Analysis of Mixed Types of Traits in Genetic Association Studies and Application to Genome-Wide Association Studies

By Ruihua Xu

B.A. in Business English, August, 2001, Sountheast University, Nanjing, China
M.S. in Statistics, May, 2008, George Mason University

A Dissertation submitted to

The Faculty of
Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

January 31, 2014

Dissertation directed by

Gang Zheng
Mathematical Statistician
National Heart, Lung and Blood Institute
National Institutes of Health

YingLei Lai
Associate Professor of Statistics

The Columbian College of Arts and Sciences of The George Washington University certifies that Ruihua Xu has passed the Final Examination for the degree of Doctor of Philosophy on Oct,18, 2013. This is the final and approved form of the dissertation.

Analysis of Mixed Types of Traits in Genetic Association Studies and Application to Genome-Wide Association Studies

Ruihua Xu

Dissertation Research Committee:

Gang Zheng, Mathematical Statistician, National Heart, Lung and Blood Institute, National Institutes of Health, Dissertation Co-Director

YingLei Lai, Associate Professor of Statistics, Dissertation Co-Director

Srinivasan Balaji, Assistant Professor of Statistics, Committee Member

Angelo Elmi, Assistant Professor of Epidemiology and Biostatistics, Committee Member

Acknowledgments

I would like to express my deepest gratitude to my thesis advisor, Dr. Gang Zheng, for his patient advising and insightful guidance throughout this research.

I would like to extend my sincere appreciation to my thesis co-advisor, Dr. YingLei Lai, for his kind advice and generous support.

I would like to thank my committee members and chair, Dr. Srinivasan Balaji , Dr. Angelo Elmi, Dr. Kai Yu, Dr. Qin Pan, Dr. Zhaohai Li for their precious time and great support. Also, many thanks to all the faculty members of the Department of Statistics and the Department of Epidemiology and Biostatistics. Special thanks to Dr. Dante Verme and Dr. Efstathia Bura, for their kind advice in selecting my course work in my Ph.D. study.

I would like to thank Dr. Colin Wu and Dr. Neal Jeffries of NIH, for their valuable comments and suggestions for this research.

Finally, I am grateful to my family and my friends for always believing in me, encouraging me to finish this work soon.

Abstract

Analysis of Mixed Types of Traits in Genetic Association Studies and Applications to Genome-Wide Association Studies

Genetic association tests aim to test if a genetic marker is associated with a trait. There are two common types of traits, one is binary (case control) trait, the other is quantitative (continuous). For a genetic outcome-dependent sampling, the binary trait and the quantitative trait are correlated. Since these two types of traits are dependent,so it is more powerful to test a joint association of the genetic marker with the both traits than to just test each single trait separately(Klei et al., 2008; Zhu et al., 2009). In this research, we proposed a simple approach to combine these two mixed types of traits into a single ordinal outcome, then apply a proportional odds model to identify the association between a genetic marker and the mixed types of traits. Performance of these mothods for analysis of mixed types of traits will be evaluated in comparison with existing methods using simulation studies. Application to a genome-wide association study of rheumatoid arthritis with outcome-dependent sampling will be presented.

Contents

List of Figures

List of Tables

Chapter 1

Introduction of Genetic Association Studies for Mixed Types of Traits

Genetic association tests aim to test if a genetic marker is associated with a trait. Usually, an biallelic marker is considered, e.g. a single nucleotide polymorphism (SNP).

There are two common types of traits, one is binary (case control) trait, such as diease or no disease, the other is quantitative (continuous), such as height or weight.

Generally, if the SNP is not associated with both binary trait and quantitative trait, then usually we will test the relationship for the SNP and binary trait or the quantitative trait seperately. For example, for binary trait, Cochran-Armitage Trend Test(ACTT)or pearson's chi-square test are used to test if there is an assoication

between a binary trait and SNP. For quantitative trait, F-test obtained from the linear regression model is used to test if there is an association between quantitative trait and SNP.

However, in a genetic outcome-dependent sampling, the binary trait and the quantitative trait are correlated. When a SNP is associated with multiple traits individually, an analysis of this SNP with these traits simultaneously (pleiotropic association) is more powerful than testing association with a single trait (Klei et al., 2008; Zhu et al., 2009).

There are already several methods for testing pleiotropic association.Xing and Xing (2009) combined the p-values of the above two tests using Fisher's combination as a test for pleiotropic association. Zheng et al. (2012a) replaced the quantitative trait (anti-CCP) of the controls by the minimum observed quantitative trait (anti-CCP) of the cases and conducted a quantitative trait analysis as if the imputed values of the controls were observed. They also combined the p-values from case-control binary trait study and quantitative trait analysis using Fisher's combination method. Due to the dependence of the two p-values, they applied the combined test with a Gamma distribution $\Gamma(1.3, 3.1)$ under the null hypothesis that there is no association between a SNP and either trait.

Motivated from a genome-wide association study (GWAS) of Genetic Analysis Workshop 16 (GAW16), a simple approach is proposed to combine binary and quantitative traits into a single ordinal outcome. Then a usual proportional odds model can be applied, which follows asymptotically a chi-squared distribution under the null hypothesis. The relation of the proposed test with the tests for the binary and quantitative traits is studied. Simulation results for normal and non-normal traits are

presented for the proposed method and some existing tests. The results are applied to GAW16 to identify genes pleiotropy for RA(rheumatoid arthritis) and anti-CCP (anti-cyclic citrullinated peptide), which is not available for controls.

This thesis is organized as follows:

Chapter 1 provides an introduction to the study topics.

Chapter 2 provides a detailed review about the background in genetic association studies and tests and some basic concepts in Genetics.

Chapter 3 provides a example for outcome-dependent sampling, a dataset of a genome-wide association study of rheumatoid arthritis with anti-cyclic citrullinated peptide, which is only available for rheumatoid arthritis patients. Furthermore, we will explore the shortcoming of the existing approaches of joint analysis for mixed types of traits.

Chapter 4 proposes our approach using proportional odds model.

Chapter 5 Simulation results for normal and non-normal traits are presented for the proposed method and some existing tests. The results are applied to GAW16 to identify genes pleiotropy for RA(rheumatoid arthritis) and anti-CCP (anti-cyclic citrullinated peptide), which is not available for controls.

Chapter 6 discusses future research.

Chapter 2

Background

2.1 Some Basic Concepts

Genetics is the study of biologic inheritance. This section will briefly introduce some basic concepts in Genetics including genes, alleles, marker, genotypes, phnotypes, Minor Allele Frequency.

A human body is made up of trillions of cells. Each cell nucleus contains an identical complement of chromosomes. Generally, human being have 23 pairs of chromosomes. Each chromosome is one long DNA molecule, and genes are functional regions of this DNA. DNA (Deoxyribonucleic Acid)is a double helix.

Genes, the basic units of inheritance, are composed of DNA and are located on chromosomes.

Each gene occupies a position along a chromosome known as a **locus**.

The genes at a particular locus can have different forms (i.e, they can be composed

of different nucleotide squence) called **alleles**.

Each person has two alleles A and B at a locus on the autosomes and the pair of the two alleles forms a **genotype** at that locus. There are four genotypes: AA, AB, BA and BB, AB and BA are not distinguished.

Markers are the special loci with the known location and the known number and types of alleles on the chromosomes. Since their locations are known, they are commonly employed as landmarks to locate disease/function loci although they may not have any function by themselves.

Phenotype is defined as any observable characteristic or trait of an individual. The trait can be continous, e.g. blood pressure, weight, height ect., descrete, including binary such as case and control, or ordinal categories related to different stages of a disease.

Minor Allele Frequency (MAF) refers to the frequency of the allele with frequency no more than 0.5 in a population. Denote the three genotypes by $G_0 = AA$, $G_1 = AB$ and $G_2 = BB$. The genotype frequencies are denoted as $g_i = Pr(G_i)$ for $i = 0, 1, 2$ and $g_0 + g_1 + g_2 = 1$

Sensitivity and Specificity Sensitivity and specificity are statistical measures of the performance of a binary classification test. Sensitivity measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition). Specificity measures the proportion of negatives which are correctly identified (e.g. the percentage of healthy people who are correctly identified as not having the condition).

Sensitivity relates to the test's ability to identify positive results. Again, consider the example of the medical test used to identify a disease. The sensitivity of a test is

Table 2.1: Example for Sensitivity and Specificity

	Disease	No Disease	Total
Screening Test Is Positive	a	b	a+b
Screening Test Is Negative	c	d	c+d
Total	a+c	b+d	a+b+c+d

the proportion of people who have the disease who test positive for it. This can also be written as:

Sensitivity $=\frac{a}{a+c}$=those who test positive/ all with disease

Specificity relates to the ability of the test to identify negative results. Consider the example of the medical test used to identify a disease. The specificity of a test is defined as the proportion of patients who do not have the disease who will test negative for it. This can also be written as:

Specificity $=\frac{b}{b+d}$=those who test negative/ all without disease

In GAW16 data, anti-CCP is a quantitative biomarker correlated with RA. It has relatively low sensitivity (about 75%) but high specificity (about 90%). So negative anti-CCP value, e.g. ≤ 20, could highly predict the negative RA because of high specificity.

2.2 Genetic Association Studies and Tests

There are two genetic studies, one is population-based studies, the other is family-based studies. In this research, we will focus on population based studies.

Genetic association tests aim to test if a genetic marker is associated with a trait. There are two common types of traits, one is binary (case control) trait, such as diease or no disease, the other is quantitative (continuous), such as height or weight.

For binary trait, generally, Cochran-Armitage Trend Test(CATT)or pearson's chi-square test are used to test if there is an assoication between a binary trait and a genetic marker. e.g. SNPs. Both of these two tests are genotype-based tests.

The Cochran-Armitage Trend Test(CATT) has an assumption that one of the alleles is the risk allele and that the risk of developing the disease increases with the number of the risk alleles in the genotype. Different trend tests are used for the four different genetic models. The trend test may not be be robust if the underlying genetic model is misspecified.

Pearson's test compares the observed genotype counts in case (r_0, r_1, r_2) and controls (s_0, s_1, s_2) with their expected values under the null hypothesis that the disease is not associated with the genotypes. Pearson's test is robust because it does not assume any genetic model. Both these two tests follow chi-squared distribution asymtotically.

For quantitative trait, F-test obtained from the linear regression model is used to test if there is an association between quantitative trait and a genetic marker.

However, in a genetic outcome-dependent sampling, the binary trait and the quantitative trait are correlated. When a SNP is associated with multiple traits individually, an analysis of this SNP with these traits simultaneously (pleiotropic association) is more powerful than testing association with a single trait (Klei et al., 2008; Zhu et al., 2009).

Xing and Xing (2009) combined the p-values from case-control binary trait study and quantitative trait analysis using Fisher's combination as a test for pleiotropic association. Zheng et al. (2012a) replaced the missing quantitative trait (anti-CCP) of the controls by the minimum observed quantitative trait (anti-CCP) of the cases

and conducted a quantitative trait analysis as if the imputed values of the controls were observed. They also combined the p-values of the two analyses using Fisher's combination method. Due to the dependence of the two p-values, they applied the combined test with a Gamma distribution $\Gamma(1.3, 3.1)$ under the null hypothesis that there is no association between a SNP and either trait.

2.3 Linkage Disequilibrium(LD) and Association

In population genetics, linkage disequilibrium is the non-random association of alleles at two or more loci. Generally, an individual's gene will be transmitted to the next generation independently of one another. However, because genes located close together on the same chromosome do tend to be transmitted together to the offsprings. If two loci tend to be inherited together, they are said to be linked. Linkage measures the deviation of the independent transmission of a locus and a phenotype or another locus within a family.

Association, on the other hand, measures such deviation in a population level and testing for association is to test if Linkage disequilibrium (LD) equals to 0. We focus on the association study through this research.

2.4 Hardy-Weinberg Principle

The Hardy-Weinberg principle, also known as the Hardy-Weinbery law or Hardy Weinberg Equilibrium (HWE), is a well known model in population genetics. It states that under random mating both allele and genotype frquency in a population remain constant or stable if no distrubing factors are introduced.

Consider a diallelic locus with alleles A and B with population frequencies q and p, respectively. Assume males and females have the same allele frquencies. Then, under random mating, the genotype frequencies $g_0 = Pr(AA)$, $g_1 = Pr(AB)$ and $g_2 = Pr(BB)$. We assume the HWE holds through this thesis.

2.5 Genetic Models

The genetic model is the relationships among the penetrances of the disease under consideration. Four common genetic models are discussed in the literature: the recessive (REC), additive (ADD), multilicative (MUL) and dominant (DOM) models. A genetic model is called REC, ADD, MUL and DOM if

$$f_1 = f_0,$$
$$f_1 = (f_0 + f_2)/2,$$
$$f_1 = \sqrt{f_0 f_2},$$
$$f_1 = f_2,$$

respectively. Here, $f_j = Pr(case|G_j)$ for j=0,1,2 is defined as the penetrance of a disease given a genotype. In this thesis, we focus on ADD (additive) Model and REC (recessive) Model.

Chapter 3

Motivation Example and Existing Problems of Current Approch

3.1 Motivation Example

Data sharing becomes common in genetic association studies, and the outcome-dependent sampling is the consequence of data sharing, under which a phenotype of interest is not measured for some subgroup. For example, when the data from the database of Genotypes and Phenotypes (dbGap)(http://www.ncbi.nlm.nih.gov/gap)are used, a specific quantitative trait of interest in one study may not be measured in another one. Data sharing is common and cost effective as data from all recent genetic association studies funded by National Institutes of Health (NIH) are available at dbGap for sharing. Genetic Analysis Workshop 16 (GAW16) is another example for data sharing (Zheng et al., 2012).

GAW16 contains about 500,000 SNPs and 868 cases(RA). There are one binary trait (Rheumatoid Arthritis) and two continuous traits (anti-CCP and IgM) in the rheumatoid arthritis dataset of Problem 1 in the Genetic Analysis Workshop 16 (GAW16): RA (rheumatoid arthritis) status, anti-CCP (anti-cyclic citrullinated peptide) and IgM. All cases in GAW16, from North American Rheumatoid Arthritis Consortium (Amos et al.,2009), have quantitative trait(anti-CCP) measurements which are greater than 20 However, 1194 controls, from New York Cancer Project included in GAW16, have no quantitative trait(anti-CCP) measurements. anti-CCP is a quantitative biomarker correlated with RA (Huizinga et al.,2005). It has relatively low sensitivity (about 75%) but high specificity (about 90%)(Coenen et al.,2007). The missing anti-CCP among controls in GAW16 is not only due to outcome-dependent sampling, but also due to the fact that the measurement is costly and that anti-CCP is often measured to confirm RA ((Rheumatoid Arthritis)(Zheng et al., 2012). So for GAW16 data, case-control association between RA and SNPs can be tested using all GAW16 data. However, association between anti-CCP and SNPs can be tested using only cases.

