NTNU – Trondheim
Norwegian University of
Science and Technology

Statistical Methods for Genetic Association Studies under the Extreme Phenotype Sampling Design

Modelling the Effects of both Common and
Rare Genetic Variants

Thea Bjørnland

Master of Science in Physics and Mathematics
Submission date: June 2014
Supervisor: Mette Langaas, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Preface

This Master's Thesis constitutes the course TMA4905 - Statistics for the Industrial Mathematics program at NTNU. The topic of this thesis, extreme phenotype sampling, evolved from the mandatory project of the course TMA4500, written in the autumn of 2013. The project I wrote in TMA4500 was focused on analysing the relationship between genetic variables and waist-hip ratio based on a dataset from HUNT (the Nord-Trøndelag Health Study). I have continued with the analysis of this dataset in Chapter 5 of this thesis, as well as in an ongoing project in collaboration with Ingrid Mostad at the Department of Clinical Nutrition, Trondheim University Hospital, and my supervisor Mette Langaas at the Department of Mathematics, NTNU.

I would like to thank my supervisor Mette Langaas for the advice, motivation and encouragement in the process of writing this thesis. I look forward to continuing our work together in my upcoming years as a Ph.D. student.

Abstract

In this thesis we investigate a concept in genetic association studies known as *extreme phenotype sampling* (EPS), where phenotype refers to physical appearances and in humans. In EPS studies, only individuals with extreme phenotypes are genotyped. Extreme phenotypes are typically defined as both ends of the spectrum of a continuously measurable trait such as weight or Body Mass Index (BMI). We introduce and develop statistical methods that apply to this design.

We investigate extreme phenotype sampling in both common and rare variant association analysis. For common variant association studies we will present methods that use the conditional model and the missing genotype model to test for genetic associations with disease. Both these methods were proposed in their simplest form by Huang & Lin (2007), and in this thesis we extend both methods to include any number of genetic and non-genetic covariates. We develop score test statistics for both these methods to test if there is an association between genetic variables and a phenotype. In order to evaluate these methods, we apply them to a dataset from the HUNT study (the Nord-Trøndelag health study) where we investigate the association between certain SNPs and waist-hip ratio.

For rare variant association studies we present five relevant methods for the cross-sectional design; (1) the collapsing method, (2) the CMC method, (3) the SKAT method, (4) the SKAT-O method, and (5) the β-SO (beta-smooth only) method. We adapt the CMC method developed by Li & Leal (2008) and the β-SO method by Fan et al. (2013), to the EPS design using the conditional model and corresponding score test. The collapsing method developed by Li & Leal (2008) and the kernel based methods developed by Wu et al. (2011) and Lee et al. (2012) have already been adapted to the extreme phenotype design based on the conditional model. We compare all five methods in an extensive simulation study. We use the software COSI to simulate rare variant genotype data.

In both common and rare variant studies we compare the cross-sectional design and the extreme phenotype sampling design. Extreme samples can theoretically be more powerful for detecting an association between a genetic variant and a phenotype, because the proportion of causal variants is enriched in extreme samples. This can be especially important for rare variant association studies. However, in this thesis we show that estimates made based on the conditional model are sensitive to violations of model assumptions in a greater degree than estimates based on a multiple linear regression model. Additionally, we show that if sample sizes are low or the proportion of causal variants included in the model is low, a random sampling method can be as powerful as an extreme phenotype sample to detect the associations.

Sammendrag

I denne oppgaven ser vi på genetiske assosiasjonsstudier der en kun analyserer individer med ekstreme fenotyper (fysisk utseende eller kjennetegn ved et menneske). Slike studier kalles *ekstrem fenotype utvalgsstudier* (extreme phenotype sampling, EPS). Ekstreme fenotyper er typisk definert som ytterpunktene i et spekter av verdier for en målbar variabel, for eksempel vekt eller BMI. Vi vil i denne oppgaven introdusere og utvikle statistiske metoder som kan brukes i EPS-studier.

Vi vil se på EPS-studier både for vanlige og sjeldne genetiske varianter. For analyser av vanlige genetiske varianter skal vi bruke to modeller; "conditional" og "missing genotype" modellene, begge utviklet av Huang & Lin (2007). Vi utvider disse modellene slik at de kan modellere fenotype som en funksjon av et vilkårlig antall genetiske og ikke-genetiske variabler. For å teste om det finnes en assosiasjon mellom fenotype og genotype, utvikler vi score test statistikker som passer til de ovennevnte modellene. Vi evaluerer kvalitetene til disse metodene ved å anvende dem på et datasett fra HUNT (Helseundersøkelsen i Nord-Trøndelag) der vi ser på forholdet mellom liv- og hofteomkretser (waist-hip ratio, WHR) som fenotype.

For analyser av sjeldne genetiske varianter vil vi presentere fem metoder som kan anvendes på tilfeldige utvalg fra en populasjon; (1) "the collapsing method", (2) "the CMC method", (3) "the SKAT method", (4) "the SKAT-O method", og (5) "the β-SO (beta-smooth only) method". Vi tilpasser CMC og β-SO metodene til bruk på EPS-studier ved å bruke conditional modellen og tilhørende score test. Collapsing metoden og SKAT metodene har allerede blitt tilpasset til dette designet av Li et al. (2011) og Barnett et al. (2013). Vi sammenligner de fem metodene i en simuleringsstudie basert på simulerte genetiske data fra programmet COSI.

