139	9.7	2.3	0.143	10.2	2.0	0.134	9.0
	SICO (T=100)	-4.9	0.137	9.4	0.8	0.144	10.3	2.1	0.135	9.1
	CE	-10.6	0.138	9.8	-9.7	0.148	11.1	1.5	0.136	9.2

Table 4.1: The relative bias (RB), standard error (SE) and mean squared error (MSE) of the estimated coefficient β_1 in *logistic model*. The true value $\beta_1^* = 0.2$. "Case" stands for unadjusted logistic regression using case sample only. "Control" stands for unadjusted logistic regression using control sample only. "CC" stands for unadjusted logistic regression using both case and control samples. "Adj CC" stands for logistic regression using both case and control samples adjusting for primary disease status. "IPW" stands for inverse probability weighted logistic regression. "SPML" stands for semi-parametric maxmium likelihood based logistic regression. "SICO (T)" stands for proposed SICO estimates with T replicate. "CE" stands for proposed CE estimates using kernel smoothing techniques.

		Logistic			Interaction			Piecewise		
n	Method	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE$\times n$
2000	Case	-24.8	0.031	6.8	13.3	0.032	3.6	-27.6	0.032	7.9
	Control	-10.6	0.037	3.5	-34.7	0.035	11.8	0.5	0.036	2.6
	CC	12.8	0.024	2.5	52.7	0.025	23.8	-8.6	0.024	1.7
	Adj CC	-18.6	0.023	3.7	-8.4	0.024	1.6	-14.2	0.024	2.6
	IPW	1.6	0.032	2.1	-0.8	0.031	1.9	-0.5	0.032	2.0
	SPML	0.2	0.026	1.3	28.3	0.024	7.7	-12.1	0.024	2.2
	CE	0.6	0.033	2.1	0.6	0.033	2.1	-0.7	0.032	2.1
500	Case	-23.3	0.065	3.2	0.4	0.062	1.9	-23.4	0.066	3.3
	Control	-11.4	0.068	2.6	-35.7	0.075	5.3	0.3	0.071	2.5
	CC	11.5	0.046	1.3	45.8	0.051	5.5	-5.9	0.050	1.3
	Adj CC	-18.2	0.046	1.7	-16.1	0.049	1.7	-11.9	0.049	1.5
	IPW	0.4	0.060	1.8	-3.1	0.066	2.2	0.0	0.064	2.1
	SPML	-1.2	0.046	1.0	21.1	0.049	2.1	-9.7	0.050	1.4
	CE	-0.8	0.060	1.8	-8.0	0.067	2.3	-0.4	0.064	2.1

Table 4.2: The relative bias (RB), standard error (SE) and mean squared error (MSE) of the estimated coefficient β_1 in *linear model*. The true $\beta_1^* = 0.2$. "Case" stands for unadjusted linear regression using case sample only. "Control" stands for unadjusted linear regression using control sample only. "CC" stands for unadjusted linear regression using both case and control samples. "Adj CC" stands for linear regression using both case and control samples adjusting for primary disease status. "IPW" stands for inverse probability weighted linear regression. "SPML" stands for semi-parametric maxmium likelihood based linear regression. "CE" stands for proposed CE estimates using one-step optimization.

4.2 Finite sample performance with quantile regression

In this section, we present the numerical studies for the finite sample performance of the proposed estimates under the quantile regression framework. Since quantile regression does not have a parametric likelihood assumption of the data, the estimating approach is considerable different from the ones for GLM. The notation is the same as before that we let $D = \{0, 1\}$ denote the primary disease status, $X = \{0, 1, 2\}$ denote a single SNP (under additive model) with minor allele frequency (MAF) 0.3, and $Z \sim N(0, 1)$ denote a covariate of interest. For the heteroscedastic continuous Y, we consider the following location scale model:

$$Y = 1 + 0.12X + 0.1Z + (1 + 0.02X)e \tag{4.3}$$

For the error term e, we consider both normal and skewed distributions. In details, $e \sim N(0, 1)$ in Quantile Model (1) and $e \sim \chi^2_1/\sqrt{2}$ in Quantile Model (2). We scale e_i in Quantile Model (2) so that it has the same error variance as in Quantile Model (1) to standardize the signal-to-noise ratio. According to the Equation (4.3), the covariate effect of X is stronger on the upper quantiles than the lower ones, while the covariate Z has constant effect at all the quantile levels. Specifically, the true X coefficient is $0.12 + 0.02Q_{e_i}(\tau)$ at the τth quantile, and the true Z coefficient is 0.1 at all quantiles. The disease model for $P(D|X, Y, Z)$ considered here is similar to the Logistic Setting that the disease prevalence follows linear logistic model (shown below). The only difference is that we adjust γ_0 to let the disease prevalence $P_0 = 5\%$.

$$P(D = 1|X, Y, Z) = \frac{\exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z)}{1 + \exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z)}$$

For both Quantile Model (1) and (2), we first simulate 500 cases and 500 controls to mimic a small case-control study, and then increase 2000 cases and 2000 controls for a large case-control study. Since the performance at different quantile levels may vary, we estimate the $X - Y$ associations at five different quantiles $\tau = (0.1, 0.25, 0.5, 0.75, 0.9)$ simultaneously. The estimates for $\tau = 0.5$ and 0.9 are shown in the tables to demonstrate the

performance at different quantiles. The selection of the best bandwidth for kernel smoothing in CE estimates is particular tricky in quantile regressions, as a bandwidth might be optimal for one quantile but not another. Therefore, in addition to 5-fold cross-validation we used in Section 4.1, we further investigate the effects of bandwidths on estimation by using the fixed bandwidths. In details, we repeatedly apply the proposed estimation procedure to a sequence of fixed bandwidths, ranging from 0.02 to 100, and then evaluate the resulting mean absolute bias with each bandwidth. To see whether the estimates from smaller sample sizes are more sensitive to bandwidth selection, we repeat this procedure on a subset of 500 cases and 500 controls.

Table 4.3 and 4.4 summarize the relative bias, standard error and mean squared error of the estimated quantile coefficients at quantile levels 0.5 and 0.9. Similar as in Section 4.4, the estimated quantile coefficients from the unadjusted traditional methods are seriously biased. Both the SICO and CE estimators produce fairly accurate estimates in all the models and at all the quantile levels with all the relative biases being controlled within 5%. Since we sample cases and controls solely depends on the disease status, the IPW also performs well in controlling the bias as expected. The mean squared errors of the SICO estimates ($T \geq 10$) are slightly smaller than IPW ones in all four scenarios. Overall, it suggests that the proposed estimating equation approach works well in performing unbiased secondary quantile analysis in case-control studies.

		$\tau = 0.5$			$\tau = 0.9$		
n		RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$
500	Case	12.2	0.087	3.9	4.7	0.116	6.7
	Control	-7.3	0.089	4.0	-6.1	0.122	7.4
	CC	41.0	0.064	3.2	31.0	0.086	4.8
	IPW	0.1	0.084	3.5	2.0	0.112	6.3
	SICO (T=1)	-0.6	0.088	3.9	-0.6	0.116	6.7
	SICO (T=10)	-1.3	0.082	3.4	-2.9	0.105	5.5
	SICO (T=100)	4.1	0.081	3.3	5.4	0.104	5.4
	CE	0.3	0.086	3.7	-0.3	0.114	6.5
2000	Case	10.8	0.041	3.8	-0.6	0.055	6.0
	Control	-13.0	0.044	4.3	-13.2	0.060	8.0
	CC	39.4	0.032	6.5	25.2	0.043	6.4
	IPW	-4.8	0.042	3.6	-2.5	0.056	6.3
	SICO (T=1)	-4.6	0.047	4.4	-1.9	0.064	8.1
	SICO (T=10)	-3.5	0.042	3.5	-1.6	0.055	6.0
	SICO (T=100)	-3.9	0.042	3.5	-2.0	0.055	6.0
	CE	-4.7	0.042	3.6	-2.0	0.056	6.3

Table 4.3: The relative bias (RB), standard error (SE) and mean squared error (MSE) of the estimated quantile coefficient in *Quantile Model (1)*. The true $\beta_{1,\tau}$ is $0.12+0.02Q_{e_i}(\tau)$. In Quantile Model (1), $e_i \sim N(0,1)$. "Case" stands for unadjusted quantile regression using case sample only. "Control" stands for unadjusted quantile regression using control sample only. "CC" stands for unadjusted quantile regression using both case and control samples. "IPW" stands for inverse probability weighted logistic regression. "SICO (T)" stands for proposed SICO estimates with T replicate. "CE" stands for proposed CE estimates using kernel smoothing techniques.

		$\tau = 0.5$			$\tau = 0.9$		
n		RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$
500	Case	-34.7	0.184	17.9	-102.2	0.372	82.1
	Control	-2.5	0.048	1.2	-19.0	0.197	19.8
	CC	36.6	0.072	3.7	57.9	0.291	46.6
	IPW	1.3	0.049	1.2	4.3	0.201	20.2
	SICO (T=1)	2.2	0.054	1.5	8.0	0.224	25.1
	SICO (T=10)	1.5	0.048	1.2	3.5	0.198	19.5
	SICO (T=100)	1.6	0.048	1.1	2.8	0.194	18.9
	CE	2.6	0.051	1.3	5.0	0.204	20.7
2000	Case	-28.8	0.086	17.4	-111.9	0.184	130.0
	Control	-1.7	0.025	1.3	-22.3	0.102	23.2
	CC	39.1	0.035	7.3	50.6	0.141	52.6
	IPW	2.0	0.025	1.3	0.2	0.099	19.6
	SICO (T=1)	2.7	0.028	1.6	0.0	0.108	23.1
	SICO (T=10)	2.0	0.026	1.3	0.2	0.098	19.3
	SICO (T=100)	2.1	0.025	1.3	0.5	0.096	18.5
	CE	2.2	0.026	1.4	-0.9	0.100	19.8

Table 4.4: The relative bias (RB), standard error (SE) and mean squared error (MSE) of the estimated quantile coefficient in *Quantile Model (2)*. The true $\beta_{1,\tau}$ is $0.12+0.02Q_{e_i}(\tau)$. In Quantile Model (2), $e_i \sim \chi^2_1/\sqrt{2}$. "Case" stands for unadjusted quantile regression using case sample only. "Control" stands for unadjusted quantile regression using control sample only. "CC" stands for unadjusted quantile regression using both case and control samples. "IPW" stands for inverse probability weighted quantile regression. "SICO (T)" stands for proposed SICO estimates with T replicate. "CE" stands for proposed CE estimates using kernel smoothing techniques.