The features of GAW16 data motivate this study, since GAW16 data includes the missing data due to the special study design with data sharing or outcome-dependence (only controls having missing data), the quantitative trait with natural grouping boundaries, and the binary and quantitative traits are both highly correlated with the SNPs.

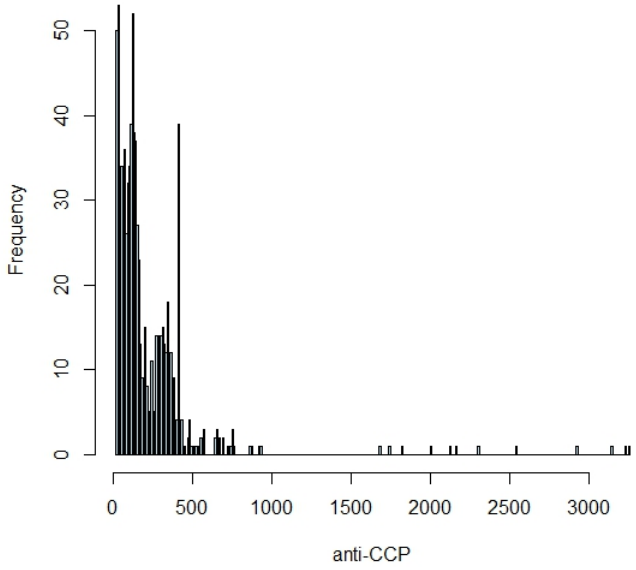
3.2 Existing Problem in Current Approach

There are some existing approachs for the analysis for the mixed types of traits. For example, Xing and Xing (2009) combined the p-values from case-control binary trait study and quantitative trait analysis using Fisher's combination as a test for pleiotropic association. Zheng et al. (2012a) replaced the missing quantitative trait (anti-CCP) of the controls by the minimum observed quantitative trait (anti-CCP) of the cases and conducted a quantitative trait analysis as if the imputed values of the controls were observed. They also combined the p-values of the two analyses using Fisher's combination method. Due to the dependence of the two p-values, they applied the combined test with a Gamma distribution $\Gamma(1.3, 3.1)$ under the null hypothesis that there is no association between a SNP and either trait.

However, the distribution of anti-CCP does not follow a normal distribution (p-value < 0.0001 using Shapiro-Wilk test).Even the log transformation or square-root applied, both p-values are still < 0.0001 using the same goodness-of-fit test). But the assumption of F test from linear regression is the outcome trait should follow normal distribution. So F-test may not be a good test to anaylze anti-CCP. So rank-based nonparametric methods are proposed to model association of anti-CCP instead of using the F-test (Li et al.,2013).

Generally, anti-CCP is regarded as positive if it is ≥ 20 and negative otherwise. It is further classified as low to weak positive if it is between 20 - 39, moderate positive if between 40 - 59, and high to strong positive if > 60 (http://en.wikipedia.org/wiki/Anti-citrullinated_protein_antibody). A large value of anti-CCP may not be as reliable as a small one. In some cases, an anti-CCP value greater than 250 would be only reported

Figure 3.1: Histogram of anti-CCP among the cases in GAW16



as > 250 and treated as 251 in the analysis(Mjaavatten et al.,2010). In GAW16, the anti-CCP measures for the cases range from 20 to over 3,000, although most of the observations are below 500. Hence, it is highly likely there are measurement errors for large anti-CCP values.

Figure 1 is the histogram of the anti-CCP from GAW16 data. The distribution of anti-CCP does not follow a normal distribution (p-value < 0.0001 using Shapiro-Wilk test).Even the log transformation or square-root applied, both p-values are still < 0.0001 using the same goodness-of-fit test). Rank-based nonparametric methods are proposed to model association of anti-CCP instead of using the F-test (Li et al.,2013). However, those approaches are only based on the genotypes and anti-CCP of RA cases without using the control data.

Chapter 4

Proposed Approach Using Proportional Odds Model

4.1 Data

Suppose the marker we are interested in has two alleles A and B with genotypes denoted as $(G_0, G_1, G_2) = (AA, AB, BB)$. For the jth individual $(j = 1, \ldots, n)$,

Table 4.1: Ordinal genetic data for mixed types of traits in genetic association studies $(C \geq 3)$.

	Ordinal outcome				
Genotype	1	2	$\cdots$	C	Total
$G_0(= AA)$	r_{01}	r_{02}	$\cdots$	r_{0C}	$r_{0\cdot}$
$G_1(= AB)$	r_{11}	r_{12}	$\cdots$	r_{1C}	$r_{1\cdot}$
$G_2(= BB)$	r_{21}	r_{22}	$\cdots$	r_{2C}	$r_{2\cdot}$
Total	$r_{\cdot 1}$	$r_{\cdot 2}$	$\cdots$	$r_{\cdot C}$	$r_{\cdot\cdot} = n$

data consist of (d_j, g_j, y_j), where d_j is the indicator for case ($d_j = 1$) or control ($d_j = 0$), g_j is the genotype, and y_j is the quantitative trait, only observed when $d_j = 1$ ($j = 1, \ldots, n$).

Then an ordinal outcome was defined as Y_d, which combines (d_j, y_j) into a single outcome. Let $Y_d = 1$ if an individual is a control and $Y_d = c$ if an individual is a case. The data is displayed in Table 4.1, where r_{ic} is the number of individuals with genotype G_i and ordinal outcome c ($i = 0, 1, 2$ and $c = 1, \ldots, C$). The data contains the case-control data, where the case and control genotype counts are $(\sum_{c=2}^{C} r_{0c}, \sum_{c=2}^{C} r_{1c}, \sum_{c=2}^{C} r_{2c})$ and (r_{01}, r_{11}, r_{21}), respectively.

4.2 Genetics Association Tests from Single Trait and Mixed Types of Traits

The genetic effect for (G_0, G_1, G_2) can be coded as a one-dimensional additive effect $(x_0, x_1, x_2) = (0, 1, 2)$, or a two-dimensional effect $(x_0, x_1, x_2) = ((0, 0), (0, 1), (1, 1))$ as well.

For the association of case-control binary trait, two score tests can be obtained by the logistic regression model with two different genetic codings. The first one with the additive effect is the trend test denoted as T_1. The other one with the two-dimensional effect is Pearson's chi-squared test denoted as T_2. Under the null hypothesis H_0, $T_k \sim \chi^2_k$ asymptotically ($k = 1, 2$), where χ^2_k is the chi-squared distribution with k degrees of freedom.

For the association with the quantitative trait, an F-test can be obtained by the

linear regression model with two different genetic codingss for the genetic effect. Here, denote the F-test as F_1 or F_2, which follow approximately χ^2_1 and χ^2_2, respectively, under H_0 when n is large enough. These F-tests can only be conducted on the quantitative traits of cases. Zheng et al. (2012a) proposed to replace the unobserved quantitative traits of controls by the minimum observed trait value of cases as if they were observed for all controls and apply the usual F-test. Denote the corresponding F-tests as F_1^* and F_2^*. Under H_0, $F_k^* \sim \chi^2_k$ asymptotically.

Single-trait associations was detected by the above tests. For the association with both binary and quantitative traits, Fisher's combination of p-values of T_1 or T_2 and F_1 or F_2 (F_1^* or F_2^*) can be used. Denote the p-values as p_{T_k}, p_{F_k} and $p^*_{F_k}$.

Denote

$$\begin{aligned} J_{uv} &= -2\log(p_{T_u}) - 2\log(p_{F_v}), \\ J^*_{uv} &= -2\log(p_{T_u}) - 2\log(p^*_{F_v}), \end{aligned}$$

for $u, v = 1, 2$. Xing and Xing (2009) considered J_{11} while Zheng et al. (2012a) considered J^*_{21}. Because controls are not used in F_1 or F_2, $J_{uv} \sim \chi^2_4$ $(u, v = 1, 2)$ under H_0 of no association with either trait. However, J^*_{uv} is no longer χ^2_4 under H_0. Zheng et al. (2012a) proposed $J^*_{21} \sim \Gamma(1.3, 3.1)$, a Gamma distribution whose shape parameter (1.3) and scale parameter (3.1) were estimated by matching the first two moments of the Gamma distribution and those estimated from the simulated data under H_0. However, it is unknown if $\Gamma(1.3, 3.1)$ can be still used if other choices of u and v are used. Furthermore, the performance of using $\Gamma(1.3, 3.1)$ is unknown when the trait is not normally distributed even when $(u, v) = (2, 1)$.

4.3 The Proportional Odds Model

For the data displayed in table 4.1, a usual proportional odds model (POM) can be applied to test mixted types of traits association:

$$\text{logit } \Pr(Y_d \leq c|G_i) = \alpha_c + \beta x_i,\ c = 1, \ldots, C-1, i = 0, 1, 2. \tag{4.1}$$

The genetic effect is coded by $(x_0, x_1, x_2) = (0, 1, 2)$. Under H_0 of no association with either trait, $\beta = 0$.

Let $y_{jc} = 1$ if an individual j has cth ordinal outcome and 0 otherwise, $j = 1, \ldots, n$ and $c = 1, \ldots, C$. Denote $\alpha_0 = -\infty$, $\alpha_C = +\infty$, $\Delta_1(x) = \exp(x)/\{1 + \exp(x)\}$ and $\Delta_2(x) = \exp(x)/\{1 + \exp(x)\}^2$.

Then the likelihood function for the data in Table 4.1 is given by

$$\begin{aligned}
L(\alpha_1, \ldots, \alpha_C, \beta) &= \prod_{j=1}^{n} \left[\prod_{c=1}^{C} \{\Pr(Y_d = c|g_j)\}^{Y_{jc}} \right] \\
&= \prod_{j=1}^{n} \left[\prod_{c=1}^{C} \{\Pr(Y_d \leq c|g_j) - \Pr(Y_d \leq c-1|g_j)\}^{Y_{jc}} \right] \\
&= \prod_{i=0}^{2} \prod_{c=1}^{C} [\exp(\alpha_c + \beta x_i)/\{1 + \exp(\alpha_c + \beta x_i)\} \\
&\quad - \exp(\alpha_{c-1} + \beta x_i)/\{1 + \exp(\alpha_{c-1} + \beta x_i)\}]^{r_{ic}} \\
&= \prod_{i=0}^{2} \prod_{c=1}^{C} \{\Delta_1(\alpha_c + \beta x_i) - \Delta_1(\alpha_{c-1} + \beta x_i)\}^{r_{ic}},
\end{aligned}$$

where g_j repensents for G_0, G_1 or G_2.

Then the log-likelihood function is as follows:

$$l = l(\alpha_1, \ldots, \alpha_C, \beta) = \sum_{i=0}^{2} \sum_{c=1}^{C} r_{ic} \log(\Delta_1(\alpha_c + \beta x_i) - \Delta_1(\alpha_{c-1} + \beta x_i)).$$

$$= \sum_{i=0}^{2} \sum_{c=1}^{C} r_{ic} \log(\Delta_{1ci} - \Delta_{1(c-1)i})$$

with $\Delta_{10i} = 0$ and $\Delta_{1Ci} = 1$ for all $i = 0, 1, 2$.

Firstly, we take derivative for α_C, for $c = 1, \ldots, C-1$, then we have $\partial l/\partial \alpha_c = \sum_{i=0}^{2} \{r_{ic}\Delta_{2ci}/(\Delta_{1ci} - \Delta_{1(c-1)i}) - r_{i(c+1)}\Delta_{2ci}/(\Delta_{1(c+1)i} - \Delta_{1ci})\}$, where $\Delta_{1c} = \exp(\alpha_c + \beta x_i)/\{1 + \exp(\alpha_c + \beta x_i)\}$ and $\Delta_{2c} = \exp(\alpha_c + \beta x_i)/\{1 + \exp(\alpha_c + \beta x_i)\}^2$

Then let $(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{C-1})$ satisfy $\partial l/\partial \alpha_c|_{H_0:\beta=0} = 0, for c = 1, \ldots, C-1$. Then we have

$$\left(\Delta_{1(c+1)} - \Delta_{1c}\right) = \frac{r_{\cdot(c+1)}}{r_{\cdot c}} \left(\Delta_{1c} - \Delta_{1(c-1)}\right), \quad c = 1, \ldots, C-1,$$

where $\Delta_{1c} = \exp(\alpha_c + \beta x_i)/\{1 + \exp(\alpha_c + \beta x_i)\}$.

Solving above equation for Δ_{1c} (or α_c), then we obtain

$\widehat{\Delta}_{1c} = \sum_{k=1}^{c} r_{\cdot k}/n$ for $c = 1, \ldots, C-1$ under H_0.

Secondly, we take derivative for β, then we have

$$\partial l/\partial \beta = \sum_{i=0}^{2} \sum_{c=1}^{C} x_i r_{ic} \{(\Delta_{2ci} - \Delta_{2(c-1)i})/(\Delta_{1ci} - \Delta_{1(c-1)i})\}$$

$$= \sum_{i=0}^{2} \sum_{c=1}^{C} x_i r_{ic} \{1 - (\Delta_{1ci} + \Delta_{1(c-1)i})\}$$

where $\widehat{\Delta}_{1c} = \sum_{k=1}^{c} r_{\cdot k}/n$ for $1 \leq c \leq C-1$, and $\widehat{\Delta}_{10} = 0$ and $\widehat{\Delta}_{1C} = 1$.

Then the score function for the above equation under the null hypothesis $\beta = 0$

and with $(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{C-1})$ can be written as follows:

$$\begin{aligned}
S_{\text{POM}} &= \frac{\partial l}{\partial \beta}|_{H_0:\beta=0;\Delta_{1c}=\widehat{\Delta}_{1c},c=1,\ldots,C-1} = \sum_{i=0}^{2}\sum_{c=1}^{C} x_i r_{ic} \frac{\widehat{\Delta}_{2c} - \widehat{\Delta}_{2(c-1)}}{\widehat{\Delta}_{1c} - \widehat{\Delta}_{1(c-1)}} \\
&= \sum_{i=0}^{2}\sum_{c=1}^{C} x_i r_{ic} \left\{1 - \left(\widehat{\Delta}_{1c} + \widehat{\Delta}_{1(c-1)}\right)\right\} \\
&= \sum_{i=0}^{2} x_i r_{i1}(1 - \widehat{\Delta}_{11}) + \sum_{i=0}^{2}\sum_{c=2}^{C} x_i r_{ic} \left\{1 - \left(\widehat{\Delta}_{1c} + \widehat{\Delta}_{1(c-1)}\right)\right\} \\
&= \sum_{i=0}^{2} x_i r_{i1}(1 - \widehat{\Delta}_{11}) - \frac{r_{\cdot 1}}{n}\sum_{i=0}^{2}\sum_{c=2}^{C} x_i r_{ic} \\
&\quad + \frac{r_{\cdot 1}}{n}\sum_{i=0}^{2}\sum_{c=2}^{C} x_i r_{ic} + \sum_{i=0}^{2}\sum_{c=2}^{C} x_i r_{ic} \left\{1 - \left(\widehat{\Delta}_{1c} + \widehat{\Delta}_{1(c-1)}\right)\right\}.
\end{aligned}$$

where $\widehat{\Delta}_{2c} = \widehat{\Delta}_{1c}(1 - \widehat{\Delta}_{1c})$ for $1 \le c \le C$. The right component of the fourth term

$$1 - \left(\widehat{\Delta}_{1c} + \widehat{\Delta}_{1(c-1)}\right) = 1 - (\sum_{k=1}^{c} r_{\cdot k} + \sum_{k=1}^{c-1} r_{\cdot k})/n$$

$$= (-r_{\cdot 1} + \sum_{k=c+1}^{C} r_{\cdot k} - \sum_{k=2}^{c-1} r_{\cdot k})/n$$

So the second term on the right hand side becomes $n^{-1}\sum_{i=0}^{2}\sum_{c=2}^{C} x_i r_{ic}(\sum_{k=c+1}^{C} r_{\cdot k} - \sum_{k=2}^{c-1} r_{\cdot k})$.