Både for analyser av vanlige og av sjeldne genetiske varianter sammenligner vi resultater fra modeller tilpasset tilfeldige utvalg og EPS-utvalg. I teorien kan det å bruke personer med ekstreme fenotyper øke studiens styrke til å oppdage årsakssammenhenger mellom genetiske variabler og fenotyper fordi kausale genetiske varianter opptrer i høyere grad blant personer med ekstreme fenotyper. Til tross for dette ser vi i våre analyser at conditional modellen er sårbar for avvik fra modellantagelser i større grad enn multiple lineære regresjonsmodeller. Vi ser også at metoder anvendt på tilfeldige utvalg kan ha like stor styrke til å oppdage årsakssammenhenger som metoder anvendt på ekstreme utvalg, når utvalgsstørrelsen er lav og det er få genetiske varianter som har en biologisk effekt under alternativhypotesen.

Contents

List of abbreviations

β-SO	beta-smooth only method
BMI	Body Mass Index
CATT	Cochran-Armitage trend test
CDCV	Common disease common variant
CDRV	Common disease rare variant
CEU	HapMap population: Western European ancestry from the CEPH collection
CMC	Combined multivariate and collapsing method
COSI	Coalescent simulator
D	Diseased
D^c	Not diseased
E	Exposed
E^c	Not exposed
EPS	Extreme phenotype sampling
FTO	Fat mass and obesity associated gene
GWAS	Genome-wide association study
HUNT	Nord-Trøndelag health survey
LD	Linkage disequilibrium
MAF	Minor allele frequency
MC4R	Melanocortin 4 Receptor gene
MLR	Multiple linear regression model
OR	Odds ratio
SKAT	Sequence kernel association test
SKAT-O	Optimal sequence kernel association test
SNP	Single nucleotide polymorphism
TSI	HapMap population: Toscans in Italy
WHR	Waist-hip ratio

Chapter 1

Introduction

Genetic association studies are statistical studies of relationships between individuals' genotypes and phenotypes (diseases or appearances). The aim of such studies is to discover regions of the human genome that are related to a particular trait or disease. Genetic association studies often target common genetic variants, which are defined as variants that are common in a population. The well known genome-wide association studies (GWAS) are examples of such common variant association studies. Other types of studies focus on rare variants. These studies assume that there exists several rare genetic variants that can have similar effects on a particular phenotype. Different statistical methods have been developed for the common and rare variant association studies.

In this thesis we investigate a concept known as *extreme phenotype sampling* (EPS), or *selective genotyping*. In EPS studies, only individuals with extreme phenotypes are genotyped. Extreme phenotypes are typically both ends of the spectrum of a continuously measurable phenotype such as weight and Body Mass Index. Extreme phenotype sampling is based on the theory that individuals with extreme phenotypes provide more information about causal genetic variants compared to individuals with average phenotypes. Extreme phenotype sampling is relevant because the power to detect causal variants in an extreme sample of a certain size is greater that in a randomly sampled group of the same size Lee et al. (2012). Thus, the EPS design can lower the cost of genetic association studies without reducing power.

We will investigate extreme phenotype sampling and relevant statistical methods in both common and rare variant association analysis. We aim to develop score test statistics for testing association between genotypes and phenotypes in common variant EPS models that include other causal variables. We aim to investigate the effectiveness of extreme sampling as well as the application of our proposed models, in real data analysis using a dataset from HUNT. Concerning rare variants, we aim to review and compare rare variant association models in a simulation study. Additionally we aim to develop EPS models based on existing rare variant models and compare these to other rare variant EPS models in a simulation study. We will also compare the power of extreme sampling to random sampling in a simulation study.

1.1 Datasets

HUNT

HUNT (the Nord-Trøndelag health study, http://www.ntnu.edu/hunt) is a study of the population of Nord-Trøndelag that begun in the 1980s. HUNT3 is the third round of the HUNT study, and includes approximately 50 000 individuals. The HUNT3 study was performed between 2006 and 2008. The participants were asked to answer questionnaires, clinical surveys were performed, and blood tests taken and stored. The participants in the study are seen to be a good representation of trends in the Norwegian population. For some areas of research the trends could also be valid for other Caucasian populations. Langhammer et al. (2012) inform that there were some groups within the population that participated less than others. These groups consist of the sickest, young adults and people with low social status. The trends found from HUNT3 data should therefore be good estimates for the majority of the population, with some reservations.

Through a joint project with nutritionist Ingrid Mostad at the Department of Clinical Nutrition, Trondheim University Hospital, we have been granted permission to analyse the HUNT dataset used by Mostad et al. (2014) in the study of waist-hip ratio and dietary habits. Due to the high cost of genotyping all participants in this dataset, only genetic information on two SNPs (rs9939609 and rs17782313) among individuals with extreme phenotypes (waist-hip ratio) is available. This dataset is therefore a relevant dataset for investigating extreme phenotype sampling methods for common variants in a real-world setting.