In Figure 4.1, we plot the mean absolute biases of the estimated quantile coefficients from Quantile Model (1) against the logarithm of their corresponding bandwidths. The horizontal line is the mean absolute bias of the estimated coefficients with CV selected bandwidth. Similarly, we plot in Figure 4.2 the mean absolute biases with fixed and CV selected bandwidth from Quantile Model (2). We found that the biases are well controlled within 0.02 regardless of the selection of bandwidth. Hence we conclude that the proposed method is not sensitive to the choice of bandwidth. The estimates are close for a fairly wide range of bandwidth. The estimates using CV selected optimal bandwidth outperform most of those with fixed bandwidths, which suggested that the proposed bandwidth selection works reasonably well. The advantage of CV selected bandwidth is more visible at the 0.5th quantile in Quantile Model (1) when the outcome is normally distributed, and at 0.1th quantile in Quantile Model (2) when the outcome follows χ^2 distribution. In other words, the selection is more helpful for the quantile levels at which the density is higher.

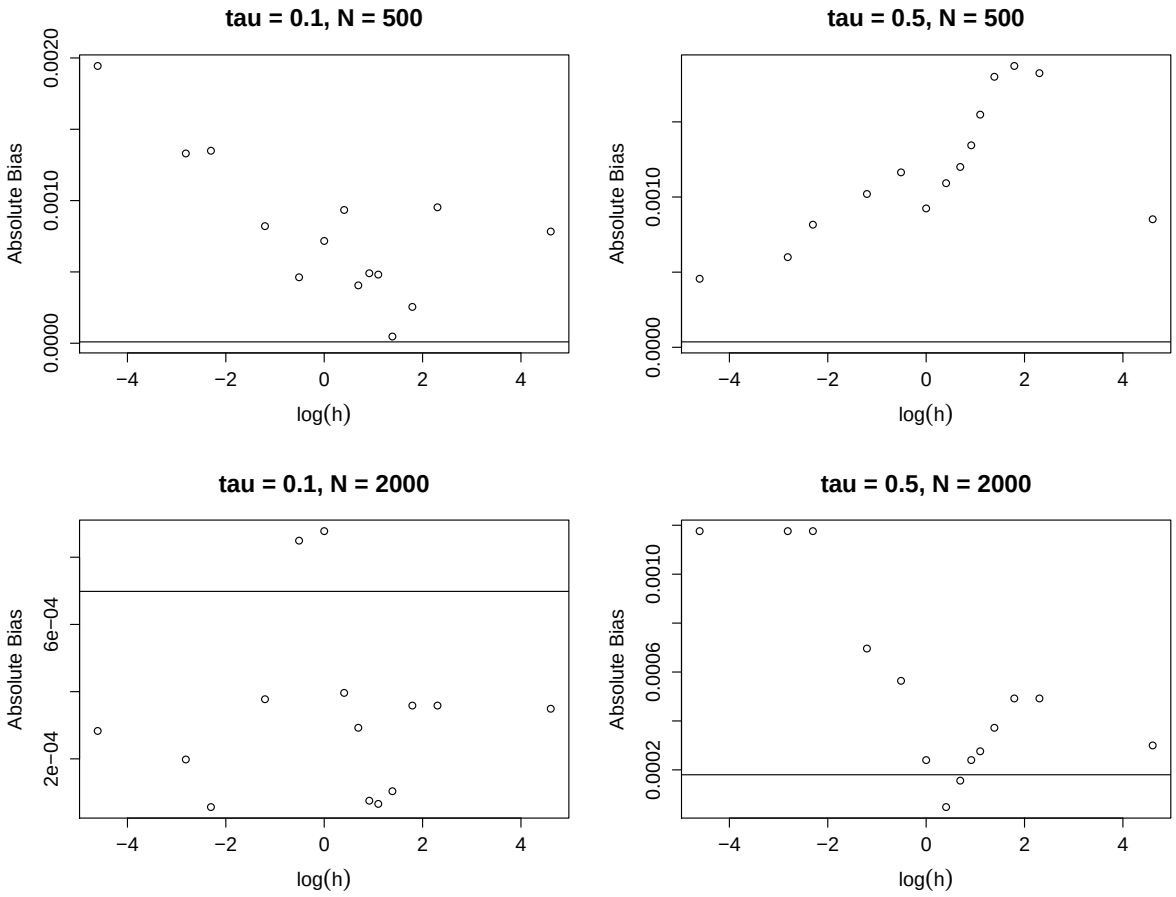


Figure 4.1: Mean absolute biases of the estimates with different bandwidths from Quantile Model (1). The horizontal line is the mean absolute bias of the estimated coefficients with CV selected bandwidth. The dots are the mean absolute biases of the estimated quantile coefficients with fixed bandwidths.

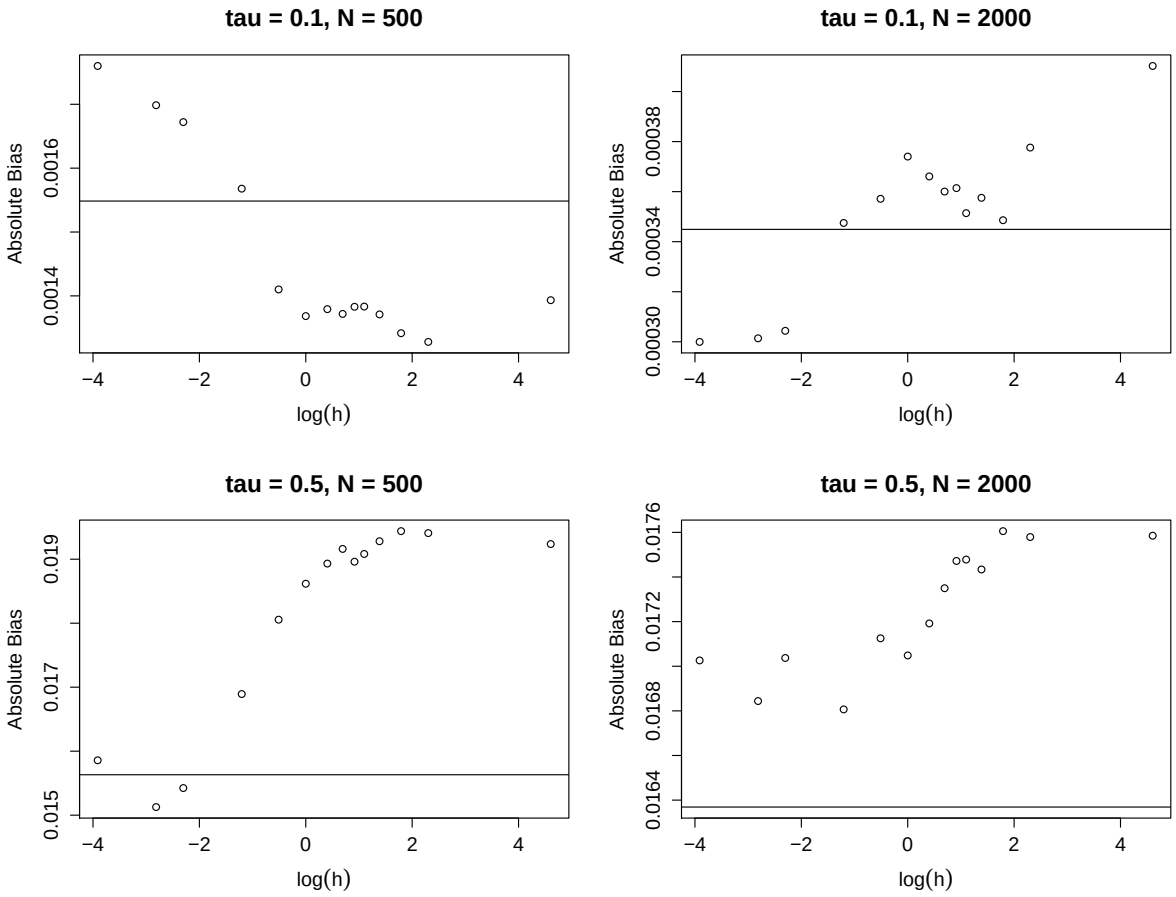


Figure 4.2: Mean absolute biases of the estimates with different bandwidths from Quantile Model (2). The horizontal line is the mean absolute bias of the estimated coefficients with CV selected bandwidth. The dots are the mean absolute biases of the estimated quantile coefficients with fixed bandwidths.

4.3 Further comparison with IPW under complex sampling schemes

The inverse-probability-weighting (IPW) technique has demonstrated comparable efficiency in Section 4.1 and slightly worse efficiency in Section 4.2 in comparison with SICO estimates for the secondary analysis of case-control data. The validation of IPW estimates, however, relies on a correct specification of the selection probabilities, which is often unknown unless nested within a large cohort study. In Section 4.1 and 4.2, we consider the simple sampling scheme that the selection probability only depends on the disease status. We have a representative random disease sample and a representative random control sample. In such case, the selection probability is homogeneous for all the cases and for all the controls, and IPW works well. In this section, we consider a complex sampling scheme that there exists an independent ancillary variable w following $N(0,1)$, and we oversample the subjects with $w > 0$. Specifically, the observations with positive w are 9 times more likely to be selected into the sample than subjects with $w < 0$ in both cases and controls. The second sampling scheme is known as stratified sampling, and is commonly used in survey designs for various reasons. For example, the National Maternal and Infant Health Survey oversampled the infants born with low birthweight ($\leq$ 2500 g) and very low birthweight ($\leq$1500 g) for the great research interests on the long term and short term health outcomes of these infants. We evaluate the finite sample performance of IPW in comparison with the proposed SICO and CE estimators under this condition.