The score function can be written as

$$\begin{aligned}
S_{\text{POM}} &= \left(1 - \frac{r_{\cdot 1}}{n}\right)\sum_{i=0}^{2} x_i r_{i1} - \frac{r_{\cdot 1}}{n}\sum_{i=0}^{2}\sum_{c=2}^{C} x_i r_{ic} \qquad (4.2) \\
&\quad + \sum_{i=0}^{2}\sum_{c=2}^{C} x_i r_{ic}\left(\frac{1}{n}\sum_{k=c+1}^{C} r_{\cdot k} - \frac{1}{n}\sum_{k=2}^{c-1} r_{\cdot k}\right). \qquad (4.3)
\end{aligned}$$

When $c = 2$, define $\sum_{k=2}^{c-1} r_{\cdot k} = 0$ in (4.3). The component in (4.2) is the score function from the logistic regression model for the case-control data and the component

in (4.3) is the score function for the POM with outcomes $c = 2, \ldots, C$ in Table 4.1. Therefore, the score function S_{POM} is a linear combination of the respective score functions for the two traits. The score function from the logistic regression model for case-control genetic association studies is equal to (4.2). The term in (4.3) is exactly the score function from the POM with ordinal outcomes of cases only.

Denote $\widehat{\Delta}_{2c} = \widehat{\Delta}_{1c}(1 - \widehat{\Delta}_{1c})$ for $1 \leq c \leq C$. The observed Fisher information matric I_n for $(\alpha_1, \ldots, \alpha_{C-1}, \beta)$, evalued under H_0 with $(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{C-1})$, can be written as

$$I_n = \begin{pmatrix} I^{(1,1)}_{(C-1)\times(C-1)} & I^{(1,2)}_{(C-1)\times 1} \\ I^{(2,1)}_{1\times(C-1)} & I^{(2,2)}_{1\times 1} \end{pmatrix},$$

where $I^{(1,1)}(c, c') = 0$ if $c' > c+1$ and $-r_{\cdot(c+1)}\Delta^2_{2c}\Delta^2_{2(c+1)}/(\Delta^2_{1(c+1)} - \Delta^2_{1c})^2$ if $c' = c+1$, $I^{(1,1)}(c, c) = n\widehat{\Delta}^2_{2c}\{(\widehat{\Delta}_{1c} - \widehat{\Delta}_{1(c-1)})^{-1} + (\widehat{\Delta}_{1(c+1)} - \widehat{\Delta}_{1c})^{-1}\}$, $I^{(1,2)}(c) = \sum_{i=0}^{2} x_i(r_{ic} + r_{i(c+1)})\widehat{\Delta}_{2c}$, $I^{(2,2)} = \sum_{i=0}^{2}\sum_{c=1}^{C-1} x_i\widehat{\Delta}_{2c}(r_{ic} + r_{i(c+1)})$. The proposed score test for $H_0 : \beta = 0$ is then given by

$$J = \frac{S^2_{\text{POM}}}{I^{(2,2)} - I^{(2,1)}\{I^{(1,1)}\}^{-1}I^{(1,2)}}. \tag{4.4}$$

Under H_0, $J \sim \chi^2_1$ asymptotically. Statistical software SAS carries out the above analysis along with the goodness-of-fit test for POM.

Chapter 5

Simulation and Application Studies

In this chapter, we will explore three betas of association results from simulated data. Then, we will explore three betas to application data.

We denote three betas as BetaCC, BetaREG and BetaPOM. The BetaCC is the estimated beta from genetic case control study. When the outcome is a binary trait, we want to know if a genetic marker, e.g. SNP is associated with the binary trait. The estimated BetaCC is log odds ratio, can be obtained from a logistic model. The BetaREG is the estimated beta from genetic quantitative study. When the outcome is a quantitative trait, we want to know if a genetic marker, e.g. SNP is associated with the quantitative trait. The estimated BetaREG is a regression coefficient, can be obtained from a linear regression model. Then we combined the binary trait and the quantitative trait into an ordinal variable as we mentioned in Chapter 4, then we apply a proportional odds model. The estimated BetaPOM is log odds ratio from the proportional odds model.

5.1 Explore Three Betas for Simulated Normal Data

For the simulation, we firstly simulated a binary trait and a quantitative trait $(Y_1, Y_2)^T$ from a bivariate normal distribution with a genetic effect a_0,a_1 , unit variance and correlation ρ. For case-control binary trait, we chose cut_N the 50th percentile of the standart normal distribution as a cut off point for case control binary trait. When it is case, we let the quantitative trait missing. Moreover, We chose minor allele frequency (MAF) as 0.15,0.30 and 0.45, assuming Hardy-Weinberg equilibrium.

We also considered a non-normal setting using $\log|Y_2|$ as a quantitative trait instead of Y_2.

Figure 5.1 is scatter plot of betaCC for simulated normal data, Figure 5.2 is scatter plot of betaREG for simulated normal data, Figure 5.3 is scatter plot of betaPOM for simulated normal data, Figure 5.4 QQ plot of betaPOM vs. betaCC for simulated normal data, it shows the log odds ratio from case-control study is similar to the log odds ratio from the proposed approach.

5.2 Explore Three Betas for Simulated Non-Normal Data

Figure 5.5 is scatter plot of betaCC for simulated non-normal data, Figure 5.6 is scatter plot of betaREG for simulated non-normal data, Figure 5.7 is scatter plot