COSI

In order to investigate existing and novel rare variant association methods, we simulate a dataset using the simulation software COSI, developed by Schaffner et al. (2005). The software is available at http://www.broadinstitute.org/∼sfs/cosi/. The COSI software simulates fictitious chromosomes of size 10 MB (megabase) in an out-of-Africa model in which it is assumed that a native African population emigrated to form separate populations (African, European, Asian). Mutations are simulated and their positions on the chromosome are reported. Additionally, the frequencies of the ancestral alleles and mutation alleles at these positions, in the current population, are reported. The benefit of the COSI simulations is that relationships between mutation sites are simulated relatively realistically so that methods that take into account relationships between SNPs and mutations along the chromosome can be evaluated. There are several available simulation softwares that create similar datasets, but as Wu et al. (2011) and Fan et al. (2013) evaluated their models using the European population generated by COSI, and since we will investigate and extend their methods, we have chosen to continue the use of COSI.

1.2 Structure of the thesis

In Chapter 2 we give an introduction to genetics and genetic association studies. We explain the basics of genetics and the two main theories of how common diseases are related to genotypes; the common disease common variant hypothesis, and the common disease rare variant hypothesis. Furthermore, we introduce mathematical notation of genetics. We also introduce the three designs that we will investigate further; the case-control design, the cross-sectional design, and the extreme phenotype sampling design. Finally, we discuss the epidemiological concept of confounding and how to deal with confounding in statistical models.

In Chapter 3 we outline the statistical models and methods that we will later adapt to genetic association studies. We introduce two types of statistical models; generalized linear models and functional linear models. We discuss the use of maximum likelihood to estimate parameters in these models. In this thesis we will focus on the score test as a means to test for association between genotypes and disease. This test is introduced in this chapter. We also outline other tests that have been used in the literature and that are used by us in data analysis or theoretical model development.

In Chapter 4 we describe the theory of common variant association studies. In this chapter we adapt the relevant methods from Chapter 3 to common variant association studies. We provide the score test statistics for the cross-sectional and the extreme phenotype sampling design. For the extreme phenotype sampling design we discuss two different models; the conditional model that only uses the extreme cases, and the missing genotype model that uses information on all individuals and considers genotypes as missing data for non-extreme subjects.

The HUNT dataset that was described above is a common variant dataset with two variants genotyped in extreme phenotype individuals. This dataset is analysed in Chapter 5, using methods from Chapter 4. Through this analysis we illuminate strengths and weaknesses of the extreme phenotype design and corresponding statistical models.

Chapter 6 contains an introduction to rare variant association studies. We discuss current methods of rare variant association methods and their adaptation to the extreme phenotype design. Some methods, such as SKAT, are established and verified for cross-sectional and extreme phenotype studies by simulation studies. A new method for rare variant association modelling that uses function linear models has proven to be more powerful than previous methods for the cross-sectional design. We adapt this method to the extreme phenotype design using the conditional model that was introduced in Chapter 3. We also discuss the so-called burden methods that are simpler, yet popular and often powerful.

In Chapter 7 we use a simulation study based on simulated genotypes from COSI to compare and verify the rare variant association models that were presented in Chapter 6. We compare the methods to each other in a cross-sectional design and an extreme phenotype sampling design. We also compare the power of the extreme phenotype design to the cross-sectional design using extreme and random samples from the same population. As the simulation studies by construction comply with the assumption of extreme phenotype sampling methods, we cannot assess their validity in real world studies, as was done for common variants in Chapter 5.

Conclusion, discussion of the results and an outline of further work that will be done on this topic is presented in Chapter 8.

Chapter 2

Introduction to genetics and genetic association studies

2.1 Cell biology and genetic inheritance

This introduction is, unless stated otherwise, based on Chapters 5 (”DNA and chromosomes”) and 19 (”Sex and genetics”) in the book by Alberts et al. (2010).

The laws of inheritance were first formulated by Gregor Mendel in the 19th century. Through experiments with different types of pea plants he learnt that so-called hereditary factors, today known as genes, govern traits of organisms. Differences in specific traits between organisms of the same species are caused by differences in genes among these organisms. Different versions of the same gene are known as *alleles*.

A cell's genetic information, and thereby an organism's genetic information, is stored in DNA (deoxyribonucleic acid). Roughly speaking, DNA is made up of two strands of sequences of specialized molecules known as *nucleotides*. The human DNA is composed of only four types of nucleotides; A, T, C and G. One DNA strand is built from approximately $3.2 \cdot 10^9$ nucleotides of these kinds. The two strands are paired with one another through hydrogen bonding between specific areas of the nucleotide molecules, and form the well-known double helix structure. The strands are connected by the binding of A and T, and C and G nucleotides, and we say that one strand is the complement of the other strand. This means that knowing the sequence of one strand automatically implies the sequence of the other strand. Differences between organisms is caused by differences in the nucleotide sequences. Between humans, these differences occur in about 0.1% of our nucleotide sequences.

Genes are generally defined as regions of DNA that code for a specific protein. Humans have approximately 25 000 genes, but these regions only cover a part of the total DNA. The remaining parts of the DNA is sometimes referred to as junk-DNA. The function of junk-DNA is not established as of today, although different hypothesis are being discussed. The term *genome* refers to an organism's complete set of DNA information.

DNA is packed and stored in the cell in structures known as *chromosomes*. Humans are sexually reproducing organisms and are therefore mainly *diploid*. This means that each cell in the human body consists of two sets of chromosomes; one set inherited from the mother and another from the father. An exception is the gametes, or germ cells (sperm in men and eggs in women). These cells are *haploid* and carry only one set of chromosomes. During sexual reproduction, a haploid germ cell and a haploid egg cell fuses together to form a diploid cell. The majority of cells in the human body are however not gametes, and generally known as somatic cells. For most of humans, the somatic cells contain 23 pairs of chromosomes, 22 of these are similar for both genders. The 23rd chromosome pair are the sex chromosomes which constitute the well-known XY pair in men and the XX pair in women.