Table 4.5 shows the relative bias, standard error, and mean squared error of the estimated coefficient β_1 under complex sampling scheme from 500 Monte-Carlo samples. We observe that when the positive w is over sampled, the IPW estimates suffer from inflated variance and bias, especially in quantile models. The proposed estimates are unaffected. As long as Y is a random sample given (D, X, Z), the resulting estimates are less affected by sampling schemes.

		Logistic			Interaction			Piecewise		
Model	Method	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE$\times n$
LogR	IPW	4.7	0.119	28.4	4.5	0.114	26.0	-1.5	0.112	25.0
	SICO (T=1)	0.8	0.080	12.9	0.1	0.081	13.0	-1.5	0.079	12.4
	SICO (T=10)	0.4	0.073	10.7	-1.4	0.072	10.5	-1.9	0.072	10.4
	SICO (T=100)	0.8	0.073	10.5	-1.4	0.072	10.4	-1.9	0.072	10.2
	CE	-3.7	0.074	11.2	-9.5	0.073	11.5	-1.9	0.072	10.3
LR	IPW	0.5	0.053	5.7	-0.7	0.048	4.6	0.0	0.052	5.5
	CE	1.2	0.033	2.1	0.1	0.031	2.1	-0.7	0.031	1.9

Table 4.5: The relative bias (RB), standard error (SE) and mean squared error (MSE) of the estimated coefficient β_1 for logistic and linear models under *complex sampling scheme*. The true value $\beta_1^* = 0.2$. "LogR" stands for the logistic regression. "LR" stands for the linear regression. "IPW" stands for inverse probability weighted quantile regression. "SICO (T)" stands for proposed SICO estimates with T replicate. "CE" stands for proposed CE estimates using kernel smoothing techniques for logistic regression and one-step optimization for linear regression.

		$\tau = 0.5$			$\tau = 0.9$		
Model	Method	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$
Quantile Model (1)	IPW	8.4	0.070	9.9	2.7	0.094	17.7
	SICO (T=10)	-0.5	0.041	3.4	0.6	0.054	5.8
Quantile Model (2)	IPW	-18.6	0.147	44.3	-106.0	0.312	251.0
	SICO (T=10)	-0.8	0.026	1.3	-1.0	0.102	20.7

Table 4.6: The relative biase (RB), standard error (SE) and mean squared error (MSE) of the estimated coefficient $\beta_{1,\tau}$ for Quantile Model (1) and (2) under *complex sampling scheme*. The true β_1 is $0.12 + 0.02Q_{e_i}(\tau)$. In Quantile Model (1), $e_i \sim N(0,1)$. In Quantile Model (2), $e_i \sim \chi_1^2/\sqrt{2}$. "IPW" stands for inverse probability weighted quantile regression. "SICO (T)" stands for proposed SICO estimates with T replicate. "CE" stands for proposed CE estimates using kernel smoothing techniques.

4.4 Type I error estimates in comparison with SPML

In Section 4.1, we observe that the SPML approach is more efficient than our proposed methods in Logistic Setting, but involves biases under Interaction and Piecewise Settings. In this section, we would like to investigate the type I error of proposed SICO estimates in comparison with SPML for the primary hypothesis $H_0 : \beta_1 = 0$ under the three settings we consider. We consider with a binary Y, a single pre-selected SNP with MAF of 10% to 50%. We simulate 100,000 Monte-Carlo samples with 2,000 cases and 2,000 controls. For the proposed SICO estimates, we consider a bootstrap procedure proposed in Section 3.6 for testing. Table 4.7 summarizes the Type I errors of the proposed SICO (T=10) and of the SPML method under the Logistic, Interaction and Piecewise settings and at the α levels of 0.05, 0.01 and 0.001. According to the table, the SICO has the correct Type I errors in all the settings we consider, while the SPML method has inflated type I error in the Interaction and the Piecewise Setting due to the deviation from the linear logistic model assumption.

		Logistic			Interaction			Piecewise		
Method	MAF	0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
SPML	0.1	0.04924	0.01008	0.00119	0.25043	0.10064	0.02311	0.09353	0.02449	0.00351
	0.2	0.04969	0.01017	0.00110	0.37414	0.17526	0.04829	0.12961	0.03881	0.00647
	0.3	0.04963	0.01038	0.00105	0.42709	0.21057	0.06665	0.15627	0.05126	0.00965
	0.4	0.05029	0.01002	0.00126	0.44216	0.22318	0.06905	0.17422	0.05983	0.01117
	0.5	0.05087	0.01051	0.00102	0.42486	0.21007	0.06234	0.17610	0.06265	0.01238
SICO	0.1	0.04944	0.01001	0.00095	0.05649	0.01219	0.00113	0.05104	0.01106	0.00119
(T=10)	0.2	0.04945	0.00997	0.00102	0.05585	0.01248	0.00108	0.05187	0.01143	0.00147
	0.3	0.04994	0.00973	0.00107	0.05903	0.01286	0.00134	0.05368	0.01123	0.00147
	0.4	0.05105	0.00958	0.00098	0.05943	0.01254	0.00132	0.05285	0.01163	0.00164
	0.5	0.04897	0.01013	0.00092	0.05912	0.01353	0.00135	0.05193	0.01181	0.00160

Table 4.7: Type I error of SICO estimates in comparison with SPML for a pre-selected SNP. "SPML" stands for semi-parametric maxmium likelihood based logistic regression. "SICO (T)" stands for proposed SICO estimates with T replicate.

4.5 The performance under biased estimated $\widehat{P}(D|X,\mathbf{Z})$

The proposed estimates require a consistently estimation of $P(D|X,\mathbf{Z})$. In previous simulation studies, we assumed that the primary disease prevalence P_0 is known, and we estimated $P(D|X,\mathbf{Z})$ assuming a linear logistic model as follows

$$P(D=1|X,\mathbf{Z}) = \exp(\gamma_0 + X\gamma_1 + \mathbf{Z}^T\boldsymbol{\gamma}_2)/\{1+\exp(\gamma_0 + X\gamma_1 + \mathbf{Z}^T\boldsymbol{\gamma}_2)\}.$$

In practice, we might encounter two issues. First, the disease prevalence often estimated from cohort studies or literature could be mis-specified. For this problem, we will demonstrate the performance of proposed estimators when estimated disease prevalence $\widehat{P}_0$ largely differs from the true value in both GLM and quantile regressions.

Second, the linear logistic model may not be a good approximation of the association between the disease status and the covariates $(X,\mathbf{Z})$. For example, when the disease prevalence is low, which is one of the main reasons to employ a case-control design, $P(D|X,Y,\mathbf{Z}) = \exp(\gamma_0 + X\gamma_1 + \mathbf{Z}^T\gamma_2 + Y\gamma_3)/\{1+\exp(\gamma_0 + X\gamma_1 + \mathbf{Z}^T\gamma_2 + Y\gamma_3)\} \approx \exp(\gamma_0 + X\gamma_1 + \mathbf{Z}^T\gamma_2 + Y\gamma_3)$. Consequently, the logistic model also holds for $P(D=1|X,\mathbf{Z})$ if Y follows an exponential family distribution. When the disease prevalence is high, however, this approximation may not work and the estimation using the linear logistic model may be biased. In the simulation studies in Section 4.1, we have already considered the performance of the proposed estimates under three disease models, Logistic, Interaction and Piecewise Setting. Here, we further investigate this problem under quantile framework by demonstrating the performance of SICO and CE estimates when the disease prevalence P_0 is high and therefore the logistic model for $P(D|X)$ is mis-specified.

Table 4.8 and 4.9 show the relative bias, standard error and mean squared error of the estimated coefficients from 500 Monte-Carlo replicates with various $\widehat{P}_0$ values used for estimation. For Table 4.8, we use data from Logistic Model (Model 4.1) and Linear Model (Model 4.2) in Section 4.1, but re-estimate the parameters based on different estimated disease prevalence $\widehat{P}_0$ ranging from $P_0/2$ to $2P_0$, where P_0 =10% is the true prevalence. For Table 4.9, we use the data from Quantile Model (1) (Model 4.3 with normal error) in

	Logistic Model (SICO (T=10))			Linear Model (CE)		
$\widehat{P}_0$	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$
$P_0/2$	-3.5	0.075	11.3	-6.0	0.032	2.4
$P_0/1.5$	-2.0	0.073	10.7	-4.3	0.032	2.1
$P_0/1.2$	-0.7	0.071	10.2	-2.7	0.031	2.0
P_0	-0.5	0.074	10.8	-1.1	0.030	1.9
$1.2P_0$	2.1	0.068	9.3	0.6	0.030	1.8
$1.5P_0$	4.0	0.066	8.7	2.9	0.029	1.7
$2P_0$	6.8	0.062	8.0	6.2	0.028	1.8

Table 4.8: The relative bias (RB), standard error (SE) and mean squared error (MSE) of the estimated coefficient β_1 with misspecifed $\widehat{P}_0$ under *Logistic Model* and *Linear Model*. $P_0 = 10\%$ is the true disease prevalence.

Section 4.2, and also re-estimate the parameters based on estimated $\widehat{P}_0$ ranging from $P_0/2$ to $2P_0$. We find the estimation bias does increase slowly as the prevalence deviates from the true one, but the differences are small even when doubling P_0.