Figure 5.1: Scatter Plot of BetaCC for Simulated Normal Data

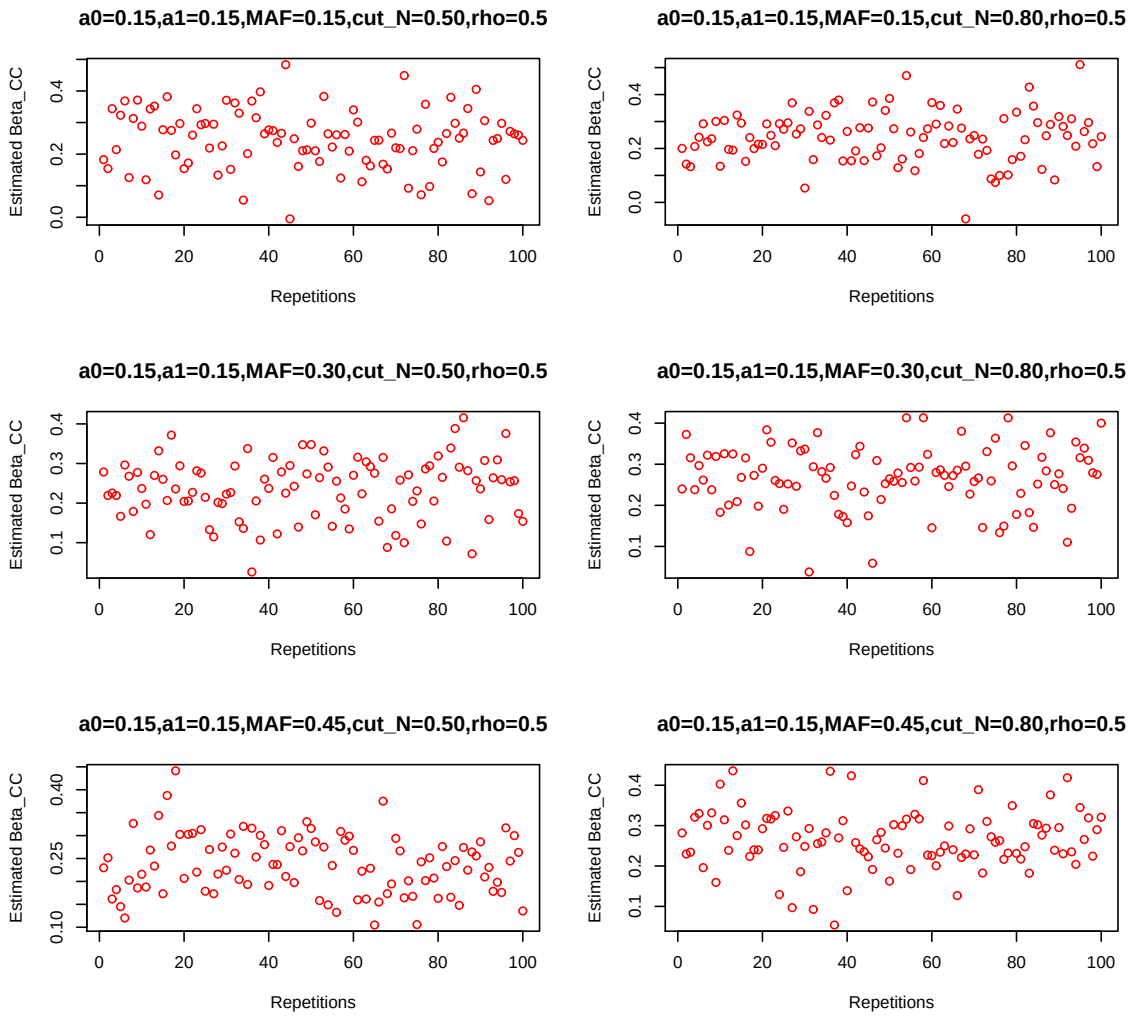


Figure 5.2: Scatter Plot of BetaREG for Simulated Normal Data

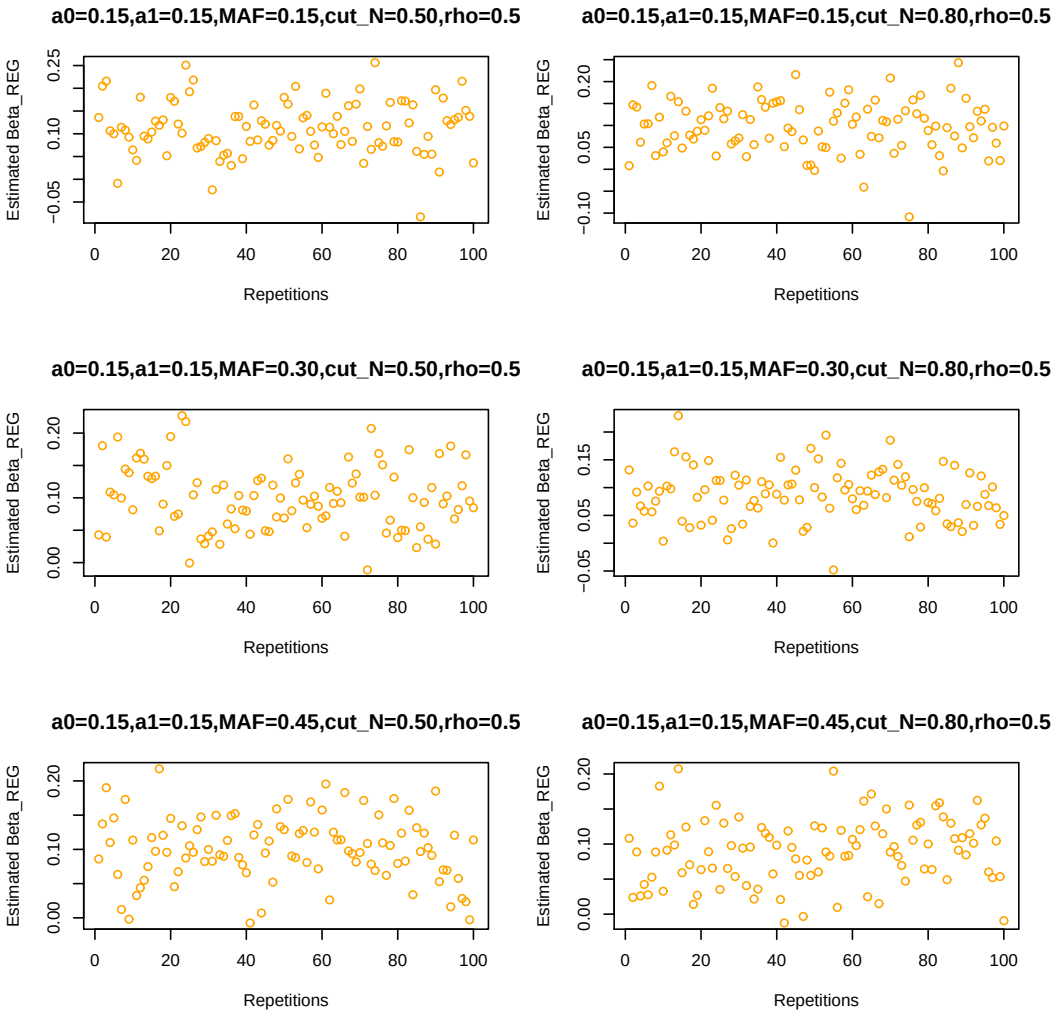


Figure 5.3: Scatter Plot of BetaPOM for Simulated Normal Data

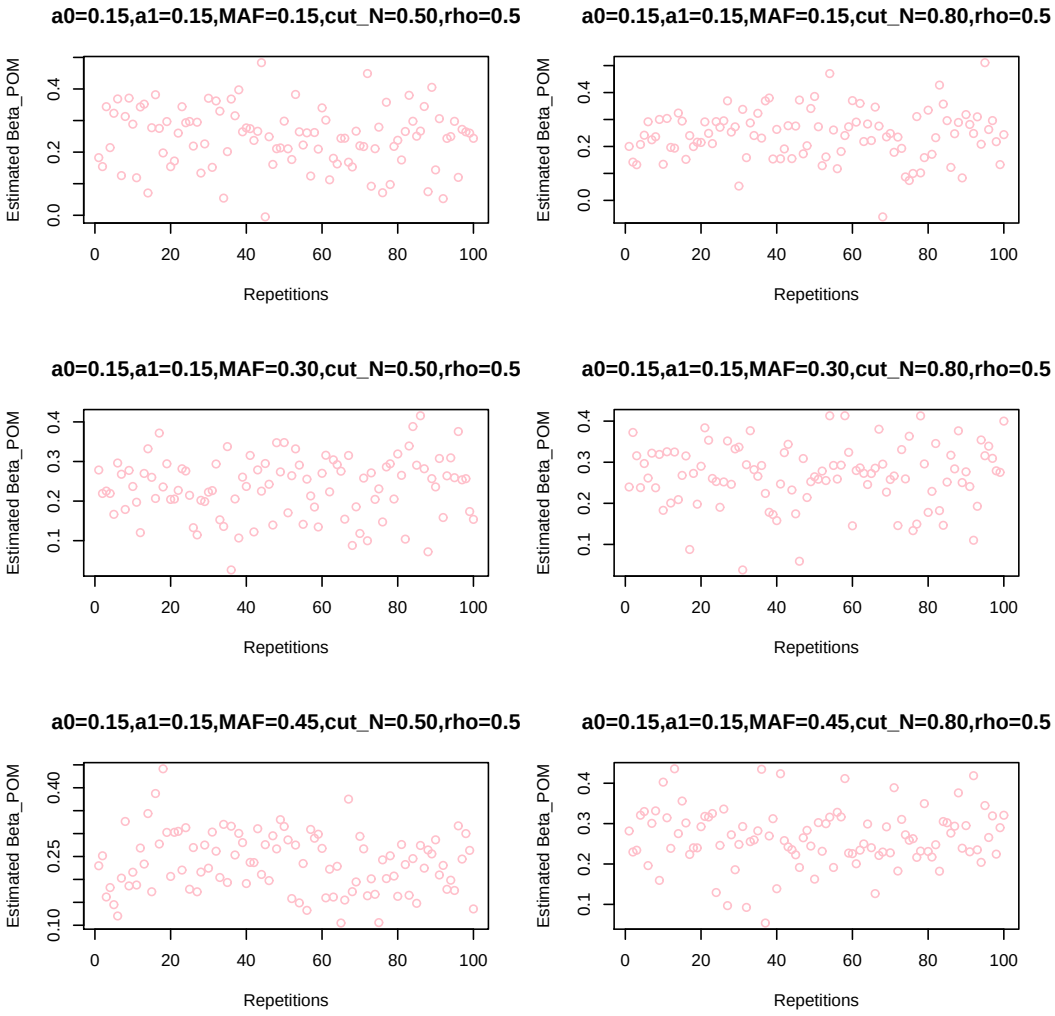
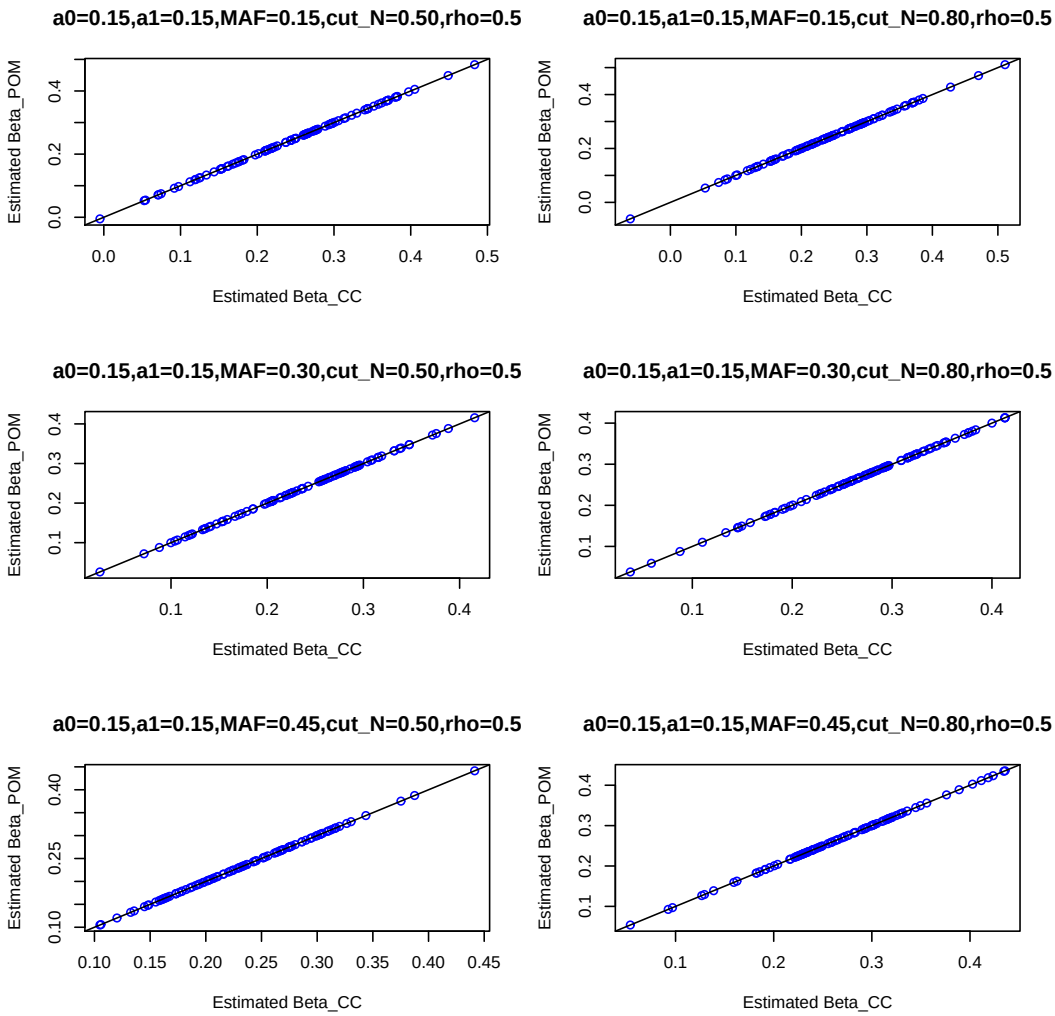


Figure 5.4: QQ Plot of BetaPOM vs. BetaCC for Simulated Normal Data



of betaPOM for simulated non-normal data, Figure 5.8 QQ plot of betaPOM vs. betaCC for simulated non-normal data, it shows the log odds ratio from case-control study is similar to the log odds ratio from the proposed approach.

5.3 Explore Three Betas for Application Data

We selected 91 SNPs from chromosome 5 with anticipated highly signal with the RA disease. QQ plot of betaPOM vs. betaCC shows the log odds ratio from case-control study is similar to the log odds ratio from the proposed approach