In females, all chromosome pairs in the somatic cells are homologs, meaning that both chromosomes carry the same genes, but possibly different versions of that gene. In males, all chromosomes pairs except the sex chromosome are homologs. As mentioned, a gene come in different versions known as alleles. However, the term allele can also be used about a genetic region that is much smaller than the gene itself. Such regions are often termed *loci.* For a specific locus, one chromosome might carry one allele, while the other chromosome carries another allele. If at some position in the genome, the two chromosomes carry equal alleles, the person is said to be *homozygous* for the trait that this region codes for. If the person carries two different alleles for a gene, the person is said to be *heterozygous* for the trait. A person's collection of alleles is known as the *genotype.* Our *phenotypes*, or appearances and traits, depend on what types of alleles our genotype consists of.

Mendel discovered through his experiments that although some pea plants carried the genetic information for both the colours yellow and green, all plants grew up as yellow. This behaviour is caused by properties known as allelic dominance or recessiveness. For a given allele pair, one allele can be dominant and the other recessive, meaning that the phenotype that the dominant allele codes for will always appear. However, the offspring of the organism might inherit only recessive variants, and therefore express a different phenotype. Today we know of more complex models for allele properties which will be introduced later.

Mendel postulated that genes are inherited independently of each other during reproduction. This is today thought to be true for genes that lie on different chromosomes, or even genes on the same chromosome that are positioned far from each other. Genes or loci that lie closely together are however likely to be inherited as one unit, a phenomenon known as co-inheritance. We say that these genes or loci have a *genetic linkage.*

Single nucleotide polymorphisms

When two or more alleles of a certain gene or locus exist in a population, and all alleles have a population frequency of more than 1%, the collection of alleles for this locus is known as a *polymorphism* (Ziegler & König 2010, page 54). In certain polymorphisms, the alleles differ from each other in only one nucleotide. For example, the sequences A-A-T-C and A-T-T-C differ only at the A/T-alleles in the position of the second nucleotide. Such variations are known as *single nucleotide polymorphisms* (SNPs). Most commonly, a SNP is *biallelic*, meaning that only two different alleles exist. The *minor allele frequency* (MAF) refers to the population frequency of the allele in a polymorphism that occurs less often. Because loci are sub-regions of a gene, a gene can consist of several SNPs.

SNPs are generally not all independent and uncorrelated. We have explained that neighbouring genes are co-inherited, and so are neighbouring SNPs. We say that SNPs are linked in blocks and these blocks are called *haplotypes*. A haplotype is a region in a chromosome where all loci are inherited together. Due to genetic linkage it is sufficient to record the allele of one SNP in an individual's haplotype in order to state with a high degree of certainty what alleles the other SNPs in the haplotype will carry. For experimental genetics research it has therefore often been sufficient to genotype only a few SNPs in order to make claims about an entire gene or genetic region. These SNPs are often termed *tag SNP*s (Li & Leal 2008). Due to co-inheritance of genes across generations dating all the way back to our ancestors, only a few haplotypes are present in the human population. This enhances the effect and precision of tag SNPs. The use of genetic linkage information and tag SNPs greatly reduces the cost of experimental genetics research as only a fraction of SNPs need to be genotyped (Li & Leal 2008). The association between the variants in a haplotype is known as *linkage disequilibrium* (LD) (Li & Leal 2008).

Rare variants

While SNPs are classified as common variants due to the lower limit on their MAFs, there are genetic variants that have MAFs below the 1% frequency threshold. These are the so-called *rare variants*. It should be noted that some researchers use a different classification of variants. Luo et al. (2013) define rare variants as variants with MAF less than 1%, low frequency variants as variants with MAFs in the range of $1-5\%$ and common variants as SNPs with a MAF above 5%. The definition of common variants as SNPs with MAF ≥ 0.05 was solidified when the HapMap Project (Gibbs et al. 2003) chose to focus on SNPs with MAFs above 0.05. We will continue to use the 0.05 threshold between common and rare variants in this thesis.

2.2 Genetic association studies

Genetic association studies are studies that aim to discover a relationship between genetic variants and a disease or phenotype. The main focus has been common diseases, which Morgenthaler & Thilly (2006) define as "diseases that may afflict 1% or more of the population during a lifetime". Currently, two major theories concerning association between genetic variants and common diseases are dominating research. These are the *common disease, common variant* (CDCV) hypothesis and the *common disease, rare variant* (CDRV) hypothesis. In order to understand these theories and their importance in genetic association studies, some terms and concepts used in genetic association research will be explained.

The *prevalence* of a disease refers to the proportion of individuals in a population who have the disease at a specific time. The *penetrance* of a genetic variant is relative to a disease or trait of interest, and refers to the probability that a person carrying this variant has the disease or trait (Schork et al. 2009). *Allelic heterogeneity* refers to the situation where several genetic variants in the same region can affect a particular disease or phenotype. Low allelic heterogeneity support the CDCV hypothesis and corresponding methods for association detection. If a tag SNP is found to have an association with a phenotype, *one* of the SNPs in the corresponding haplotype block is assumed to be causal for the phenotype, if not the tag SNP itself. The CDRV hypothesis, on the other hand, assumes extreme allelic heterogeneity in regions of interest (Schork et al. 2009). It is in CDRV studies assumed that a gene that is associated with a disease is associated through *multiple* variants simultaneously. These associations may be of different strength and work in opposite directions.