In Table 4.10, we simulate the data according to Quantile Model (2) (Model 4.3 with χ^2 error) in Section 4.2, but the true disease prevalence P_0 increases from 5% to 30%. In all the cases, the relative biases from the SICO estimates are smaller than 4%, which indicates its robustness against the deviation from the logistic $P(D|X, \mathbf{Z})$. The CE estimates are relatively more sensitive to the bias under Quantile Model (2) with higher disease prevalence.

		$\tau = 0.5$			$\tau = 0.9$		
$\widehat{P}_0$	Method	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$
$P_0/2$	CE	-9.1	0.043	4.0	-7.9	0.059	7.2
	SICO (T=10)	-8.7	0.043	3.9	-7.5	0.058	6.9
$P_0/1.5$	CE	-7.8	0.043	3.9	-6.4	0.059	7.0
	SICO (T=10)	-7.2	0.043	3.8	-5.9	0.057	6.6
$P_0/1.2$	CE	-6.3	0.043	3.8	-4.4	0.058	6.7
	SICO (T=10)	-5.9	0.042	3.7	-4.2	0.056	6.4
P_0	CE	-4.7	0.042	3.6	-2.0	0.056	6.3
	SICO (T=10)	-4.5	0.042	3.6	-2.7	0.056	6.2
$1.2P_0$	CE	-3.7	0.042	3.5	-1.1	0.055	6.1
	SICO (T=10)	-2.9	0.042	3.5	-0.9	0.055	6.0
$1.5P_0$	CE	-1.1	0.041	3.4	1.2	0.055	6.0
	SICO (T=10)	-0.4	0.041	3.4	1.7	0.054	5.7
$2P_0$	CE	3.4	0.040	3.3	5.2	0.052	5.5
	SICO (T=10)	3.4	0.040	3.3	5.8	0.052	5.4

Table 4.9: The relative bias (RB), standard error (SE) and mean squared error (MSE) of the estimated quantile coefficients $\beta_{1,\tau}$ with misspecifed $\widehat{P}_0$ under *Quantile Model (1)* at quantile levels 0.5 and 0.9. $P_0 = 5\%$ is the true disease prevalence.

		$\tau = 0.5$			$\tau = 0.9$		
Prev	Method	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$
0.05	SICO (T=10)	2.0	0.026	1.3	0.2	0.098	19.3
	CE	2.2	0.026	1.4	-0.9	0.100	19.8
0.1	SICO (T=10)	-0.1	0.024	1.1	-2.3	0.089	15.9
	CE	1.5	0.025	1.2	8.8	0.095	18.4
0.2	SICO (T=10)	0.2	0.021	0.9	-3.6	0.081	13.2
	CE	2.6	0.022	1.0	8.5	0.089	16.2
0.3	SICO (T=10)	0.2	0.019	0.7	-1.9	0.079	12.4
	CE	3.1	0.020	0.9	12.3	0.086	15.4

Table 4.10: The relative bias (RB), standard error (SE) and mean square error of the estimated coefficients $\beta_{1,\tau}$ under *Quantile Model (2)* at quantile levels 0.5 and 0.9. "SICO (T)" stands for proposed SICO estimates with T replicate. "CE" stands for proposed CE estimates using kernel smoothing techniques.

Chapter 5

Applications

5.1 Overview

In this chapter, we apply the proposed new estimating equation based approach to conduct the secondary analysis in two different contexts, the Risk Assessment of Cerebrovascular Event study and the New York University Bellevue Asthma Study. In the first example, we consider the highly prevalent complex disease diabetes as the binary secondary phenotype using the young onset stroke case-control data. Diabetes is strongly associated with the young onset stroke, and for the SNPs that are also associated with stroke, the traditional methods may be largely bias. We demonstrate the performance of our new estimating equations based approach in comparison with comparable existing methods in the literature. In addition, we investigated the algorithm problem of SPML approach we encounter in this real data example. In the second example, we consider the association of TSLP gene with a continuous secondary phenotype, serum IgE level, using asthma case-control data. We first consider the mean level associations, and apply IPW, SPML and the proposed CE estimate for the secondary analysis under the linear regression framework. Second, we extend the analysis to quantile regression to obtain a more comprehensive picture of the genetic association with the secondary outcome. We demonstrate the attractive properties of secondary quantile regression in comparison of mean regression in different SNPs we analyzed. These examples clearly present the value of the new estimating equation based

approach in the secondary trait analysis in genetic case-control studies.

5.2 Application to Risk Assessment of Cerebrovascular Events (RACE) Study

To illustrate the application of these methods, we select two SNPs, rs6712932 and rs1990760, from the Risk Assessment of Cerebrovascular Events (RACE) Study to estimate their genetic association with diabetes. RACE is a GWA study available in the dbGaP database that includes 1,220 young onset stroke cases and 1,273 controls from the Risk Assessment of Cerebrovascular Events Study in Pakistan [Cornelis et al., 2010]. The SNPs rs6712932 and rs1990760 have been found genome-wide significant in several previous studies with diabetes [Salonen et al., 2007; Todd et al., 2007]. SNP rs6712932-G is reported to be a protective factor for type-2 diabetes with OR=0.66 (CI: 0.54 - 0.79) [Salonen et al., 2007] and SNP rs1990760-G is reported to be protective from type-1 diabetes with OR=0.85 (CI: 0.81 - 0.90) [Todd et al., 2007]. In this section, we evaluate the association between these two SNPs and diabetes using this case-control dataset. For notation, we let $D = \{0,1\}$ denote the primary case-control status of young onset stroke, $Y = \{0,1\}$ denote the binary secondary phenotype of diabetes, $X = \{0,1,2\}$ denote the count of minor alleles for each of the two SNPs, and Z denote a continuous variable, the propensity score [Guo and Fraser, 2010] developed from a set of covariates including age, gender, smoking status, coronary artery disease, myocardial infarction and the top 10 principle components from population stratification using EIGENSTRAT [Price et al., 2006a]. The association between secondary phenotype diabetes Y and each of the pre-selected SNPs X is modeled by the following model

$$P(Y=1) = \frac{\exp(\beta_0 + \beta_1 X + \beta_2 Z)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 Z)},$$

where β_1 is the coefficient of interest that relates diabetes to the SNP.

Both types of diabetes are known to be risk factors for stroke [Peters et al., 2014; Sundquist and Li, 2006]. In this dataset, we only have information on whether a subject

has diabetes without further specification on the type of diabetes. In the primary analysis of the risk factors for young onset stroke, we estimate the OR of having diabetes is 3.18 (p-value$<$ 0.0001) for the young onset stroke. In addition, both SNPs are associated with young onset stroke with marginal per-minor-allele OR as 1.14 (p-value=0.024) and 0.87 (p-value=0.015), respectively. After adjusting for diabetes, the associations between these SNPs and stroke remain significant (rs6712932: OR=1.16; p-value=0.017. rs1990760: OR=0.84; p-value=0.003). This is a scenario where the commonly-used tradition methods may provide biased estimation for the association between these two SNPs and diabetes, and we should apply appropriate approaches for estimation.

We evaluated the association between diabetes and the SNPs using all the methods we considered in the simulations, including (1) regression using cases only, (2) regression using controls only, 3) regression using combined case-control sample, (4) regression using case-control sample adjusting for case-control studies, (5)IPW, (6) SPML, (7) the proposed SICO estimates and (8) the proposed CE estimates. The prevalence of stroke in adult Pakistan population is needed for the estimating equations and the SPML approach, and we estimated it to be approximately 3.6% by using our best knowledge from the literature [Pakistan Stroke Society, 2006]. For IPW approach, there is no clear selection probability, so we used the disease prevalence to approximate the ratio of selection probabilities between cases and controls and used bootstrap method to construct bootstrap standard errors and p-values for inference. For the proposed SICO and CE methods, bootstrap tests were used to construct the standard error and calculate the p-value.

Table 5.1 presents the results of the association between the SNPs and diabetes using the eight different approaches. For SNP rs6712932, all of these methods yielded similar effect estimates suggesting no significant association between rs6712932 and diabetes. We observe the directions for the genetic association with stroke and diabetes are opposite, which might also suggest the marker has little or no true association with diabetes in this general population. When there is no or little association between the tested marker and the secondary trait, the biases of the traditional methods are small, which is consistent with the findings in Monsees et al. [2009]. Among the all the theoretical unbiased methods we

consider, SPML approach is the most efficient, followed by CE estimate and SICO estimate, and IPW is the least efficient.

For the SNP rs1990760, there is a significant protecting association among controls. While in the same direct, the association among cases has different value and is not significant. Regression on the case-control sample ignoring the sampling scheme or adjusting for the primary disease produce point estimates (-0.126 and -0.159) that are closer to the one among cases (-0.112) than that of controls (-0.247). Since cases constitutes a small percentage of general population, it suggests the estimates from the two methods may be biased from the true association in general population. The proposed SICO and IPW detect similar significant protecting associations between SNP rs1990760 and diabetes that are closer to the controls ($\hat{\beta} = -0.227$, p-value=0.0374 for SICO; $\hat{\beta} = -0.228$, p-value=0.0365 for IPW). We expect their estimates are closer to the true value in the population. CE estimate is more efficient but potential contains some biases ($\hat{\beta} = -0.142$, p-value=0.0379). This is because the estimation process involves kernel smoothing techniques, and it does not work well when there are few observations in the neighborhood of (X, Z). In this particular example, the covariate Z has heavy tail among cases.

Interestingly, we observe that the SPML method fails to generate an estimate for SNP rs1990760 in Table 5.1 due to an algorithm problem in their software. To further investigate the problem of the SPML approach, we apply the method to the top 1000 SNPs selected from the association analysis with the primary disease (p-value<1.5E-3). The failure rate of the SPML approach in generating an estimate is presented in figure 5.1. We can see that when the SNP is strongly associated with the primary disease, which is also the case that the traditional methods are likely to bias, the SPML approach is highly likely to fail. For example, among the top 20 SNPs (p-value<1.5E-10) , the failure rate is as high as 40%. This is because when the continuous covariate enters into the algorithm, it is treated as a high-dimensional nuisance parameter of the likelihood function that has to be profiled out at each value. When we conduct the same analysis without adjusting for the covariates, the SPML approach works well.