Figure 5.5: Scatter Plot of BetaCC for Simulated Non-Normal Data

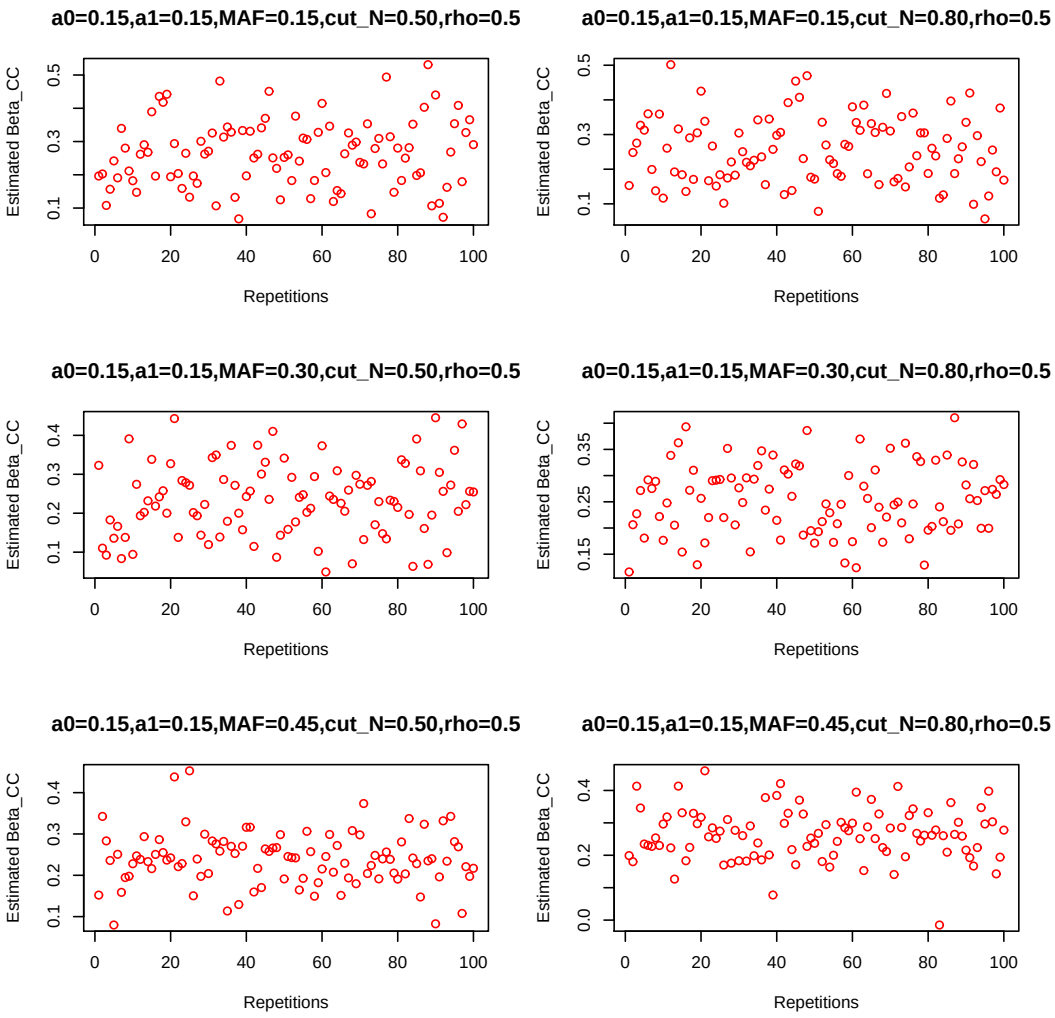


Figure 5.6: Scatter Plot of BetaREG for Simulated Non-Normal Data

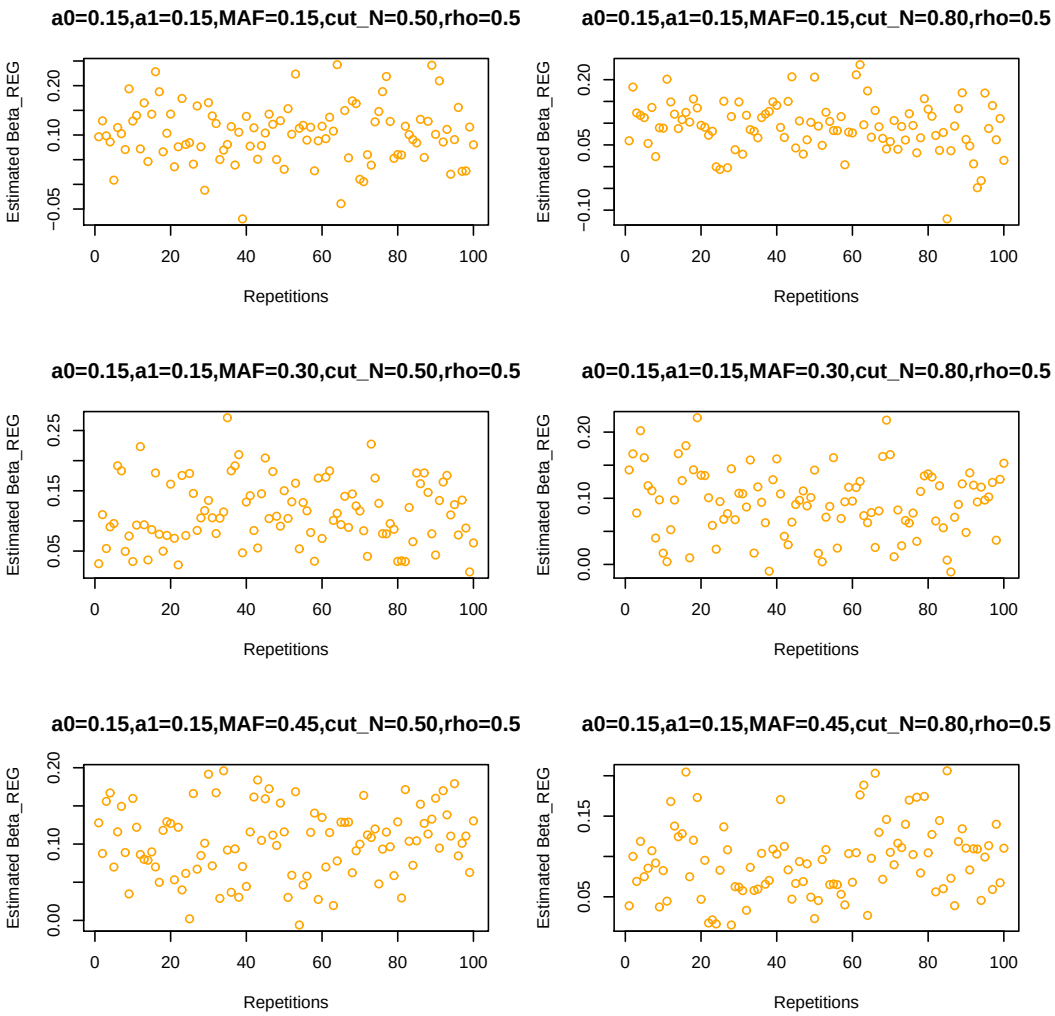


Figure 5.7: Scatter Plot of BetaPOM for Simulated Non-Normal Data

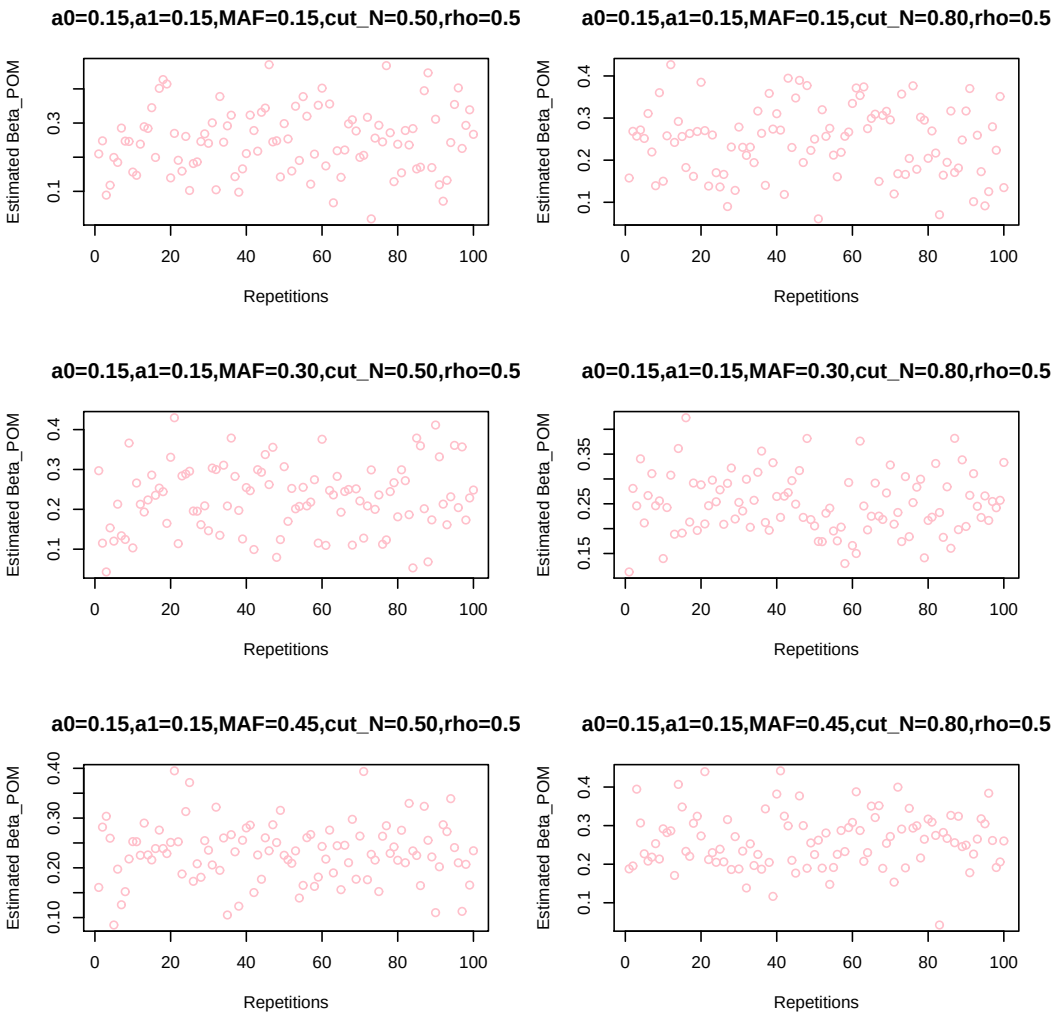


Figure 5.8: QQ Plot of BetaPOM vs. BetaCC for Simulated Non-Normal Data

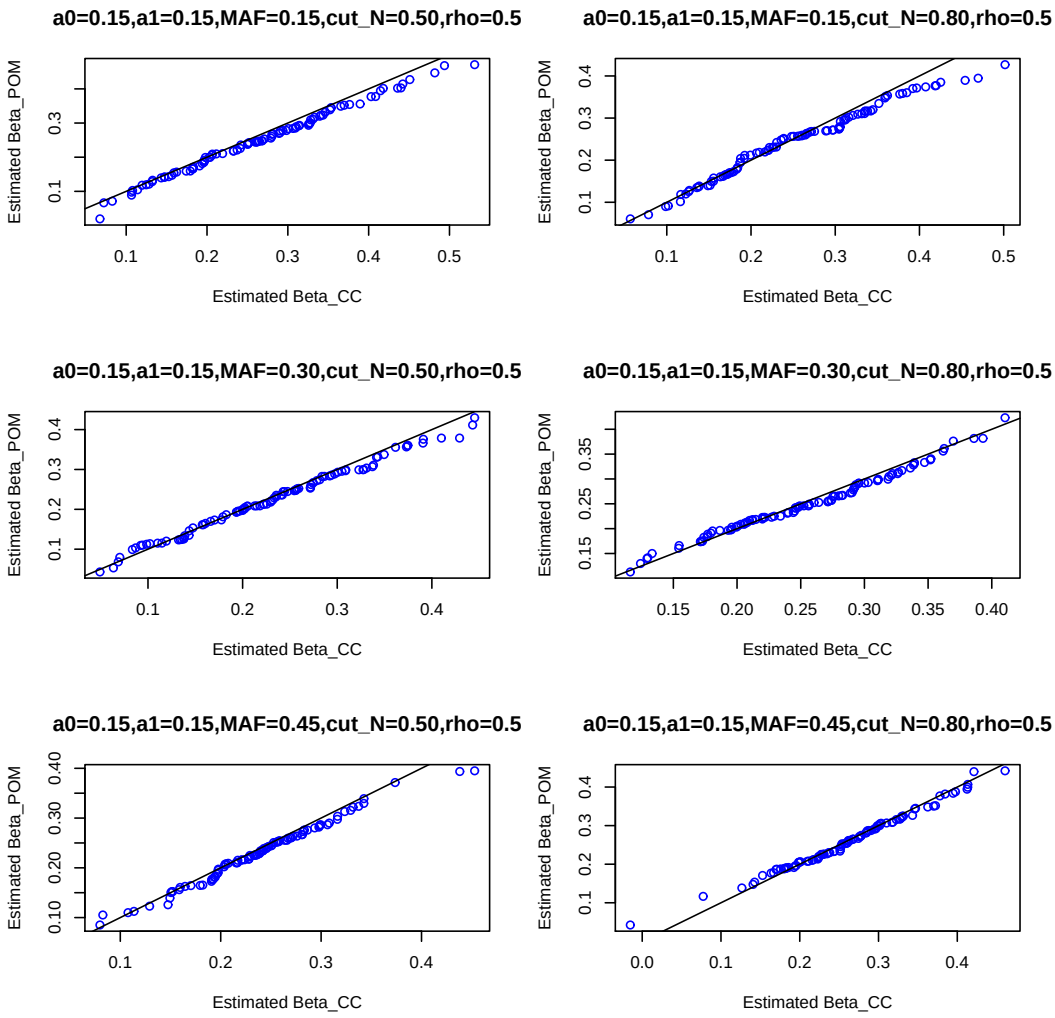
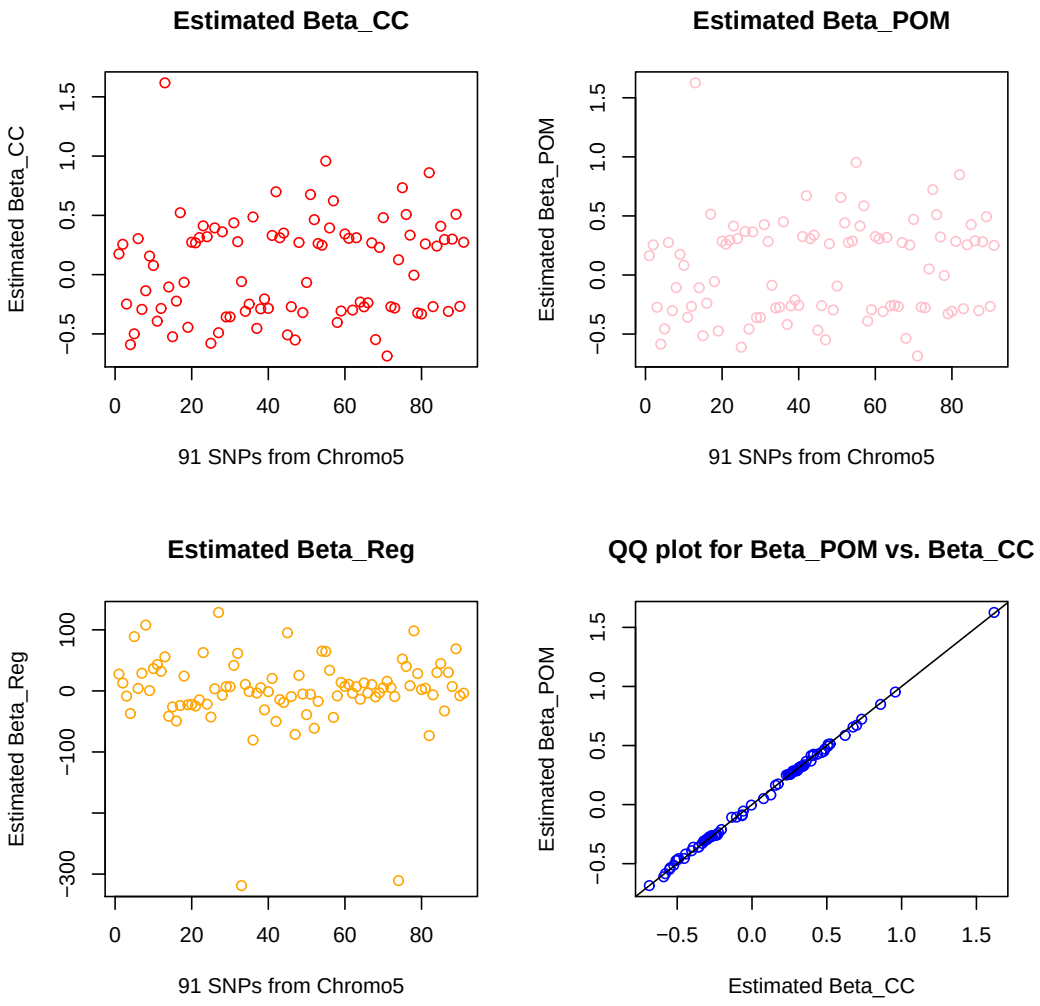


Figure 5.9: Application to 91 SNPs



5.4 Simulation Result of Type I Error

Firstly, a genotype g was simulated for one individual from the multinominal distribution with probabilities $((1-p)^2, 2p(1-p), p^2)$ given the minor allele frequency (MAF) p assuming Hardy-Weinberg equilibrium (HWE). Then mixed traits was simulated as $(Y_1, Y_2)^T$ from a bivariate normal distribution with zero means, unit variances and correlation ρ. Let t be a latent threshold such that $d = 1$ (case) if $Y_1 > c$ and $d = 0$ (control) otherwise. Let Y_2 be a quantitative trait if $d = 1$. But Y_2 was not used if $d = 0$. t was chosen as a cut off point, either the 50th or the 80th percentile of the standard normal distribution. A total of n samples (d_j, g_j, y_j) were simulated, where $d_j = 1$ for $j = 1, \ldots, r$ and $d_j = 0$ for $j = r+1, \ldots, n$, and y_j was not available for $d_j = 0$. We also considered a non-normal setting using $\log|Y_2|$ as a quantitative trait instead of Y_2 but kept d as before.

The first simulation of type I error rate based on normal distributed simulated data is to examine if the distribution $\Gamma(1.3, 3.1)$ can be used for J^*_{uv} with other choices of (u, v)besides J^*_{21} considered by Zheng et al. (2012a). The type I error rates of J^*_{uv} with $(u, v) = (1, 1)$, $(2, 2)$, $(2, 1)$and the proposed test J for a normally distributed trait with norminal level $\alpha = 5\%$ or 1% are reported in Table 5.1. As expected, J^*_{21} and J both have good control of type I error under H_0, but J^*_{11} and J^*_{22} have inflated type I error rates when the same Gamma distribution is used. The results show that all type I error rates are consistent across various ρ, MAFs and cut-off points t. The results also imply that different Gamma distributions may need to be identified when different test statistics are combined while the simple χ^2_1 can be used for the proposed test.

Table 5.1: Type I error rates (%) of J^*_{uv}, $(u,v) = (1,1)$, $(2,2)$ and $(2,1)$, with $\Gamma(1.3, 3.1)$ and J with χ^2_1 with 10,000 replicates. Data are normally distributed. The correlation of the traits is ρ. The cutoff point is t. The nominal level is α.

			$\alpha = 5\%$				$\alpha = 1\%$			
ρ	MAF	t	J^*_{11}	J^*_{22}	J^*_{21}	J	J^*_{11}	J^*_{22}	J^*_{21}	J
0.50	0.15	50th	5.84	5.83	4.95	5.11	1.49	1.57	1.24	0.98
		80th	5.57	5.38	4.69	4.65	1.50	1.35	1.05	0.99
	0.30	50th	5.47	5.51	4.53	4.73	1.40	1.45	1.10	0.97
		80th	6.05	5.65	4.99	5.18	1.62	1.40	1.27	1.02
	0.45	50th	5.90	5.35	4.68	4.75	1.75	1.43	1.24	1.11
		80th	6.11	5.99	5.25	5.34	1.47	1.56	1.13	1.01
0.95	0.15	50th	5.71	5.67	4.91	5.28	1.49	1.44	1.21	0.96
		80th	5.96	5.84	5.20	5.21	1.28	1.36	1.02	1.00
	0.30	50th	5.39	5.45	4.68	5.12	1.30	1.28	0.99	0.89
		80th	5.78	5.64	4.80	5.12	1.41	1.23	1.01	1.11
	0.45	50th	5.76	5.58	4.69	4.96	1.44	1.24	1.09	1.11
		80th	5.04	5.26	4.38	4.68	1.21	1.29	1.00	0.89

Table 5.2: Type I error rates (%) of J_{11} with χ^2_4, J^*_{21} with $\Gamma(1.3, 3.1)$ and J with χ^2_1 with 10,000 replicates. Data are non-normally distributed. The correlation of the traits is ρ. The cutoff point is t. The nominal level is α.

			$\alpha = 5\%$			$\alpha = 1\%$		
ρ	MAF	t	J_{11}	J^*_{21}	J	J_{11}	J^*_{21}	J
0.5	0.15	50th	5.32	4.38	5.07	1.01	0.78	0.96
		80th	5.11	3.91	5.28	0.85	0.66	0.94
	0.3	50th	4.97	4.12	5.01	0.81	0.86	0.94
		80th	4.88	3.59	4.93	0.99	0.79	1.04
	0.45	50th	4.84	4.25	4.81	1.03	0.82	0.91
		80th	5.14	4.35	5.19	0.94	0.75	1.06
0.95	0.15	50th	4.89	2.78	5.20	0.94	0.26	0.90
		80th	5.10	5.66	5.21	0.88	1.29	0.91
	0.3	50th	4.69	2.67	4.76	1.07	0.32	1.16
		80th	5.10	5.35	5.14	1.06	1.14	0.79
	0.45	50th	5.12	3.12	5.16	0.99	0.39	0.93
		80th	4.94	5.09	5.05	1.05	1.34	1.06

The results for the non-normally distributed trait are presented in Table 5.2, where three tests are considered: J_{11}, J_{21}^* and J. Although type I error rates of all three tests are consistent across ρ, MAFs and t, type I error rates for J_{21}^* tend to be conservative when non-normally distributed traits are used. Results in both tables demonstrate that the proposed test J has more robust type I error than J_{uv}^*, where the latter depends on the choice of (u, v) and the underlying distribution.