Genetic association studies come in many forms and apply different statistical models and methods for analysis. In this section we will briefly explain the ideas behind common variant and rare variant association testing. In later chapters, a detailed description of statistical models and methods that can be applied in these studies will be given. We should note that although the CDCV and CDRV hypothesis appear as opposing theories, it is probable that a combination of the two is realistic (Li & Leal 2008).

Common variants

Most genetic association studies performed to date are based on the CDCV hypothesis. The argumentation behind this hypothesis is summarized by Alberts et al. (2010, page 681) in the following manner: "Because mutations that destroy the activity of a key gene are likely to have disastrous effects on the fitness of the mutant individual, they tend to be eliminated from the population by natural selection and so are rarely seen. Genetic variants that make for slight differences in a gene's function, on the other hand, are much more common". The idea is therefore that common diseases and phenotypes are likely to be caused by common variants (SNPs), and not rare variants, as mutations would not be rare if they did not cause great difficulties (rare diseases) for the carrier.

A well-known type of common variant genetic association study is the genome-wide association study (GWA study). In a typical GWA study, diseased and not diseased individuals are genotyped for a selection of tag SNPs along the genome, and statistical tests for homogeneity between the groups are used to determine whether the frequency of individuals with certain alleles are significantly different in the two groups. To date, hundreds of GWA studies have been performed, resulting in a good mapping of SNPs that are associated with common diseases. However, as stated by Schork et al. (2009), "more than $90 - 95\%$ of the heritable component of a disease has been left unexplained after extensive GWAS interrogation".

The HapMap project (Gibbs et al. 2003) is aimed to map linkage disequilibrium and thus provide suggestions for useful tag SNPs for different genes in several human populations, thus enabling further tag SNP and GWA studies.

Rare variants

The CDRV hypothesis has become more relevant due to the limitations of the common variant association studies. This idea is not concerned with the theory of one variant predisposing for one disease, but rather that a collection of rare variants, each with a relatively high penetrance, can be causal for a disease (Schork et al. 2009). This is in line with the assumption of extreme allelic heterogeneity, which according to Li & Leal (2008) implies that "a disease is caused collectively by multiple rare variants with moderate to high penetrances".

As with the HapMap project for common variant association testing, the 1000 Genomes project (www.1000genomes.org) was initiated to "facilitate the search for rare variants in different genes, if not the entire genome" (Schork et al. 2009).

Common experimental designs are often not suitable for rare variant association testing. For example, sampling diseased and healthy individuals who are to be genotyped and thereafter compared, becomes expensive due to required sample size. Due to the rare nature of the variants, large sample sizes are required in order to obtain enough information to discover an association. In addition, the use of tag SNPs is not advisable as its use is low-powered for rare variants (Li & Leal 2008).

Association vs. effect

The aim of a genetic association study is either to find whether an effect is present, try to quantify this effect, or a combination of both. Rare variant association studies are relatively new in research and have mainly been focusing on statistical methods for detecting an association. Common variant association studies are much more established and have focused both on detection and quantification.

2.3 Mathematical definitions for genetic models

There are aspects of a genetic model that are easily generalized by mathematical definitions. In the following we will introduce some of the most widely used concepts in genetics and statistics.

The odds ratio

A common measure of the severity of a disease between different groups of people is the odds ratio (OR). The odds ratio refers to the odds of being diseased under some exposure compared to the odds under another exposure, often referred to as exposed versus unexposed. A good example is the ratio of the odds for lung cancer among smokers (exposed) versus non-smokers (unexposed).

We define π_e as the probability of being diseased under some exposure. The odds for exposed individuals is then defined as

$$\text{odds}_e = \frac{\pi_e}{1 - \pi_e}.$$

If $\text{odds}_e > 1$, the probability of being diseased is greater than the probability of not being diseased, among exposed individuals. Let π_u be the probability of being diseased among unexposed individuals, and define odds_u in the same manner as above. The odds ratio is defined as

$$\text{OR} = \frac{\text{odds}_e}{\text{odds}_u} = \frac{\pi_e/(1 - \pi_e)}{\pi_u/(1 - \pi_u)}. \tag{2.1}$$

If $\text{OR} = 1$ then the exposure has no effect on development of disease. If $\text{OR} > 1$ we expect the disease to develop more often among exposed individuals.

One very important property of the odds ratio is that it is symmetric. In the above definition we discussed π_e, the probability of being diseased (D), when exposed (E). Thus $\pi_e = P(D|E)$ and $1 - \pi_e = P(D^c|E)$. Superscript c denotes the complement of an event.

Consider now the exposure as the random event. Then the odds of being exposed among the diseased is given by

$$\text{odds}_d = \frac{P(E|D)}{P(E^c|D)} = \frac{P(E \cap D)/P(D)}{P(E^c \cap D)/P(D)} = \frac{P(D|E)P(E)}{P(D|E^c)P(E^c)} = \frac{\pi_e}{\pi_u}\frac{P(E)}{P(E^c)},$$

and for nondiseased $\text{odds}_{d^c} = \frac{P(E|D^c)}{P(E^c|D^c)} = \frac{1-\pi_e}{1-\pi_u}\frac{P(E)}{P(E^c)}$. The odds ratio becomes

$$\text{OR} = \frac{\text{odds}_d}{\text{odds}_{d^c}} = \frac{\pi_e/\pi_u}{(1-\pi_e)/(1-\pi_u)} = \frac{\pi_e/(1-\pi_e)}{\pi_u/(1-\pi_u)},$$

which confirms the symmetric property of the odds ratio.