In summary of this example, when the tested marker X has little or no association with

	rs6712932				rs1990760			
	Est $\hat{\beta}_1$	OR	S.E.	p-value	Est $\hat{\beta}_1$	OR	S.E.	p-value
Case-control	-0.060	0.94	0.074	0.4166	-0.126	0.88	0.068	0.0648
Case	-0.100	0.90	0.092	0.2800	-0.112	0.89	0.088	0.2024
Control	-0.125	0.88	0.132	0.3441	-0.247	0.78	0.118	0.0368
Stratified case-control	-0.108	0.90	0.076	0.1530	-0.159	0.85	0.071	0.0246
IPW	-0.113	0.89	0.134	0.4009	-0.228	0.80	0.109	0.0365
SPML	-0.099	0.91	0.075	0.1888	NA	NA	NA	NA
SICO (T=10)	-0.098	0.91	0.127	0.4381	-0.227	0.80	0.109	0.0374
CE	-0.106	0.90	0.089	0.2336	-0.142	0.86	0.068	0.0379

Table 5.1: The association between two pre-selected SNPs (rs6712932 and rs1990760) and diabetes in a young onset stroke case-control sample. "Case-control" stands for unadjusted logistic regression using both case and control samples. "Case" stands for unadjusted logistic regression using case sample only. "Control" stands for unadjusted logistic regression using control sample only. "Stratified" stands for logistic regression using both case and control samples adjusting for primary disease status. "IPW" stands for inverse probability weighted logistic regression. "SPML" stands for semi-parametric maxmium likelihood based logistic regression. "SICO (T)" stands for proposed SICO estimates with T replicate. "CE" stands for proposed CE estimates using kernel smoothing techniques.

the secondary trait Y, all methods provide valid estimates. When the tested marker X is potentially associated with the secondary trait Y, the traditional methods may contain biases. Among the theoretical justified unbiased approaches, IPW and SICO estimates provide robust and similar estimation for the association between the tested marker and secondary trait. SPML is more efficient, but is subject to biases from violation of model assumptions and algorithm problems. CE estimate requires kernel smoothing approximations in logistic regression and might be biased if the sample is unbalanced in cases and controls. In the next example, we consider CE estimate for linear regression, which does not need the kernel smoothness for its estimation. It provides similar point estimate but is more efficient

than IPW approach, which indicates its value in analyzing the secondary trait.

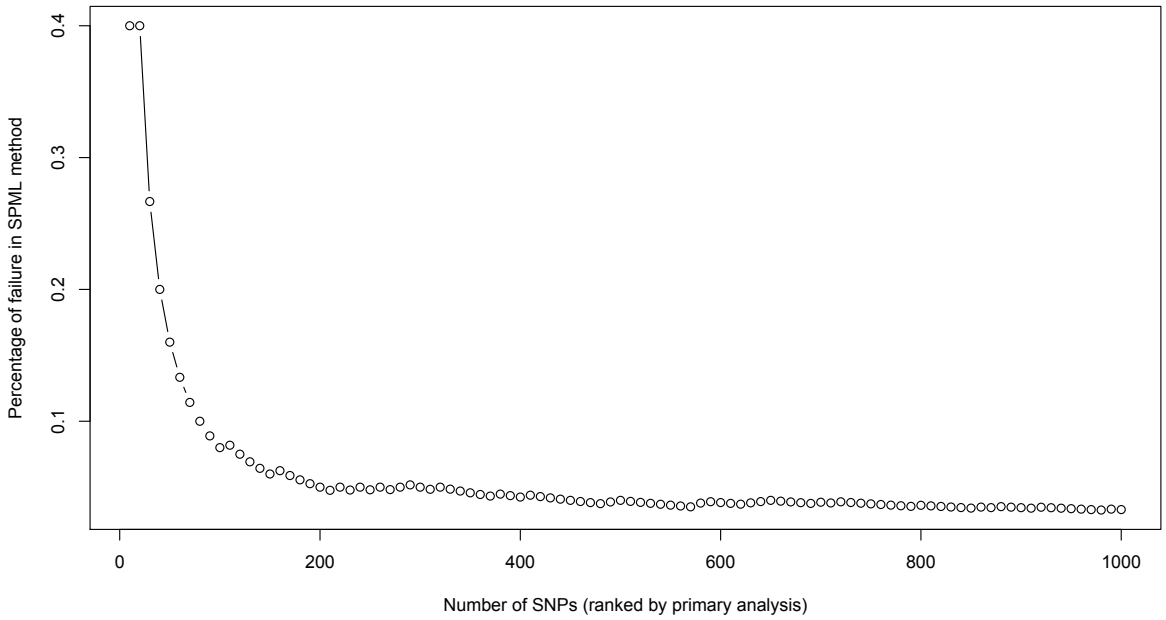


Figure 5.1: The failure rate of SPML method

5.3 Application to New York University Bellevue Asthma Study

In this section, we apply the proposed methods to study the association between the Thymic stromal lymphopoietin (TSLP) gene and serum IgE levels from the New York University Bellevue Asthma Registry [Liu et al., 2011]. The study consisted of 387 asthmatics and 212 healthy controls, and measured 10 tag SNPs in the TSLP gene. IgE is a class of antibody that is elevated in various allergic diseases. Understanding the genetic basis of IgE paves to way to recognize the mechanism of TSLP in affecting asthma and other allergic diseases. Therefore, the purpose of the secondary analysis is to identify the TSLP SNPs that are associated with elevated serum IgE level. Since log serum IgE level is approximately normally distributed among cases and controls, we first consider least square regression and compare the results with existing novel methods including IPW and SPML. Second, as elevated IgE level instead of mean IgE level plays an essential role in allergic diseases, we are also interested to apply the quantile regression to further investigate the genetic association with upper quaniles of the IgE. We illustrate the two types of regressions separately.

5.3.1 Mean regression

We denote X as the minor allele count for each of the 10 TSLP SNPs, Z as a continuous variable derived as the first principal component score from 213 ancestry informative markers to adjust for population stratification, and Y as the log serum IgE level. Then the least square model we consider is as follows

$$Y = \beta_0 + \beta_1 X + \beta_2 Z.$$

Three approaches are used to estimate the coefficient β_1: the IPW approach, SPML approach, and proposed CE approach. We calculate the overall asthma prevalence as 10.1% based on 6 birth cohort studies, and this information is used to approximate selection probability in IPW, and estimate $P(D|X,Y,Z)$ in SPML and $P(D|X)$ in CE. The standard

errors and p-values in proposed CE estimates were calculated using bootstrap, i.e. we bootstrap cases and controls separately, and re-apply the entire estimating procedure to the bootstrap case-control sample.

The resulting estimated coefficients are summarized in Table 5.2. The point estimates of the CE coefficients are very similar to IPW, while the estimates from SPML are largely different in many SNPs. We know from simulations in Chapter 4 that IPW is robust and SPML is efficient but potentially biased with mis-specified $P(D|X,Y,Z)$. To understand if SPML is potentially biased, we further tested the $X-Y$ interactions in $P(D|X,Y,Z)$ to understand the underlying models, and some of the interactive effects are significant (SNPs rs2289278 and rs10035870). Therefore, we believe that the SPML approach is substantively biased due to the violation of the model assumptions, and the proposed CE estimates as well as IPW provides relative unbiased estimations. The Table 5.2 also shows that the proposed CE estimates are more efficient than IPW approach in analyzing the genetic associations. In details, IPW only detects one significant SNP rs10035870 at α-level 0.05 , while CE identifies three (rs11466741, rs11466743 and rs10035870). In summary, the proposed new estimating equation based approach (CE estimator in particular) combines the advantages of IPW and SPML estimators that its is robust and fairly efficient in estimating the marker-secondary trait associations. Therefore, it is useful in real data analysis to discover potential SNPs.

5.3.2 Quantile regression

In this section, we consider the quantile regression for the association between TSLP gene and the upper quantiles of the log serum IgE level. We are particularly interested in discovering the SNPs that are associated with elevated IgE levels, and quantile regression is able to present a comprehensive picture on the where the effects of the SNPs exist on the distribution of serum IgE levels. The quantile model we consider is as follows:

$$Q_Y(\tau) = \beta_{0,\tau} + \beta_{1,\tau}X + \beta_{2,\tau}Z,$$

SNP	Method	Est.	S.E.	p-value	SNP	Method	Est.	S.E.	p-value
rs2289276	IPW	0.09	0.14	0.500	rs2289278	IPW	-0.29	0.22	0.193
	SPML	0.06	0.10	0.569		SPML	0.05	0.16	0.764
	CE	0.10	0.08	0.252		CE	-0.29	0.16	0.077
rs1898671	IPW	-0.12	0.17	0.505	rs11241090	IPW	0.14	0.33	0.658
	SPML	-0.12	0.10	0.262		SPML	0.20	0.22	0.377
	CE	-0.11	0.13	0.380		CE	0.14	0.24	0.554
rs11466741	IPW	0.20	0.12	0.112	rs10035870	IPW	0.63	0.28	0.024
	SPML	0.03	0.09	0.721		SPML	0.03	0.21	0.904
	CE	0.20	0.09	0.032		CE	0.63	0.22	0.004
rs11466743	IPW	-0.61	0.34	0.073	rs11466749	IPW	0.07	0.24	0.779
	SPML	-0.25	0.29	0.382		SPML	0.14	0.15	0.343
	CE	-0.61	0.29	0.034		CE	0.07	0.16	0.678
rs2289277	IPW	0.13	0.12	0.282	rs11466750	IPW	-0.05	0.16	0.748
	SPML	0.03	0.09	0.739		SPML	0.08	0.12	0.484
	CE	0.13	0.09	0.165		CE	-0.05	0.13	0.671

Table 5.2: Estimated mean allelic effects on log serum IgE level in *linear regression*. "IPW" stands for inverse probability weighted least squared regression. "SPML" stands for semi-parametric maximum likelihood least squared regression. "CE" stands for proposed CE estimates using one-step optimization

where the X is the minor allele count for each of the 10 TSLP SNPs, Z is a continuous variable derived as the first principal component score from 213 ancestry informative markers to adjust for population stratification, and Y is the log serum IgE level.