5.5 Simulation Result of Power Comparison

Simulation procedures under an alternative hypothesis H_1 were similar to those under H_0 except that the traits were simulated conditional on the simulated genotype. After genotype g was simulated, $(Y_1, Y_2)^T$ were simulated from a bivariate normal with means $(\mu_1, \mu_2)^T$, unit variances and correlation ρ, where $\mu_i = -a_i$, d_i or a_i correponding to $g = G_0$, G_1 or G_0, respectively $(i = 1, 2)$. Under the additive model, $d_i = 0$, and under the recessive model, $d_i = -a_i$, $i = 1, 2$, where a_1 and a_2 were pre-specified genetic effects. In the above setting, the non-genetic effect was set to 0 without loss of generality. When $a_1 = a_2 = 0$, the above simulation is identical to the one under H_0. The definition of the case-control data and non-normal traits were the same as under H_0.

Then we focus on testing mixed types of traits associations using the proposed test J, J_{11} and J_{21}^* because their type I error rates are not inflated. The comparisons between the joint analyses using J_{11} and J_{21}^* with single-trait analyses using the trend test, Pearson's test and the F-test were presented before and are not repeated here. The performance of the tests depend on the choices of (a_1, a_2) and whether or not

Table 5.3: Empirical power (%) of J_{11}, J_{21}^* and J with 10,000 replicates under the additive model. The correlation of the traits is $\rho = 0.50$. The cutoff point is t. The genetic effects are (a_1, a_2). The nominal level is 5%.

				Normal			Non-normal		
a_1	a_2	MAF	t	J_{11}	J_{21}^*	J	J_{11}	J_{21}^*	J
0.10	0.15	0.15	50th	60.2	50.7	43.0	32.3	19.9	43.7
			80th	64.3	58.8	47.8	45.0	19.5	58.4
		0.30	50th	81.9	72.0	62.1	50.4	30.9	63.7
			80th	84.6	78.3	67.6	66.2	30.7	77.1
		0.45	50th	87.6	79.1	69.5	58.0	36.1	70.2
			80th	89.2	83.0	71.8	73.5	36.8	83.1
0.15	0.10	0.15	50th	62.0	70.1	76.1	57.2	50.2	67.2
			80th	71.9	80.0	82.0	69.2	57.9	78.2
		0.30	50th	84.3	89.3	92.5	80.1	73.3	86.7
			80th	90.1	94.2	95.1	87.8	79.6	92.7
		0.45	50th	89.6	93.9	96.0	85.8	79.5	91.4
			80th	93.4	96.3	97.2	92.5	85.1	95.6
0.15	0.15	0.15	50th	74.5	76.1	76.2	58.7	46.2	69.3
			80th	80.9	84.2	81.5	73.5	52.3	83.3
		0.30	50th	92.6	93.1	92.9	82.2	69.5	89.7
			80th	95.3	96.5	95.0	90.7	74.0	95.8
		0.45	50th	96.3	96.3	96.1	87.4	76.5	93.6
			80th	97.1	97.8	96.8	94.4	80.4	97.8

the traits follow the normal distribution.

The results for normally and non-normally distributed traits are reported in Table 5.3 under the additive model with $\rho = 0.50$. When the traits follow normal distributions, the results in Table 5.3 show that J_{11} is most powerful when $a_1 < a_2$ and J is most powerful when $a_2 < a_1$ while the three tests have similar performance when $a_1 = a_2$. However, J dominates J_{11} and J_{21}^* when the traits do not follow normal distributions.

The power comparison implies that the proposed test J is more useful than existing joint tests when the association due to the binary trait is much stronger than that

Table 5.4: Empirical power (%) of J_{11}, J_{21}^* and J with 10,000 replicates under the additive model. The correlation of the traits is $\rho = 0.95$. The cutoff point is t. The genetic effects are (a_1, a_2). The nominal level is 5%.

				Normal			Non-normal		
a_1	a_2	MAF	t	J_{11}	J_{21}^*	J	J_{11}	J_{21}^*	J
0.10	0.15	0.15	50th	63.7	56.1	43.8	52.5	25.2	58.5
			80th	69.8	62.9	48.3	70.5	54.0	68.5
		0.30	50th	85.0	77.8	63.5	75.6	46.4	78.9
			80th	88.5	82.7	66.9	89.5	74.4	86.7
		0.45	50th	90.3	83.0	69.0	82.6	57.7	84.8
			80th	93.1	87.7	73.8	93.2	79.8	90.3
0.15	0.10	0.15	50th	55.8	62.0	76.8	56.8	38.2	63.2
			80th	68.7	69.3	81.7	69.9	74.1	65.5
		0.30	50th	79.6	84.4	93.4	79.4	62.5	84.3
			80th	87.8	88.3	95.4	89.2	90.6	85.2
		0.45	50th	85.4	89.6	96.1	85.1	70.1	89.8
			80th	92.4	92.5	97.5	92.5	94.1	90.0
0.15	0.15	0.15	50th	70.0	75.1	77.1	64.5	38.9	76.9
			80th	75.5	81.6	82.5	73.7	80.2	82.4
		0.30	50th	90.2	92.5	93.5	86.5	64.7	93.1
			80th	92.3	95.1	95.6	91.4	94.3	95.5
		0.45	50th	94.0	95.7	96.3	91.4	74.3	95.9
			80th	95.2	97.3	97.5	95.2	96.9	97.6

of the quantitative trait and when the quantitative trait is not normally distributed. In the former case, grouping the quantitative trait does not result in noticeable loss of power. In the latter case, the proposed test is based on the ordinal outcome and more robust with respect to the underlying trait distribution.

The results for normally and non-normally distributed traits are reported in Table 5.4 under the additive model with $\rho = 0.95$. When the traits follow normal distributions, the results in Table 5.4 show that J_{11} is most powerful when $a_1 < a_2$ and J is most powerful when $a_2 < a_1$, the three tests have similar performance when

Table 5.5: Empirical power (%) of J_{11}, J_{21}^* and J with 10,000 replicates under the recessive model. The correlation of the traits is $\rho = 0.50$. The cutoff point is t. The genetic effects are (a_1, a_2). The nominal level is 5%.

				Normal			Non-normal		
a_1	a_2	MAF	t	J_{11}	J_{21}^*	J	J_{11}	J_{21}^*	J
0.10	0.15	0.15	50th	10.1	12.2	9.2	7.1	7.8	8.6
			80th	11.2	14.4	9.5	8.5	7.8	11.0
		0.30	50th	41.6	44.1	29.3	21.5	21.2	29.9
			80th	48.0	52.8	32.6	32.4	25.5	43.5
		0.45	50th	82.3	79.5	61.4	49.2	42.8	61.9
			80th	86.2	87.3	67.5	69.4	51.8	80.0
0.15	0.10	0.15	50th	9.8	18.7	13.9	9.6	14.4	11.1
			80th	12.5	25.1	16.6	11.5	18.0	14.7
		0.30	50th	41.0	67.0	55.6	36.5	53.6	46.0
			80th	53.4	81.3	65.4	52.1	67.7	62.5
		0.45	50th	83.2	95.4	92.8	77.6	87.0	85.3
			80th	91.5	98.7	96.2	90.4	94.8	94.3
0.15	0.15	0.15	50th	12.8	20.4	14.2	9.8	13.0	12.7
			80th	15.5	27.5	16.4	12.1	17.5	16.8
		0.30	50th	55.1	73.2	56.8	38.4	50.5	49.8
			80th	67.0	85.2	67.1	56.5	63.8	68.8
		0.45	50th	92.2	97.0	92.1	80.4	85.5	88.4
			80th	96.8	99.2	96.2	93.2	92.8	97.0

$a_1 = a_2$. However, J dominates J_{11} and J_{21}^* when the traits do not follow normal distributions.

The results for normally and non-normally distributed traits are reported in Table 5.5 under the recessive model with $\rho = 0.50$. When the traits follow normal distributions, the results in Table 5.5 show that J_{21}^* is most powerful. However, J dominates J_{11} and J_{21}^* when the traits do not follow normal distributions.

The results for normally and non-normally distributed traits are reported in Table 5.6 under the recessive model with $\rho = 0.95$. When the traits follow normal

Table 5.6: Empirical power (%) of J_{11}, J_{21}^* and J with 10,000 replicates under the recessive model. The correlation of the traits is $\rho = 0.95$. The cutoff point is t. The genetic effects are (a_1, a_2). The nominal level is 5%.

				Normal			Non-normal		
a_1	a_2	MAF	t	J_{11}	J_{21}^*	J	J_{11}	J_{21}^*	J
0.10	0.15	0.15	50th	10.8	12.2	8.3	9.0	6.5	11.0
			80th	13.0	15.7	9.5	11.6	14.1	13.4
		0.30	50th	45.9	49.0	29.7	36.3	26.6	41.7
			80th	55.5	58.7	33.8	52.0	51.0	52.6
		0.45	50th	85.1	83.6	62.1	75.6	62.3	79.6
			80th	91.7	91.0	70.3	90.2	84.1	88.1
0.15	0.10	0.15	50th	9.8	16.7	14.1	9.6	11.7	10.9
			80th	12.2	21.4	17.2	12.9	23.0	11.7
		0.30	50th	36.7	62.6	56.4	36.3	48.0	42.6
			80th	51.7	74.7	68.5	52.4	77.0	48.0
		0.45	50th	78.0	92.8	92.7	76.8	83.4	82.5
			80th	89.2	96.8	96.7	90.6	97.8	86.2
0.15	0.15	0.15	50th	11.6	19.3	13.4	10.5	12.1	13.9
			80th	13.0	24.7	16.4	13.1	25.2	17.2
		0.30	50th	49.6	72.0	56.9	43.9	48.6	56.7
			80th	60.6	83.8	68.6	57.9	82.7	68.5
		0.45	50th	89.7	97.2	92.9	85.7	86.3	92.7
			80th	94.0	99.1	96.5	93.1	98.7	96.5

Table 5.7: Proportions of 10,000 replicates under H_0 that the goodness-of-fit test for POM assumption is rejected at 0.05 level.

ρ	MAF	t	Normal	Non-normal
0.5	0.15	50th	5.30%	5.31%
		80th	5.00%	5.73%
	0.30	50th	5.12%	5.35%
		80th	5.20%	5.16%
	0.45	50th	5.14%	5.39%
		80th	5.52%	5.09%
0.95	0.15	50th	5.51%	5.45%
		80th	5.78%	5.91%
	0.30	50th	5.54%	5.47%
		80th	5.26%	4.94%
	0.45	50th	5.33%	5.14%
		80th	5.39%	4.90%

distributions, the results in Table 5.6 show that J_{21}^* is most powerful. When the traits do not follow normal distributions, J dominates J_{11} and J_{21}^* when $a_1 = a_2$.

5.6 The POM Assumption and Related Tests

Under $H0$, the rejection rates for testing the POM assumption were simulated with various ρ, MAF and t for the normal and non-normal traits.

In Table 5.7, the results show the POM assumption is rejected at about 5% with range from 4.94% to 5.91%.

The simulated rejection rates under H1 are reported in Table 5.8, which show the POM assumption is rejected more often than under H0. Under the normal tratis, the rejection rate is between 6% and 7%, while under the non-normal traits, the rate is still between 6% and 7% when a1 = a2, but much higher (up to 36%) when $a1 \neq a2$.

Table 5.8: Proportions of 10,000 replicates under H_1 that the goodness-of-fit test for POM assumption is rejected at 0.05 level.

					Additive Model		Recessive Model	
a_1	a_2	ρ	MAF	t	Normal	Non-normal	Normal	Non-normal
0.10	0.15	0.50	0.15	50th	5.86%	6.80%	5.63%	5.83%
				80th	5.08%	8.67%	4.95%	6.70%
			0.30	50th	5.71%	6.65%	5.18%	6.73%
				80th	4.91%	9.24%	5.40%	8.39%
			0.45	50th	5.80%	6.66%	5.79%	7.28%
				80th	5.68%	9.76%	5.23%	10.63 %
		0.95	0.15	50th	5.67%	14.44%	6.06%	7.34%
				80th	5.73%	17.37%	5.29%	7.68%
			0.30	50th	5.82%	19.01%	6.05%	11.93%
				80th	5.65%	22.90%	5.88%	13.26%
			0.45	50th	5.76%	20.23%	5.62%	20.48%
				80th	5.98%	24.51%	5.79%	25.36%
0.15	0.10	0.50	0.15	50th	6.54%	6.76%	6.11%	5.69%
				80th	5.55%	7.09%	5.06%	5.34%
			0.30	50th	6.16%	8.27%	6.62%	6.43%
				80th	5.80%	8.65%	5.24%	6.67%
			0.45	50th	5.79%	9.31%	7.36%	8.46%
				80th	6.15%	9.77%	5.53%	8.82%
		0.95	0.15	50th	6.40%	9.37%	6.21%	5.75%
				80th	6.06%	19.96%	6.14%	6.21%
			0.30	50th	6.70%	13.51%	6.37%	7.96%
				80th	6.15%	32.64%	5.58%	15.65%
			0.45	50th	6.42%	14.23%	6.72%	13.36%
				80th	5.92%	35.46%	6.32%	36.46%
0.15	0.15	0.50	0.15	50th	5.93%	6.15%	5.83%	5.95 %
				80th	5.42%	6.43%	5.29%	6.11 %
			0.30	50th	6.44%	6.79%	6.21%	6.42 %
				80th	5.78%	6.18%	5.01%	6.10 %
			0.45	50th	5.90%	7.11%	6.78%	6.92 %
				80th	5.28%	6.17%	5.87%	6.38 %
		0.95	0.15	50th	6.57%	6.39%	5.93%	6.31 %
				80th	6.38%	5.61%	5.78%	6.10 %
			0.30	50th	6.37%	6.34%	6.57%	6.43 %
				80th	5.80%	6.19%	5.99%	5.73 %
			0.45	50th	6.48%	5.89%	6.73%	7.30 %
				80th	6.40%	5.71%	6.35%	5.70 %

Table 5.9: Type I error rates (%) of J, J_G, J^*, and J_A with 10,000 replicates under H_0. The nominal level is 5%. Test J^* is the same as J except that the replicates without the POM assumption (at the same 5% level) are deleted.