Genetic models

We will focus on modelling a phenotype Y as a function of exposures X and G. Here G represents genetic factors and X represents other possibly causative factors referred to as non-genetic factors.

We will hereafter consider biallelic loci. The two alleles of the loci will be referred to as a low-risk allele, denoted by a, and a high-risk allele, denoted by A. Risk refers to the probability of developing a particular disease or phenotype. An individual's genotype for a biallelic SNP is usually referred to as aa, aA or AA according to the alleles in the individual's two chromosomes. These genotypes will be indexed as 0, 1 and 2, according to the number of high-risk alleles, in the following.

We let p denote the population frequency of the a-allele. Consequently, the A-allele frequency q must be given by $q = 1 - p$. The minimum of these frequencies is the minor allele frequency (MAF). This frequency should should be above 1% for the alleles to classify as a SNP (Ziegler & König 2010, page 54). Genome-wide population studies such as the HapMap Project can be used as references for SNP MAFs in different populations.

Consider a population that is closed for immigration and where mating occurs and is successful independent of genotypes. Based on allele frequencies p and q, we can write down the genotype frequencies;

$$\begin{aligned} g_0 &= p^2, \\ g_1 &= pq + qp = 2pq, \text{ and} \\ g_2 &= q^2. \end{aligned} \tag{2.2}$$

For a given locus, the high-risk allele can be recessive, dominant or neither. If A is recessive, the probability of being diseased is the same for individuals with zero or one high-risk allele, and higher for individuals with two high-risk alleles. For dominant A, the probabilities are equal for individuals with genotypes aA and AA, and lower for individuals with zero high-risk alleles. As a third alternative, the probability of being diseased increases with the number of high-risk alleles in the genotype. Examples of such models are additive and multiplicative models.

Consider a population in which a biallelic loci exists, and that satisfies the assumptions made above. Let f_0, f_1 and f_2 represent frequencies of diseased individuals among individuals with genotypes aa, aA and AA, respectively. For a given genotype, f_i is therefore the conditional probability of being diseased and represents the penetrance of the genotype;

$$f_0 = P(D|aa), \quad f_1 = P(D|aA), \quad f_2 = P(D|AA).$$

Formally, *recessive* models are defined by;

$$f_0 = f_1 < f_2,$$

and *dominant* models are defined by;

$$f_0 < f_1 = f_2.$$

A *monotone* model is defined by

$$f_0 < f_1 < f_2,$$

and examples of such models include the *additive* model;

$$f_1 = (f_0 + f_2)/2,$$

and the *multiplicative* model;

$$f_1 = \sqrt{f_0 f_2}.$$

When modelling the relationship between genotype and disease it is common to code the genetic variable $g = (aa, aA, AA)$ as $(0, 1, 2)$ for a monotone model; $(0, 0, 1)$ for a recessive model; and $(0, 1, 1)$ for a dominant model.

2.4 Experimental designs

This section on study-design is based on Chapter 6 (”Types of Epidemiological Studies”) in the book by Rothman et al. (2008).

Epidemiologists separate between two major classes of designs; the experimental design and the non-experimental design. The experimental design is typical for the natural sciences where the scientist manipulates conditions and aims to estimate the effect these manipulations have on the observations. In epidemiology, such designs are often infeasible or unethical. For example, one cannot prescribe genotypes to individuals, and it can be unethical to give one set of individuals a drug that could cure a disease, and the other group a placebo. Epidemiologists and other medical researchers must therefore make use of non-experimental methods. Among such methods are the case-control design and the cross-sectional study. These studies are not randomized and systematic errors are likely to occur. Controlling for confounding is therefore particularly important. The theory of confounders will be explained later in this chapter.

We refer to a prospective design as a study where individuals are followed over time and disease occurrences during the period of follow-up is recorded. The retrospective design is a design where the exposure is measured at the present moment or based on individual records, and the disease status at the present time is recorded as the outcome. The major disadvantage to a good prospective study is financial and efficiency issues. It is expensive and time-consuming to follow large groups of people over time, especially for less common diseases. However, a prospective design can separate between short-term disease and long-term disease, and often estimate true frequencies directly. The retrospective design is often preferred to the prospective design due to lower costs and higher efficiency. A drawback of this design is that the only disease information available is the prevalence of the disease.

We will in this thesis focus on three different non-experimental retrospective designs for association studies; the case-control design, the cross-sectional design, and the extreme phenotype sampling design. These designs are chosen because the majority of studies in the literature follow these designs. For retrospective designs such as these, estimation of disease frequencies under different exposures must be performed carefully in accordance with the sampling procedure.