To evaluate effects of the TSLP gene variants on different levels of IgE, we estimated the model at quantile levels of 0.15, 0.25, 0.5, 0.75 and 0.85, respectively. Three approaches were used to estimated the quantile coefficients: the IPW approach and the proposed SICO and CE methods. SPML approach is not considered for quantile regression as it is based on likelihood function and can not be applied to non-parametric regressions. Similar to the simulation studies, we use a Gaussian kernel and select the bandwidth using 5-fold cross-validation for the CE estimates. The resulting estimated quantile coefficients are summarized in Table 5.3. All the p-values in Table 5.3 were calculated using bootstrap, i.e. we bootstrap cases and controls separately, and re-apply the entire estimating procedure to the bootstrap case-control sample. The estimated quantile coefficients from the three approaches are comparable. However, due to the small sample size in this particular example, the bootstrap standard errors of the CE estimates and IPW estimates are much bigger than the ones from SICO estimates. Consequently, the SICO estimates are more powerful to detect the quantile associations with small sample sizes.

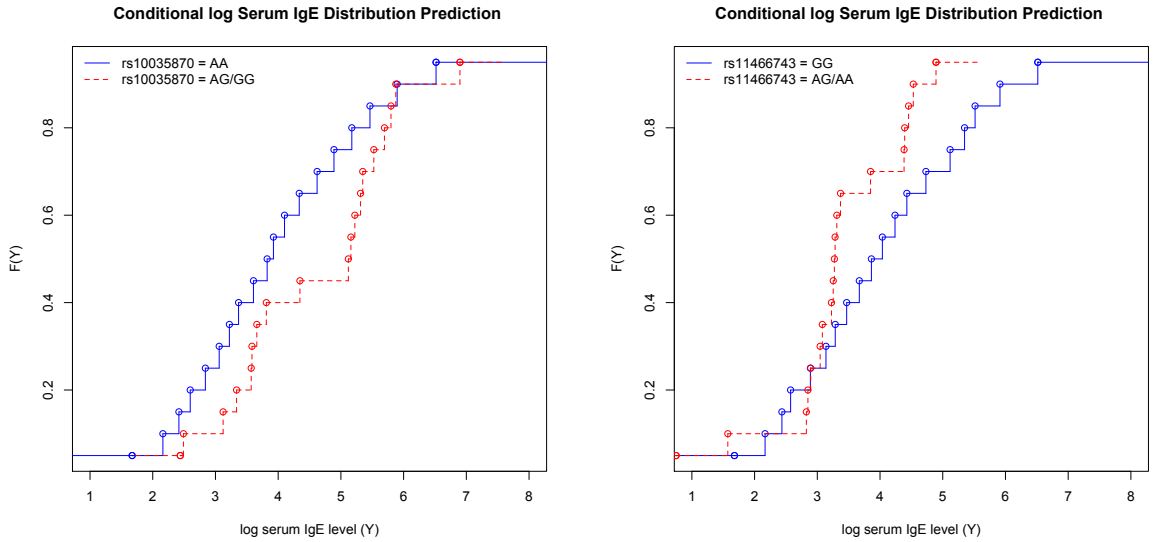
From the mean coefficients output for CE estimates in Table 5.2, we observed that SNPs rs11466741, rs11466743 and rs10035870 had significant associations with mean serum IgE level, with p-values of 0.032, 0.034 and 0.004, respectively. The results from quantile regressions also indicated significant association with these SNPs, and these associations remain significant even after a conservative Bonferroni correction for estimating different quantile levels and the number of SNPs. Moreover, quantile analysis presented a more comprehensive picture on the effects of the SNPs and suggested that the SNPs have different impact on the distribution of serum IgE level. For example, having one or two A allele of SNP rs11466743 decreases the mean of IgE value by 0.61. Based on the quantile analysis, however, this SNP has no effect on the lower quantiles (0.15th and 0.25-th quantiles) of IgE value, but significantly decreases the median and upper (0.75th and 0.85th) quantiles of IgE by 0.6, 1.2 and 1.1, respectively. In addition, elevated serum IgE level indicates

hypersentitive allergic effect and thus it is important to know the TSLP effects on the upper quantile of serum IgE level. Specifically, the propose method showed that SNPs rs2289276, rs2289278, rs2289277 and rs11466750 have significant association with 75th quantile of log serum IgE level; however, the mean regression did not indicate significant association, illustrating the potential for the new approach to discover new associations.

Moreover, to see how genetic variants impact the distribution of serum IgE level, we estimate the quantile coefficients on a fine grid of quantile levels. In Figures 5.2(a) and 5.2(b), we plot the estimated conditional distribution functions with different genotypes at SNPs rs10035870 and rs11466743, respectively. Specifically, the solid curve in Figure 5.2(a) is the estimated quantile function for the patients whose genotype at rs10035870 is AA, and the dashed line is that of those whose genotype is AG/GG at rs10035870. In Figure 5.2(b), the solid curve is the estimated quantile function with genotype GG at rs11466743, and the dashed line is that of genotype AG/AA.

Both SNPs were found to have significant impact on the distribution of serum IgE level. Based on Figure 5.2(a), rs10035870 has strong positive effect on the entire distribution of serum IgE level, and thus subjects with the mutation allele of rs10035870 tend to have higher serum IgE level in general. In contrast, SNP rs11466743 only has strong impact on the median and upper quantiles, but makes little difference at the lower quantiles of serum IgE level. As indicated in Figure 5.2(b), the subjects with genotype AG/AA in rs11466743 are less likely to have a very high serum IgE level compared to those with genotype GG, however, they also have equal chance to have low IgE serum level. For example, for the subjects with rs11466743 genotype AG/AA, the probability of hypersensitive allergic effects ($Y > 5$) is nearly zero. However, for the subjects with genotype GG, this probability is approximately 30%. However, the probabilities of log IgE serum level ($Y < 3$) are 25% for all the genotypes.

In the original case-control asthma study, we found that the SNP rs1898671 was associated with the asthma disease risk. When examining the genetic association with the serum IgE levels, we identified different associated SNPs. Asthma is an allergic immune disorder that is usually diagnosed based on the pattern of symptoms, response to therapy over time



(a) rs10035870 (b) rs11466743

Figure 5.2: The estimated distribution functions of log serum IgE level associated with SNP rs10035870 and rs11466743.

and spirometry. With allergic effects, serum IgE levels may be normal or sub-normal. In addition, the elevation of serum IgE levels may be caused by different allergic diseases than asthma. Therefore, it is nature to find out the analysis of the secondary outcome serum IgE levels discovers different SNPs from the primary analyses.

		$\tau = 0.15$		$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.85$	
SNP	Method	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value
rs2289276	IPW	-0.1	5.0E-01	-0.1	6.9E-01	0.4	1.1E-02	0.5	6.1E-02	0.1	6.9E-01
	SICO	-0.1	6.4E-02	0.0	9.7E-01	0.4	3.8E-07	0.3	8.7E-04	0.0	7.5E-01
	CE	-0.1	3.4E-01	0.0	9.2E-01	0.4	2.6E-02	0.3	1.1E-01	-0.1	7.4E-01
rs1898671	IPW	-0.1	5.4E-01	-0.3	9.5E-02	-0.3	2.1E-01	0.3	3.9E-01	-0.1	7.1E-01
	SICO	-0.2	9.2E-03	-0.2	3.9E-03	-0.2	8.8E-03	0.2	6.7E-02	0.0	9.8E-01
	CE	-0.1	5.1E-01	-0.2	2.6E-01	-0.2	4.8E-01	0.2	6.0E-01	0.2	5.9E-01
rs11466741	IPW	-0.1	6.5E-01	0.1	5.1E-01	0.5	3.6E-03	0.4	1.5E-02	0.2	3.9E-01
	SICO	0.0	4.1E-01	0.1	1.7E-01	0.4	8.4E-08	0.3	3.5E-04	0.1	3.6E-01
	CE	-0.1	6.8E-01	0.1	4.7E-01	0.4	1.1E-02	0.4	4.3E-03	0.0	9.6E-01
rs11466743	IPW	0.4	6.7E-01	0.0	9.4E-01	-0.5	8.2E-02	-1.1	3.6E-02	-1.1	1.4E-02
	SICO	0.4	1.6E-01	0.0	9.7E-01	-0.6	1.2E-03	-1.2	1.2E-11	-1.1	5.4E-08
	CE	0.4	6.6E-01	-0.1	8.9E-01	-0.6	7.4E-02	-1.2	3.5E-03	-1.2	5.0E-04
rs2289277	IPW	-0.1	5.5E-01	-0.1	7.1E-01	0.3	2.2E-01	0.3	1.3E-01	0.2	4.6E-01
	SICO	-0.1	7.1E-02	0.0	5.5E-01	0.3	1.7E-04	0.3	5.7E-03	0.1	3.7E-01
	CE	-0.1	6.5E-01	0.0	9.2E-01	0.3	2.0E-01	0.3	5.2E-02	0.0	9.1E-01
rs2289278	IPW	-0.1	7.1E-01	-0.3	2.7E-01	-0.1	7.6E-01	-0.5	2.3E-01	-0.5	3.1E-01
	SICO	-0.1	1.6E-01	-0.3	2.1E-03	-0.1	2.2E-01	-0.5	3.4E-04	-0.3	7.1E-02
	CE	-0.1	6.6E-01	-0.3	2.5E-01	-0.2	5.1E-01	-0.5	1.6E-01	0.0	9.7E-01
rs11241090	IPW	0.4	4.7E-01	0.4	3.4E-01	-0.3	5.9E-01	-0.1	9.4E-01	0.4	6.1E-01
	SICO	0.4	3.2E-02	0.3	2.3E-02	-0.3	1.6E-01	0.0	9.3E-01	0.3	2.8E-01
	CE	0.4	3.4E-01	0.4	3.6E-01	-0.3	5.4E-01	-0.2	7.6E-01	0.5	4.8E-01
rs10035870	IPW	0.7	1.1E-01	0.5	2.3E-01	0.8	2.6E-01	0.8	2.0E-02	0.4	3.1E-01
	SICO	0.8	1.1E-09	0.6	9.0E-07	0.9	5.4E-07	0.8	3.1E-04	0.4	3.7E-02
	CE	0.7	4.2E-02	0.5	1.1E-01	0.7	2.9E-01	0.7	6.8E-03	0.3	4.5E-01
rs11466749	IPW	-0.2	3.8E-01	-0.2	5.3E-01	0.1	8.8E-01	0.2	6.9E-01	0.3	5.6E-01
	SICO	-0.2	2.2E-02	-0.2	6.8E-02	0.0	7.1E-01	0.1	2.7E-01	0.3	7.9E-03
	CE	-0.2	3.3E-01	-0.2	4.2E-01	0.0	9.2E-01	0.1	8.3E-01	0.4	2.6E-01
rs11466750	IPW	-0.1	5.0E-01	-0.1	6.6E-01	-0.3	8.2E-02	-0.2	5.1E-01	-0.1	8.7E-01
	SICO	-0.2	3.3E-02	-0.1	5.5E-02	-0.3	2.1E-04	-0.3	4.2E-03	0.1	4.2E-01
	CE	-0.1	5.0E-01	0.0	8.4E-01	-0.3	1.1E-01	-0.3	3.4E-01	0.2	5.9E-01