Trait	ρ	MAF	t	J	J_G	J^*	J_A	POM fails
Normal	0.5	0.15	50th	5.16	5.25	5.12	8.08	5.64%
			80th	4.94	4.99	4.94	7.16	5.39%
		0.30	50th	5.10	4.84	5.14	7.72	4.97%
			80th	4.92	5.17	4.91	7.30	5.31%
		0.45	50th	5.19	4.62	5.16	7.68	5.04%
			80th	4.60	4.58	4.65	6.83	5.16%
	0.95	0.15	50th	4.70	4.65	4.81	7.58	5.71%
			80th	4.75	4.62	4.76	7.64	5.65%
		0.30	50th	5.14	4.49	5.05	7.78	4.99%
			80th	5.08	4.83	5.07	8.26	5.75%
		0.45	50th	4.87	4.81	4.92	7.85	4.99%
			80th	5.13	4.69	5.09	7.97	4.93%
Non-normal	0.5	0.15	50th	5.21	4.57	5.15	7.94	5.45%
			80th	5.04	5.01	5.01	8.18	6.02%
		0.30	50th	5.00	5.20	5.01	8.29	5.57%
			80th	5.08	4.87	5.12	8.04	5.35%
		0.45	50th	5.19	4.89	5.19	8.48	5.59%
			80th	4.65	4.93	4.54	7.75	5.16%
	0.95	0.15	50th	4.68	4.61	4.72	7.54	5.49%
			80th	4.83	4.37	4.80	7.55	5.29%
		0.30	50th	5.23	4.93	5.23	8.08	5.15%
			80th	5.02	4.89	5.02	8.11	5.33%
		0.45	50th	5.16	5.16	5.23	8.56	5.61%
			80th	4.97	4.69	5.05	7.96	5.26%

Table 5.9 shows the type I error rates for normally and non-normally distributed traits under the null with $\rho = 0.50, 0.95$. POM assumption failure rate is about 5% from 4.93% to 6.02%. J , J_G and J^* have good control of type I error rate, but J_A has a inflated type I error.

Table 5.10 shows the empirical power rates for normally distributed traits under the alternative for additive model with $\rho = 0.50, 0.95$. POM assumption failure rate is about 6% from 5.13% to 6.62%. J and J^* have similar power which is higher than J_G. J_A has the highest power, slightly higher than J and J^*, but with a inflated type I error.

Table 5.11 shows the empirical power rates for non-normal distributed traits under the alternative for additive model with $\rho = 0.50, 0.95$. POM assumption failure rate is much higher, up to 34.46%, when $a0 \neq a1$. J and J^* have similar power which is higher than J_G. J_A has the highest power, slightly higher than J and J^*, but with a inflated type I error.

Table 5.12 shows the empirical power rates for normally distributed traits under the alternative for recessive model with $\rho = 0.50, 0.95$. POM assumption failure rate is about 5% to 7%. J and J^* have similar power which is higher than J_G. J_A has the highest power, slightly higher than J and J^*, but with a inflated type I error.

Table 5.13 shows the empirical power rates for non-normal distributed traits under the alternative for recessive model with $\rho = 0.50, 0.95$. POM assumption failure rate is much higher, to up 34.35%, when $a0 \neq a1$. J and J^* have similar power which is higher than J_G. J_A has the highest power, slightly higher than J and J^*, but with a inflated type I error.

Table 5.10: Empirical power (%) of J, J_G J^* and J_A with 10,000 replicates under the additive model and with normal traits. The nominal level is 5%. Test J^* is the same as J except that the replicates without the POM assumption (at the same 5% level) are deleted.

a_1	a_2	ρ	MAF	t	J	J_G	J^*	J_A	POM fails
0.10	0.15	0.50	0.15	50th	42.53	29.35	42.57	44.86	5.76%
				80th	47.62	38.66	47.77	49.69	5.14%
			0.30	50th	62.43	46.11	62.68	64.30	5.68%
				80th	66.39	55.46	66.24	67.57	5.13%
			0.45	50th	70.67	53.60	70.72	72.15	5.90%
				80th	72.11	61.69	72.03	73.17	5.27%
		0.95	0.15	50th	43.75	27.01	43.97	46.44	5.98%
				80th	47.87	30.21	47.82	50.02	5.41%
			0.30	50th	63.53	43.37	63.37	65.20	5.97%
				80th	68.56	46.82	68.38	69.87	5.64%
			0.45	50th	70.35	48.91	70.55	71.94	5.73%
				80th	73.62	52.55	73.64	74.97	5.79%
0.15	0.10	0.50	0.15	50th	76.56	61.60	76.54	77.66	6.05%
				80th	81.65	72.97	81.67	82.48	5.26%
			0.30	50th	92.85	83.54	92.83	93.25	6.29%
				80th	94.88	90.63	94.81	95.07	5.34%
			0.45	50th	95.46	89.12	95.49	95.73	5.86%
				80th	97.17	94.35	97.21	97.36	5.69%
		0.95	0.15	50th	76.98	57.64	77.13	78.35	6.62%
				80th	82.25	63.98	82.25	83.20	5.90%
			0.30	50th	93.16	80.79	93.30	93.64	6.41%
				80th	95.33	85.28	95.27	95.57	6.48%
			0.45	50th	96.10	87.18	96.09	96.33	6.39%
				80th	97.33	89.31	97.30	97.45	5.85%
0.15	0.15	0.50	0.15	50th	76.67	61.16	76.64	77.87	6.27%
				80th	82.02	74.27	81.97	82.82	5.56%
			0.30	50th	93.00	83.53	93.01	93.41	6.04%
				80th	94.58	90.00	94.52	94.81	5.73%
			0.45	50th	95.89	89.08	95.92	96.13	6.33%
				80th	97.05	94.11	97.00	97.16	5.56%
		0.95	0.15	50th	77.81	57.66	78.11	79.13	6.38%
				80th	83.12	64.25	83.04	83.88	5.87%
			0.30	50th	93.02	80.84	93.15	93.55	6.41%
				80th	95.46	85.43	95.41	95.67	6.20%
			0.45	50th	96.09	86.79	96.08	96.27	6.14%
				80th	97.14	89.43	97.14	97.29	5.96%

Table 5.11: Empirical power (%) of J, J_G J^* and J_A with 10,000 replicates under the additive model and with non-normal traits. The nominal level is 5%. Test J^* is the same as J except that the replicates without the POM assumption (at the same 5% level) are deleted.

a_1	a_2	ρ	MAF	t	J	J_G	J^*	J_A	POM fails
0.10	0.15	0.50	0.15	50th	42.22	26.60	42.42	44.93	6.31%
				80th	57.94	40.60	57.86	60.58	8.89%
			0.30	50th	62.24	42.86	62.30	64.33	6.60%
				80th	77.07	60.20	76.95	78.89	9.49%
			0.45	50th	71.15	50.00	71.31	72.94	6.66%
				80th	82.69	66.90	82.80	84.35	9.92%
		0.95	0.15	50th	59.73	46.66	59.74	63.99	14.50%
				80th	68.98	56.47	69.11	73.07	16.47%
			0.30	50th	79.54	68.89	79.55	82.99	18.88%
				80th	87.07	79.32	87.22	89.81	22.44%
			0.45	50th	84.92	75.65	85.19	87.84	20.13%
				80th	90.80	83.60	90.81	92.76	23.21%
0.15	0.10	0.50	0.15	50th	66.61	48.37	66.72	68.65	6.69%
				80th	77.82	59.96	77.70	79.18	6.96%
			0.30	50th	86.61	71.59	86.72	87.78	8.56%
				80th	93.17	81.67	93.04	93.65	8.92%
			0.45	50th	91.06	78.47	90.96	91.73	9.52%
				80th	95.65	86.89	95.66	96.00	8.62%
		0.95	0.15	50th	62.47	48.56	62.65	65.89	9.46%
				80th	64.29	61.48	64.12	71.13	20.43%
			0.30	50th	83.71	70.95	83.60	85.51	12.52%
				80th	84.58	82.94	84.46	89.24	31.79%
			0.45	50th	89.92	78.91	89.84	91.22	14.68%
				80th	89.70	86.94	89.79	93.39	36.81%
0.15	0.15	0.50	0.15	50th	70.36	50.94	70.35	71.96	6.44%
				80th	84.31	66.22	84.28	85.15	5.93%
			0.30	50th	89.61	75.03	89.66	90.28	6.45%
				80th	95.60	85.89	95.67	95.92	6.28%
			0.45	50th	93.61	81.72	93.62	94.03	6.40%
				80th	97.73	90.72	97.73	97.87	6.42%
		0.95	0.15	50th	77.49	57.51	77.68	78.74	6.67%
				80th	83.23	64.80	83.55	84.32	5.62%
			0.30	50th	93.01	80.37	93.17	93.60	6.68%
				80th	95.67	85.86	95.67	95.92	6.08%
			0.45	50th	96.17	86.68	96.21	96.44	6.60%
				80th	97.45	89.78	97.41	97.55	5.37%

Table 5.12: Empirical power (%) of J, J_G J^* and J_A with 10,000 replicates under the recessive model and with normal traits. The nominal level is 5%. Test J^* is the same as J except that the replicates without the POM assumption (at the same 5% level) are deleted.

a_1	a_2	ρ	MAF	t	J	J_G	J^*	J_A	POM fails
0.10	0.15	0.50	0.15	50th	9.21	7.30	9.15	12.07	5.37%
				80th	9.68	8.43	9.55	11.87	5.15%
			0.30	50th	28.55	19.92	28.53	31.47	5.75%
				80th	33.42	26.33	33.48	35.72	4.90%
			0.45	50th	62.09	46.33	62.13	64.08	5.86%
				80th	67.82	57.84	67.85	69.29	5.71%
		0.95	0.15	50th	8.23	6.91	8.26	11.33	5.88%
				80th	9.44	7.54	9.35	12.55	5.95%
			0.3	50th	29.88	18.39	29.73	32.93	6.10%
				80th	34.63	21.89	34.52	37.53	5.58%
			0.45	50th	61.72	41.35	61.52	63.41	5.55%
				80th	68.69	48.36	68.59	70.14	5.76%
0.15	0.10	0.50	0.15	50th	13.19	10.56	13.16	16.44	5.95%
				80th	16.12	13.01	16.04	18.70	5.32%
			0.30	50th	56.04	42.04	55.91	58.45	7.01%
				80th	66.87	57.26	66.86	68.17	4.77%
			0.45	50th	92.25	83.31	92.15	92.72	7.27%
				80th	96.47	93.13	96.52	96.70	5.51%
		0.95	0.15	50th	13.61	10.05	13.58	16.96	6.10%
				80th	16.82	11.46	16.76	19.91	5.68%
			0.30	50th	57.40	38.68	57.17	59.56	6.48%
				80th	68.13	48.15	68.16	69.91	6.06%
			0.45	50th	92.85	80.34	92.83	93.30	7.16%
				80th	96.12	87.52	96.11	96.32	5.86%
0.15	0.15	0.50	0.15	50th	13.35	9.86	13.23	16.20	5.71%
				80th	16.90	13.71	16.96	19.38	5.21%
			0.30	50th	56.67	41.32	56.34	58.74	6.16%
				80th	67.04	56.65	67.23	68.47	5.07%
			0.45	50th	92.42	83.59	92.39	92.89	7.01%
				80th	96.03	92.64	96.08	96.27	5.40%
		0.95	0.15	50th	14.35	9.69	14.16	17.53	6.03%
				80th	16.58	11.19	16.74	19.74	5.51%
			0.30	50th	56.37	38.86	56.56	58.91	6.49%
				80th	68.38	47.85	68.47	70.15	5.86%
			0.45	50th	93.16	81.22	93.24	93.67	6.86%
				80th	96.74	88.27	96.79	96.99	6.15%

Table 5.13: Empirical power (%) of J, J_G J^* and J_A with 10,000 replicates under the recessive model and with non-normal traits. The nominal level is 5%. Test J^* is the same as J except that the replicates without the POM assumption (at the same 5% level) are deleted.

a_1	a_2	ρ	MAF	t	J	J_G	J^*	J_A	POM fails
0.10	0.15	0.50	0.15	50th	9.40	7.22	9.36	12.46	5.69%
				80th	11.57	8.67	11.65	14.83	6.49%
			0.30	50th	29.68	19.03	29.52	32.97	6.48%
				80th	42.76	29.91	42.83	46.59	8.64%
			0.45	50th	63.34	44.24	63.56	65.65	6.90%
				80th	79.71	64.41	79.69	81.64	10.54%
		0.95	0.15	50th	10.61	8.92	10.52	14.42	7.66%
				80th	13.54	10.76	13.61	17.66	8.07%
			0.3	50th	42.12	32.52	41.75	47.35	11.81%
				80th	53.60	42.89	53.43	59.04	14.38%
			0.45	50th	78.37	69.59	78.30	82.37	20.43%
				80th	87.82	82.35	87.35	90.35	25.16%
0.15	0.10	0.50	0.15	50th	11.09	8.36	11.05	14.27	5.28%
				80th	14.90	10.82	14.95	18.10	5.55%
			0.30	50th	45.11	30.75	45.00	47.99	6.12%
				80th	62.12	43.76	62.07	64.15	6.20%
			0.45	50th	84.53	69.77	84.55	85.73	8.16%
				80th	94.59	84.96	94.61	95.09	9.06%
		0.95	0.15	50th	10.29	8.22	10.40	13.52	5.55%
				80th	11.88	11.11	12.02	16.40	6.39%
			0.30	50th	42.81	31.26	42.86	46.95	8.04%
				80th	48.29	45.18	48.26	56.16	16.17%
			0.45	50th	82.26	69.89	82.29	84.53	13.23%
				80th	85.68	85.58	85.31	90.52	36.15%
0.15	0.15	0.50	0.15	50th	11.92	8.61	11.93	15.03	5.78%
				80th	18.13	11.93	18.13	21.52	6.25%
			0.30	50th	49.96	33.21	50.03	52.88	6.70%
				80th	69.85	50.34	69.74	71.43	6.10%
			0.45	50th	88.89	73.71	88.76	89.52	6.87%
				80th	96.74	89.56	96.78	96.99	6.95%
		0.95	0.15	50th	14.20	10.18	14.16	17.42	5.89%
				80th	16.64	11.08	16.52	19.50	5.27%
			0.30	50th	56.26	37.95	56.17	58.57	6.26%
				80th	68.79	48.57	68.86	70.45	5.68%
			0.45	50th	92.97	80.80	92.87	93.40	7.69%
				80th	96.89	88.20	96.89	97.08	6.00%

5.7 Application to GAW16 Data

In this section, we will compare our proposed approach with the other two existing approachs in a RA data. The RA data (Problem 1 of GAW16–the Genetic Analysis Workshop 16) is the initial whole-genome association study, comprising 868 cases and 1,194 controls, more than 500,000 SNPs are genotyped.

For illustration, we considered 32,045 SNPs on chromosome 5 and 33,797 SNPs on chromosome 6. For chromosome5, we check the POM assumption using the Bonferroni corrected significance level with 1.56×10^{-6}, and the level 10^{-4}. The numbers of SNPs failed the POM test are 148 (0.46%) with the level 10^{-4} and 107 (0.33%) with the Bonferroni correction.

Then we deleted the above 148 SNPs which couldn't satisfy the assumption and analyzed the remaining SNPs on chromosome 5 for pleiotropic association. First, we selected three sub-sets of SNPs whose p-values based on J, J_{11} and J_{21}^* are $< 10^{-4}$, respectively. The numbers of SNPs in the three sub-sets are 69, 48 and 29, respectively, indicating J can detect more SNPs with p-values $< 10^{-4}$ than J_{11} or J_{21}^*. Combining all the unique SNPs of the three sub-sets, we obtained 91 SNPs.