The case-control design

A case-control study is a retrospective study that evaluates binary outcomes. It is particularly useful in studies with a clear separation of diseased and not diseased. A case group is selected based on records of disease, and independent of the individual cases' exposure to causal events. A control group is sampled from the undiseased population. The control selection in this design is critical. According to Rothman et al. (2008, page 116) this sampling procedure must follow two basic rules; controls must be sampled from the source population from which the cases arose; and the controls must be sampled independent of exposure. The sample disease frequencies in such studies are by design much higher than in the general population. The odds ratio is a useful disease measure in case-control studies. Mathematically we partition the population into two sub-populations; diseased and not diseased individuals, so that the union of these sub-groups form the entire population. Thus, the controls in a case-control study must be sampled form the sub-population of not diseased individuals, while the cases must be sampled from the diseased sub-population.

The cross-sectional design

Rothman et al. (2008, page 97) define a cross-sectional study as "a study that includes as subjects all persons in the population at the time of ascertainment or a representative sample of all such persons, selected without regard to exposure or disease status". The disease or phenotype status in a cross-sectional study is measured at the same time as the exposure. The disease or phenotype can be analysed as a discrete or continuous variable.

The extreme phenotype sampling design

The extreme phenotype sampling (EPS) design began as a common variant association design under the name *selective genotyping.* The idea of this design is to only genotype individuals in the extreme ends of the phenotype spectrum. Huang & Lin (2007) claim that this design increases the power of association detection compared to random samples of the same size, and it can also reduce costs. An appropriate statistical model for this design was proposed by Huang & Lin (2007). If all assumptions are satisfied by the dataset, the estimates would be the same as those of other models applied to a cross-sectional design.

Although Huang & Lin (2007) show that the effect estimates based on their model are less biased than when other statistical models are applied to the extreme sample dataset, the model has been used mainly for testing and not for estimating effects.

The idea of genotyping individuals with extreme phenotypes has later been used in rare variant association testing, where the low frequencies of variants makes it important to maximize power. The EPS design and the model proposed by Huang & Lin (2007) has been explored in rare variant association studies by Li et al. (2011) and Barnett et al. (2013), among others.

2.5 Confounding in genetic association studies

As mentioned, the non-experimental designs are at risk of causing false conclusions due to lack of randomization of subjects and exposures. Perhaps there is an underlying cause for both the outcome and the exposure that causes a spurious association?

There are three epidemiological terms we should be aware of. A *confounder* is a common cause for both the exposure and the phenotype. If the confounder is not included in the statistical model, the effect of the exposure will be over- or underestimated. A *collider* is a common consequence of the exposure and the phenotype, while a *mediator* is a consequence of the exposure and cause of the phenotype. Including colliders or mediators in the model can introduce a bias in the effect estimate of the exposure. Epidemiologists investigate the effect of one exposure at the time, and only include confounders as additional covariates in the model. A fourth important concept in epidemiology is *effect measure modification*, also known as heterogeneity of effects. This heterogeneity becomes apparent if across strata of some exposure, the effects of another exposure vary.

In statistical analysis, the aim is often to create a model that explains as much of the variance in the data as possible. It is therefore of interest to include several exposures in the same model. We define a *nonconfounder* as an exposure that is a probable cause for the outcome, but not associated with the exposures already included in the model. Nonconfounders can be included without difficulties in a statistical model. Effect measure modification between exposures can be model by an interaction term in a statistical model and are therefore relatively easy to handle. Caution must be made with exposures that are thought to be colliders, mediators or confounders. Only expert biological or medical knowledge can determine the nature of such variables.

Population stratification

The term *population stratification* refers to subgroups of a population between which the allele frequencies differ. These subgroups are important confounders in genetic models as they can be a cause for a phenotype and a cause for genetic exposure. Consider two sub-populations with MAFs q and q' for some locus of interest. The population penetrances satisfy

$$\begin{aligned} f_0 &= f_1 = f_2 \\ f'_0 &= f'_1 = f'_2, \end{aligned}$$

such that there is no increased risk of disease depending on genotypes. Assume that due to cultural differences, one of the sub-populations is more prone to disease than the other.

If these groups are considered as one population, the increased risk would appear as a genetic effect although it is in fact the culture that is the risk factor.

Population stratification is an issue in most genetic association studies where subjects are related and form subgroups of allelic differences and family cultures. As it is often the case that relations between subjects are not known, population substructure must be estimated and controlled for by appropriate methods. Price et al. (2006) propose a novel method for controlling for population stratification in a case-control study by the use of principal components. One can also use principal component analysis (PCA) to control for population stratification in other types of studies. We note that the PCA approach assumes that for each individual in the study, genotypes of a relatively large number of loci are known, and that many of these loci are located on other chromosomes or far away from the locus one is testing.

Chapter 3

Statistical models and methods

We will in this chapter consider statistical modelling of both discrete and continuous data. We will deal with models with a dependent variable Y and covariates Z which can be separated into non-genetic (X) and genetic (G) exposures. We will consider the cases where Y is dichotomous (discrete with two levels), as well as Y continuous. We will present two modelling methods; generalized linear models and functional linear models, and explain how to estimate the parameters of these models by maximum likelihood methods. In addition, we will present tests for association between Y and Z in general, and Y and G in particular.

The models presented in this chapter are appropriate in ideal cases, where the sample sizes are large, and the number of parameters is relatively small. The models apply to randomly sampled data without substructures. These models are not necessarily directly applicable to the genetic association studies that we aim to investigate, but form the basis for more advanced methods that we will investigate and develop in this thesis.