Table 5.3: The estimated allelic effects on log serum IgE level in *quantile regression* at quantile levels of 0.15, 0.25, 0.5, 0.75, and 0.85. "IPW" stands for inverse probability weighted quantile regression. "SICO" stands for proposed SICO estimates with 10 replicate. "CE" stands for proposed CE estimates using kernel smoothing techniques.

Chapter 6

Conclusions and future work

In this paper, we propose a general framework to estimate the genetic association with secondary phenotype in case-control studies. In this chapter, we discuss and summarize the advantages as well as some limitations of the proposed approach. In the first section, some important conclusions are listed. We then point out a few possible future directions on this topic in the second section.

6.1 Conclusions

We propose a new estimating equation based approach to estimate the association between genetic variants and secondary phenotype in the case-control designs. It combines observed and counter-factual outcomes to constitute unbiased estimating equations. Compared with the existing unbiased approaches, including the IPW and SPML, it has the following attractive features.

First, it can accommodate various types of phenotypes (e.g., binary, continuous and ordinal), SNPs models (additive, dominant or recessive), and regressions (e.g., least square regression, GLM, quantile regression). Likelihood function based approaches, such as SPML approach, lack of the flexibility and diversity.

Second, it can easily accommodate covariates, including population substructure, an important confounding in genetic studies. When we encounter a large number of covariates, it is easy to conduct variable selection using cross-validation or introduce penalty

functions, like Lasso, Elastic Net, and SCAD.

Third, it relaxes the common conditions on the disease prevalence models. Except for the IPW approach, most secondary trait analyses using the entire case-control sample assume logistic models of $P(D|X,Y,\mathbf{Z})$. Our proposed estimating equations approach is more general, and only assumes that the disease probability follows a logistic model with just $(X,\mathbf{Z})$; it is not affected by the underlying distribution of Y. In section 4.1, while most existing methods are biased when $P(D|X,Y,Z)$ doesn't follow the logistic model in the Piecewise Setting, our proposed estimating equations are robust to the model mis-specification. Certainly, there is a price to be paid for the flexibility of the model assumptions. When the underlying $P(D|X,Y,\mathbf{Z})$ satisfies the model assumptions in the SPML approach, the SPML method shows greater efficiency than our proposed approach. As expected, there is a trade-off between the robustness and efficiency of the models, and depending on the underlying true model a different method may be optimal.

Fourth, the proposed approach is not sensitive to sampling schemes. Although the IPW approach is a simple and flexible method that works for any models, it requires knowing the additional information on sampling probabilities. As shown in the simulations in Section 4.3, the resulting estimates may not be efficient under some types of sampling schemes. For example, under the complex sampling schemes, where the some sampling variables are unrelated to the disease status, the IPW estimates are very inefficient.

Fifth, it requires little external information to be known. In secondary trait analyses, it is hard to obtain unbiased estimation without any external information. The proposed method only requires the disease prevalence P_0 to be specified. In contrast, the IPW approach needs the selection probabilities, which is often hard to obtain when the case-control sample is not nested within a larger cohort study. The bias correction approach by [Wang and Shete, 2011] also needs the prevalence of the secondary phenotype in addition to P_0. In principle, the SPML approach might be able to estimate β_1 without external information on disease prevalence, but the resulting inference is unstable Li and Gail [2012] and their publicly available software at http://www.bios.unc.edu/ lin/softward/SPREG/ requires knowing P_0. Thus, no unbiased approaches are able to perform without external

information of P_0 or selection probability. When mis-specified, theses methods are subject to biases. However, simulations showed the new estimating equations are fairly robust to such mis-specification.

Finally, it is computationally simple and straightforward. The estimation can be achieved by weighted regressions. Hence the computation does not require special software.

In summary, the construction of the estimating equations is straightforward and computationally efficient by simulating pseudo observations and evaluating the expected counterfactual estimating function. It provides robust and fairly efficient unbiased estimation for the variant-secondary phenotype association. It has appropriate type I error rate and is robust to disease prevalence mis-specification. It can be extended to multiple study designs and can be applied to multiple regressions, including regressions with no parametric likelihood function, such as quantile regression.

6.2 Future extension

We would like to extend the proposed methods to accommodate a wide range of case-control studies that are biased in secondary analysis. For example, we consider to joint analyze multiple case-control studies at the same time to improve power. We also consider the nested/matched case-control design that aims at improving the efficiency of traditional case-control design. Finally, we consider expanding the secondary analysis from GWAS to sequencing studies that focus on the rare variants. In this section, we will explain the issues we might encounter in these directions, and our preliminary ideas of secondary analysis for the further study.

6.2.1 Secondary analysis in multiple genetic case-control studies

One direction to expand the new estimating equations is to joint analyze of the same secondary trait from multiple case-control studies. As most of the case-control studies are powered for the primary analysis, we might have difficulties to detect the important SNPs for the secondary traits using one case-control sample. Meanwhile, a lot of secondary

traits, such as height, weight, blood pressure and common diseases, are widely asked in the case-control studies. It is natural for the researchers to consider using the data from multiple case-control studies to investigate the associations between the genetic markers and the same secondary trait of interest. In the example from the introduction [Lettre et al., 2008], researchers analyzed the secondary trait height using six GWAS focusing on diabetes, cardiovascular diseases and cancers.

When we consider joint analysis of multiple case-control studies, there are a number of issues we need to be aware of. First, different GWAS might use different genotyping platforms, and their SNPs may not overlap each other. Therefore, imputation of the 'hidden' variants is needed to combine the studies. Imputation is the process of predicting a set of genotypes that are not directly assayed in a sample of individuals. Imputation methods can infer the alleles of 'hidden' variants and use those inferences to test the hidden variants for association. As imputed SNPs can lead to false positives if they are poorly performed, or even well performed, the major discovery based on imputed SNPs should be verified, probably by conducting new studies.

Second, by combining the multiple case-control studies, researchers might accidentally mix the subjects from different subgroups. If genotype frequencies differ between these subgroups and sampling favors certain subgroups over others, the sample estimate may be biased. Even there is no bias, variance of the estimate can be affected, and can affect the validity of the association test results. Therefore, the adjustment for population stratification is particularly important in this scenario when we combine multiple samples to correct the bias and variance. There are a number of approaches to adjust for population stratification, including genomic control [Devlin and Roeder, 1999], STRUCTURE [Pritchard et al., 2000], Eigenstrat [Price et al., 2006b] and EMMAX [Kang et al., 2010], and among them Eigenstrat is the most popular approach nowadays in the literature.

After dealing with these issues, one could expand the estimating equations to incorporate multiple diseases, and conduct the secondary analysis in multiple genetic case-control studies. For the case-control studies when no overlap in participants, we could simply write the new estimating equations conditional on the primary disease status separately for each

study, and take the summation their equations. Therefore, for SICO approach, one can simulate the pseudo observations $\tilde{y}_i$ and estimate $\widehat{P}(d_i|x_i, \mathbf{z}_i)$ for each sample and conduct the weighted regressions on the combined multiple original samples and counter-factual pseudo samples. For CE approach, one can calculate the conditional expectation of the $S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$ over $\tilde{y}_i$ and estimate $\widehat{P}(d_i|x_i, \mathbf{z}_i)$ separately for each sample, and conduct weighted regressions using combined multiple samples by taking the weights from each individual studies. For multiple case-control studies have overlapped observations, and the details of the joint analysis, we refer to future research to handle them.

6.2.2 Secondary analysis in nested/matched case-control designs

Another direction of extending the estimating equations is to adapt more complex designs. A lot of GWAS use nested/matched case-control designs to improve the efficiency. A nested/matched case-control study is a special type of a case-control study in which only a subset of controls from the cohort are compared to the cases. Unlike in traditional case-cohort study that cases are compared to a random subset of controls, in a nested/matched case-control study, some number of controls are selected for each case from that case's matched risk set. By matching on factors such as age and selecting controls from relevant risk sets, the nested case control model is generally more efficient than a case-cohort design with the same number of selected controls.