The left panel of Figure 2 plots the p-values, in $-\log_{10}$ scale, of J_{11} vs. J (row 1), J_{21}^* vs. J (row 2), and J_{11} vs. J_{21}^* (row 3). In the plots, solid squares, plus, and open circles represent the p-values of J_{11}, J_{21}^* and J, respectively. The results show J has more smaller p-values than J_{11} (65 vs. 26), and both J and J_{11} have more smaller p-values than J_{21}^*. Comparing the p-values of J_{11} and J within the three sub-sets, the numbers of SNPs for which J has smaller p-values are 60, 26, and 18, respectively. Hence, even for the SNPs that were selected based on the performance of J_{11}, there are still 26 SNPs out of 48 SNPs, with which J has smaller p-values than J_{11}. The

right panel of the figure presents the Q-Q plots of all the SNPs on chromosome 5, which shows J and J_{11} are overall comparable except in the tail, where J tends to have smaller p-values.

For chromosome6, we check the POM assumption using the Bonferroni corrected significance level with 1.48×10^{-6}, and the level 10^{-4}. The numbers of SNPs failed the POM test are 262 (0.78%) with the level 10^{-4} and 195 (0.58%) with the Bonferroni correction.

Then we deleted the above 262 SNPs which couldn't satisfy the assumption and analyzed the remaining SNPs on chromosome 6 for pleiotropic association. First, we selected three sub-sets of SNPs whose p-values based on J, J_{11} and J_{21}^* are $< 10^{-4}$, respectively. The numbers of SNPs in the three sub-sets are 505, 433 and 322, respectively, indicating J can detect more SNPs with p-values $< 10^{-4}$ than J_{11} or J_{21}^*. Combining all the unique SNPs of the three sub-sets, we obtained 541 SNPs.

The left panel of Figure 3 plots the p-values, in $-\log_{10}$ scale, of J_{11} vs. J (row 1), J_{21}^* vs. J (row 2), and J_{11} vs. J_{21}^* (row 3). In the plots, solid squares, plus, and open circles represent the p-values of J_{11}, J_{21}^* and J, respectively. The results show J has more smaller p-values than J_{11} (409 vs. 132), and both J and J_{11} have more smaller p-values than J_{21}^*. Comparing the p-values of J_{11} and J within the three sub-sets, the numbers of SNPs for which J has smaller p-values are 399, 306, and 233, respectively. Hence, even for the SNPs that were selected based on the performance of J_{11}, there are still 306 SNPs out of 433 SNPs, with which J has smaller p-values than J_{11}. The right panel of the figure presents the Q-Q plots of all the SNPs on chromosome 6, which shows J and J_{11} are overall comparable except in the tail, where J tends to have smaller p-values.

Figure 5.10: Plots (the left panel) of the p-values of J_{11} (solid square) and J (open circle) (row 1), J_{21}^* plus) and J (open circle) (row 2), and J_{11} (solid square) and J_{21}^* (plus) (row 3) for 91 SNPs of chromosome 5 whose p-values are less than 10^{-4} with at least one of the three tests. Q-Q plots (the right panel) of the Clog10 (p-values) of the three tests for chromosome 5.

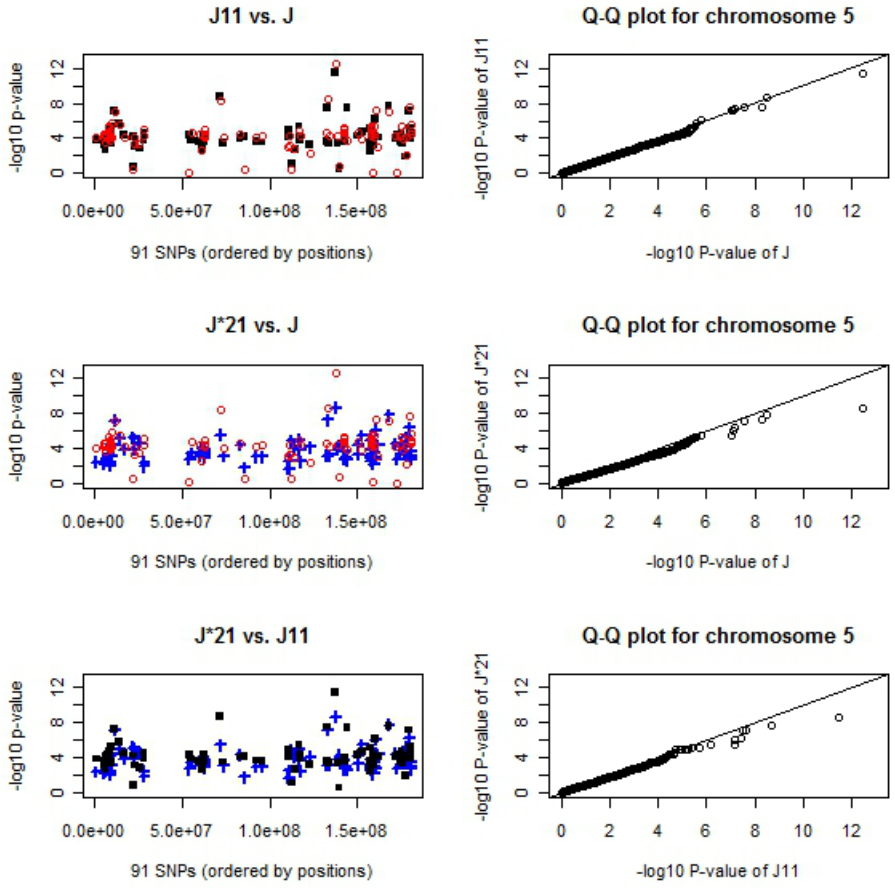
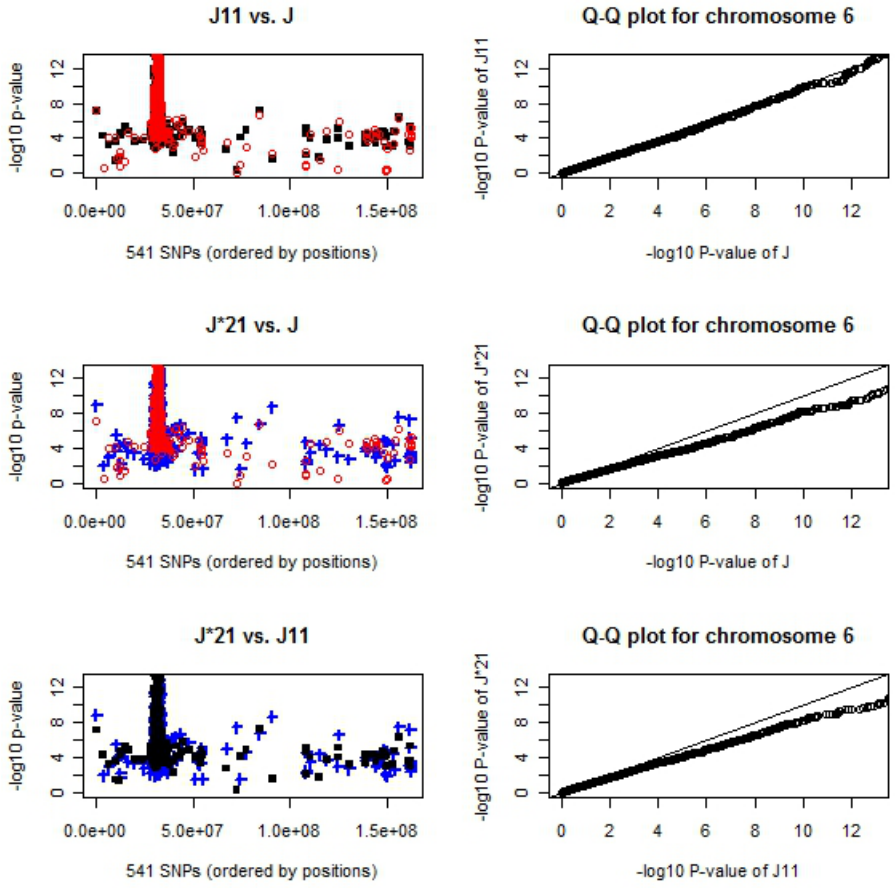


Figure 5.11: Plots (the left panel) of the p-values of J_{11} (solid square) and J (open circle) (row 1), J_{21}^* (plus) and J (open circle) (row 2), and J_{11} (solid square) and J_{21}^* (plus) (row 3) for 541 SNPs of chromosome 6 whose p-values are less than 10^{-4} with at least one of the three tests. Q-Q plots (the right panel) of the Clog10 (p-values) of the three tests for chromosome 6.



Chapter 6

Future Research

In the future research, two parameters will be set up for a usual proporitional odds model (POM) rather than one parameter.

As we recall, for the data displayed in Table 4.1, a usual proportional odds model (POM) is applied to test mixted types of traits association:

$$\text{logit } \Pr(Y_d \leq c|G_i) = \alpha_c + \beta x_i,\ c = 1, \ldots, C-1, i = 0, 1, 2. \tag{6.1}$$

The genetic effect is coded by $(x_0, x_1, x_2) = (0, 1, 2)$. Under H_0 of no association with either trait, $\beta = 0$.

If two parameter as considered under H_0 of no association with either trait, $\beta_1 = \beta_2 = 0$. Under assumption that there is no POM between control and case group1 but having POM among cases. Thus, we will set up the model as follows.

$$\text{logit } \Pr(Y_d \leq 1|G_i) = \alpha_1 + \beta_1 x_i,\ i = 0, 1, 2. \tag{6.2}$$

$$\text{logit}\ \Pr(Y_d \leq c|G_i) = \alpha_c + \beta_2 x_i,\ c = 2, \ldots, C-1, i = 0, 1, 2. \tag{6.3}$$

The genetic effect is coded by $(x_0, x_1, x_2) = (0, 1, 2)$.

likelihood function of two parameters for the data in Table 4.1 is given by

$$L(\alpha_1, \ldots, \alpha_{C-1}, \beta) = \prod_{j=1}^{n} \left[\prod_{c=2}^{C} \{\Pr(Y_d = c|g_j)\}^{Y_{jc}} \right] \left[\{\Pr(Y_d = 1|g_j)\}^{Y_{j1}} \right]$$

$$= \prod_{i=0}^{2} \prod_{c=2}^{C} \{\Delta_1(\alpha_c + \beta x_i) - \Delta_1(\alpha_{c-1} + \beta x_i)\}^{r_{ic}} \prod_{i=0}^{2} \{\Delta_1(\alpha_c + \beta_2 x_i) - \Delta_1(\alpha_1 + \beta_1 x_i)\}^{r_{i1}}$$

where c=2,. . . , C-1, i=0,1,2.

where g_j takes G_0, G_1 or G_2. Denote $\Delta_1(\alpha_c + \beta x_i) = \Delta_{1ci}$ and similar notation for Δ_2. Denote $\alpha_0 = -\infty$, $\alpha_C = +\infty$, $\Delta_1(x) = \exp(x)/\{1 + \exp(x)\}$ and $\Delta_2(x) = \exp(x)/\{1 + \exp(x)\}^2$.

Then sorce function and fisher infromation will be derived based on this likelihood of two parameters.

References

Agresti, A.,(1984) *Analysis of Ordinal Categorical Data*. Wiley.

Agresti, A. (2002). *Categorical Data Analysis. Second Ed*. Wiley.

Amos, C.I., Chen, W.V., Seldin, M.F., Remmers, E.F., Taylor, K.E., Criswell L.A., Lee, A.T., Plenge, R.M., Kastner, D.L., and Gregersen, P.K. (2009). Data for genetic analysis workshop 16 problem 1, association analysis of rheumatoid arthritis data. *BMC Proc.* **3(Suppl 7)** S2.

Coenen, D., Verschueren, P., Westhovens, R., and Bossuyt, X. (2007). Technical and diagnostic performance of 6 assays for the measurement of citrullinated protein/peptide antibodies in the diagnosis of rheumatoid arthritis. *Clin. Chem.* **53** 498–504.

Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet. Epidemiol.* **32** 9–19.

Kost, J.T. and McDermott, M.P. (2002). Combining dependent P-values. *Stat. Probab. Lett.* **60** 183–190.

Huizinga, T.W., Amos, C.I., van der Helm-van Mil, A.H., Chen, W., van Gaalen, F.A., Jawaheer, D., Schreuder, G.M., Wener, M., Breedveld, F.C., Ahmad, N., Lum, R.F., de Vries, R.R., Gregersen, P.K., Toes, R.E., and Criswell, L.A. (2005). Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. *Arthritis Rheum.* **52** 3433–3438

Li, Q., Li, Z., Zheng, G., Gao, G., and Yu, K. (2013). Rank-based robust tests for quantitative trait genetic association studies. *Genet. Epidemiol.* **37**, 358–365.

Mjaavatten, M.D., van der Heijde, D., Uhlig, T., Haugen, A.J., Nygaard, H., Sidenvall, G., Helgetveit, K., and Kvien, T.K. (2010). The likelihood of persistent arthritis increases with the level of anti-citrullinated peptide antibody and immunoglobulin M rheumatoid factor: a longitudinal study of 376 patients with very early undifferentiated arthritis. *Arthritis Res. Therapy* **12** R76.

O'Reilly, P.F., Hoggart, C.J., Pomyen, Y., Calboli, F.C.F., Elliott, P., Jarvelin, M.-R., and Coin, L.J.M. (2012). MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE* **7(5)** e34861.

Sasieni, P. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.

Xing, C. and Xing, G. (2009). Power of selective genotyping in genome-wide association studies of quantitative traits. *BMC Proc.* **3(Suppl 7)** S23.

Yang, Q., Wu, H., Guo, C.Y., and Fox, C.S. (2010). Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet. Epidemiol.* **34** 444-454.

Zaykin, D.V., Westfall, P.H., Young, S.S, Karnoub, M.A., Wagner, M.J., and Ehm, M.G. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Hered.* **53** 79–91.

Zhang, H.P., Liu, C.-T., and Wang, X. (2010). An association test for multiple traits based on the generalized Kendalls tau. *J. Am. Assoc. Statist.* **105** 473–481.

Zhang, H., Wacholder, S., Qin, J., Hildesheim, A., and Yu, K. (2011). Improved genetic association tests for an ordinal outcome representing the disease progression process. *Genet. Epidemiol.* **35** 499–505.

Zheng, G., Joo, J., Zaykin, D., Wu, C.O., Geller, N.L. (2009). Robust tests in genome-wide scans under incomplete linkage disequilibrium. *Stat. Sci.* **24** 503–516.

Zheng, G., Wu, C.O., Kwak, M., Jiang, W., Joo, J., and Lima, J.A.C. (2012). Joint analysis of binary and quantitative traits with data sharing and outcome-dependent sampling. *Genet. Epidemiol.* **36** 263–273.

Zhu, W. and Zhang, H.P. (2009). Why do we test multiple traits in genetic association studies? (with discussion), *J. Korean Statist. Soc.* **38** 1-10.