To conduct the secondary analysis for a nested case-control sample, one must take into account the way in which controls are sampled from the cohort. It is common that researchers treat the cases and selected controls as the original cohort and performing a logistic regression. This can result in biased estimates as the controls are not a representative sample of the general population. We could possibly expand our proposed approach to account for the missing covariates among those who are not selected into the study from the population. This would improve the estimation of $\widehat{P}(d_i|x_i, \mathbf{z}_i)$, and results in an unbiased estimates for the general population. Further research needs to be carried out in this direction for the nested/matched case-control studies.

6.2.3 Secondary analysis in sequencing studies

The availability of high-throughput sequencing enables rapid sequencing of large stretches of DNA base pairs spanning entire genomes, and produces new statistical questions to analyze the association between the secondary trait and rare variants. While the proposed approaches provide a general framework for case-control studies that are not limited to GWAS, they could not be simply applied sequencing studies. One major reason is that the rare variants have very low frequency (MAF$<$ 1%), so there are few occurrences of the variants in dataset of reasonable sample size. Second, the rare variants are numerous that over 95% of variants in a region have MAF$<$ 1%, and this increased penalty for multiple testing. Finally, the effect sizes of the rare variants are not expected to be very large with expected odds ratio roughly between $4-5$. As a result, standard association tests and regressions used in GWAS have very low power to detect the rare variants.

Therefore, when conducing the secondary analysis for sequencing studies, we need first consider the grouping strategies to handle the rare variants and then apply the proposed approaches to the treated genetic information for estimation. For example, we could incorporate region-based analysis, such as burden tests and sequence kernel association test [Wu et al., 2011], to aggregate the rare variants within a region, and then apply our estimating equations to the aggregated data to detect the effects in regions. We refer to future research for handling the sequencing data.

In summary, while proposed for GWA case-control studies, the new estimating equation based approach can be extended in a number of directions to accommodate the secondary analysis for many types of genetic studies. Further study is needed to realize these ideas.

Bibliography

Burrows, B., Martinez, F. D., Halonen, M., Barbee, R. A., and Cline, M. G. (1989). Association of asthma with serum ige levels and skin-test reactivity to allergens. *New England Journal of Medicine*, 320(5):271–277. PMID: 2911321.

Cornelis, M. C., Agrawal, A., Cole, J. W., Hansel, N. N., Barnes, K. C., Beaty, T. H., Bennett, S. N., Bierut, L. J., Boerwinkle, E., Doheny, K. F., Feenstra, B., Feingold, E., Fornage, M., Haiman, C. A., Harris, E. L., Hayes, M. G., Heit, J. A., Hu, F. B., Kang, J. H., Laurie, C. C., Ling, H., Manolio, T. A., Marazita, M. L., Mathias, R. A., Mirel, D. B., Paschall, J., Pasquale, L. R., Pugh, E. W., Rice, J. P., Udren, J., van Dam, R. M., Wang, X., Wiggs, J. L., Williams, K., and Yu, K. (2010). The gene, environment association studies consortium (geneva): maximizing the knowledge obtained from gwas by collaboration across studies of multiple conditions. *Genetic Epidemiology*, 34(4):364–372.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004.

Edwards, A. O., Ritter, R., Abel, K. J., Manning, A., Panhuysen, C., and Farrer, L. A. (2005). Complement factor h polymorphism and age-related macular degeneration. *Science*, 308(5720):421–424.

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R., Elliott, K. S., Lango, H., Rayner, N. W., et al. (2007). A common variant in the fto gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826):889–894.

Guo, S. and Fraser, M. W. (2010). Propensity score analysis. *Statistical methods and applications*.

Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., Spencer, K. L., Kwan, S. Y., Noureddine, M., Gilbert, J. R., et al. (2005). Complement factor h variant increases the risk of age-related macular degeneration. *Science*, 308(5720):419–421.

He, J., Li, H., Edmondson, A. C., Rader, D. J., and Li, M. (2011). A gaussian copula approach for the analysis of secondary phenotypes in case–control genetic association studies. *Biostatistics*, page kxr025.

Hjartåker, A., Langseth, H., and Weiderpass, E. (2008). Obesity and diabetes epidemics. In *Innovative Endocrinology of Cancer*, pages 72–93. Springer.

Jiang, Y., Scott, A., and Wild, C. (2006). Secondary analysis of case-control data. *Statistics in medicine*, 25(8):1323–1339.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., Eskin, E., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., et al. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Kraft, P. (2007). Analyses of genome-wide association scans for additional outcomes. *Epidemiology*, 18(6):838.

Lee, A., McMURCHY, L., and Scott, A. (1997). Re-using data from case-control studies. *Statistics in medicine*, 16(12):1377–1389.

Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., Eyheramendy, S., Voight, B. F., Butler, J. L., Guiducci, C., et al. (2008). Identification of ten

loci associated with height highlights new biological pathways in human growth. *Nature genetics*, 40(5):584–591.

Li, H. and Gail, M. (2012). Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies. *Human Heredity*, 73(3):159–173.

Lin, D. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic epidemiology*, 33(3):256–265.

Liu, M., Rogers, L., Cheng, Q., Shao, Y., Fernandez-Beros, M. E., Hirschhorn, J. N., Lyon, H. N., Gajdos, Z. K., Vedantam, S., Gregersen, P., et al. (2011). Genetic variants of tslp and asthma in an admixed urban population. *PloS one*, 6(9):e25099.

Mendel, G. (1865). Experiments in plant hybridization. *Verhandlungen des naturforschenden Vereins Brünn.) Available online: www. mendelweb. org/Mendel. html (accessed on 1 January 2013)*.

Monsees, G., Tamimi, R., and Kraft, P. (2009). Genome-wide association scans for secondary traits using case-control samples. *Genetic epidemiology*, 33(8):717–728.

Morjaria, J. B. and Polosa, R. (2009). Off-label use of omalizumab in non-asthma conditions: new opportunities. *Expert Review of Respiratory Medicine*, 3(3):299–308. PMID: 20477322.

Nagelkerke, N. J., Moses, S., Plummer, F. A., Brunham, R. C., and Fish, D. (1995). Logistic regression in case-control studies: The effect of using independent as dependent variables. *Statistics in medicine*, 14(8):769–775.

Pakistan Stroke Society (2006). Information about stroke.

Peters, S. A. E., Huxley, R. R., and Woodward, M. (2014). Diabetes as a risk factor for stroke in women compared with men: a systematic review and meta-analysis of 64 cohorts, including 775âĂŁ385 individuals and 12âĂŁ539 strokes. *The Lancet*, (0):–.

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

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006a). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904 – 909.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006b). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Regan, E. A., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. A., Beaty, T. H., Curran-Everett, D., Silverman, E. K., and Crapo, J. D. (2010). Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43. PMID: 20214461.

Richardson, D., Rzehak, P., Klenk, J., and Weiland, S. (2007). Analyses of case-control data for additional outcomes. *Epidemiology*, 18(4):441–445.

Risch, N., Merikangas, K., et al. (1996). The future of genetic studies of complex human diseases. *Science-AAAS-Weekly Paper Edition*, 273(5281):1516–1517.

Salonen, J. T., Uimari, P., Aalto, J.-M., Pirskanen, M., Kaikkonen, J., Todorova, B., HyppÃűnen, J., Korhonen, V.-P., Asikainen, J., Devine, C., Tuomainen, T.-P., Luedemann, J., Nauck, M., Kerner, W., Stephens, R. H., New, J. P., Ollier, W. E., Gibson, J. M., Payton, A., Horan, M. A., Pendleton, N., Mahoney, W., Meyre, D., Delplanque, J., Froguel, P., Luzzatto, O., Yakir, B., and Darvasi, A. (2007). Type 2 diabetes whole-genome association study in four populations: The diagen consortium. *The American Journal of Human Genetics*, 81(2):338 – 345.

Scott, A. and Wild, C. (2001). Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, 96(1):3–27.

Scott, A. and Wild, C. (2002). On the robustness of weighted methods for fitting models to case–control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):207–219.

Sundquist, K. and Li, X. (2006). Type 1 diabetes as a risk factor for stroke in men and women aged 15âĂŞ49: a nationwide study from sweden. *Diabetic Medicine*, 23(11):1261–1267.

Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S. F., Payne, F., Lowe, C. E., Szeszko, J. S., Hafler, J. P., Zeitels, L., Yang, J. H. M., Vella, A., Nutland, S., Stevens, H. E., Coleman, H. S. G., Maisuria, M., Meadows, W., Smink, L. J., Healy, B., Burren, O. S., Lam, A. A. C., Ovington, N. R., Allen, J., Adlem, E., Leung, H.-T., Wallace, C., Howson, J. M. M., Guja, C., Ionescu-Tirgoviste, C., GET1FIN, Simmonds, M. J., Heward, J. M., Gough, S. C., Consortium, T. W. T. C. C., Dunger, D. B., Wicker, L. S., , and Clayton, D. G. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics*, 39:857 – 864.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of {GWAS} discovery. *The American Journal of Human Genetics*, 90(1):7 – 24.

Wang, J. and Shete, S. (2011). Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genetic epidemiology*, 35(3):190–200.

Wang, J. and Shete, S. (2012). Analysis of secondary phenotype involving the interactive effect of the secondary phenotype and genetic variants on the primary disease. *Annals of Human Genetics*.

Wei, Y., Pere, A., Koenker, R., and He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in medicine*, 25(8):1369–1382.

Wei, Y., Song, X., Liu, M., Ionita-Laza, I., and Reibman, J. (2015). Quantile regression in the secondary analysis of case-control data. *Journal of the American Statistical Association*, 0(ja):00–00.

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. *The American Journal of Human Genetics*, 89(1):82–93.

Yang, J., Loos, R. J., Powell, J. E., Medland, S. E., Speliotes, E. K., Chasman, D. I., Rose, L. M., Thorleifsson, G., Steinthorsdottir, V., Mägi, R., et al. (2012). Fto genotype is associated with phenotypic variability of body mass index. *Nature*, 490(7419):267–272.