3.1 Statistical models

We will in the following give an introduction to generalized linear models. We will explain how such models can fitted to datasets where Y is dichotomous or continuous. We will thereafter introduce the concept of functional linear models. Functional models are applicable when the covariates of the linear model have a temporal quality, such as having time or spacial qualities in addition to their observed values.

3.1.1 Generalized linear models

In the following, we use the notation of McCullagh & Nelder (1989) and assume that a random event Y is influenced by p covariates, $Z_1, \ldots, Z_p$. The different covariates may have a different number of attainable values. Let K be the number of unique combinations of the different levels of these exposures, and let $\mathbf{Y} = (Y_1, \ldots, Y_K)^T$ be the corresponding outcomes. Thus, Y_k is the predicted outcome under exposures $Z_{1k}, \ldots, Z_{pk}$, $k = 1, \ldots, K$. As an example, if we are investigating the effect of one covariate Z ($p = 1$) with three levels, we have $K = 3$ unique combinations. This yields the random vector $\mathbf{Y} = (Y_1, Y_1, Y_3)^T$, which elements are responses to the exposures Z_{11}, Z_{12} and Z_{13}, respectively.

According to McCullagh & Nelder (1989, page 27), generalized linear models consist of three components:

1. We assume that any observation $\mathbf{y} = (y_1, \dots, y_K)^T$ is a realization of a random variable $\mathbf{Y} = (Y_1, \dots, Y_K)^T$, which constitutes the *random component*. The components of $\mathbf{Y}$ are independent and follow distributions from some exponential family with possibly different parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)^T$. Let μ_k denote the expected value of each Y_k, i.e. $\mu_k = \mathrm{E}(Y_k)$, $k = 1, \dots, K$.

2. For each observed y_k, the corresponding p covariate levels $Z_{1k}, \dots, Z_{pk}$, of the p covariates, are known. The *systematic component* is a linear predictor $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)^T$ given by
$$\boldsymbol{\eta} = \sum_{j=1}^{p} \beta_j \mathbf{Z}_j, \tag{3.1}$$
where $\mathbf{Z}_j = (Z_{j1}, \dots, Z_{jK})^T$, such that $\eta_k = \sum_{j=1}^{p} \beta_j Z_{jk}$, $k = 1, \dots, K$. The coefficients $\beta_1, \dots, \beta_p$ are unknown and must be estimated.

3. Because the properties of the random component may not be directly reflected by the linear predictor $\boldsymbol{\eta}$, the *link function* $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ is introduced as a link between the mean of the random component and the systematic component.

Normal distributed random components

The bell-shaped normal distribution is widely applicable to data where Y is a continuous measurement of some natural phenomena. The generalized linear model with normally distributed random components is commonly known as a multiple linear regression model.

If the components of $\mathbf{Y}$ are assumed to follow a normal distribution with expected values μ_k, $k = 1, \dots, K$, and constant variance σ^2, we have the generalized linear model

$$\eta_k = \alpha + \sum_{j=1}^{p} \beta_j Z_{jk},$$

where α is some appropriate intercept. We have that $\mu_k = \mathrm{E}(Y_k) = \alpha + \sum_{j=1}^{p} \beta_j Z_{jk}$ such that the appropriate link function is simply the identity $g(\mu_k) = \mu_k$.

The multiple linear regression model is expressed as

$$Y = \alpha + \boldsymbol{\beta}^T \mathbf{Z} + \epsilon,$$

where ϵ follows a $N(0, \sigma^2)$ distribution. For any individual i with exposures $\mathbf{Z}_i = (Z_{1i}, \dots, Z_{pi})$, the outcome Y_i is predicted by

$$\hat{Y}_i = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{Z}_i,$$

where the estimated parameters $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ can be found by maximum likelihood estimation.

Binomial distributed random components

If we aim to fit a linear model to data where Y has two levels, i.e. is dichotomous, we can use a generalized linear model with binomial distributed random components.

A Bernoulli trial is an experiment where the response Z takes one of two possible values. A typical example is a coin toss where the outcome is either heads or tails. We often refer to one outcome as a success, and define the random variable γ as

$$\gamma = \begin{cases} 1 & \text{if the outcome is a success, and} \\ 0 & \text{otherwise.} \end{cases}$$

We define π, the probability of success, such that $P(\gamma = 1) = \pi$ and $P(\gamma = 0) = 1 - \pi$. The binomial distribution is defined as the distribution of the total number of successes in a series of independent and identically distributed Bernoulli trials. Letting Y denote the number of successes in n trials, the binomial distribution is defined by

$$P(Y = y|n, \pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}. \tag{3.2}$$

If the random component $\mathbf{Y}$ of a generalized linear model is assumed to follow a binomial distribution with parameters $\boldsymbol{\pi} = (\pi_1, n_1, \ldots, \pi_K, n_K)^T$, where $n_1, \ldots, n_K$ are known, we have the generalized linear model

$$\eta_k = \alpha + \sum_{j=1}^{p} \beta_j Z_{jk}. \tag{3.3a}$$

For the binomial distribution, we aim to model Y_k/n_k as opposed to Y_k. We have that $\mu_k = \mathrm{E}(Y_k/n_k) = \pi_k$. Since this linear model has no limits on the real line, and π_i should represent a probability, a link function $g(\pi_k) = \eta_k$ should map the real line into the interval $[0, 1]$. According to McCullagh & Nelder (1989, page 108) any such link function could be used. We will use the so-called *logit* link function

$$g(\pi_k) = \log\left(\frac{\pi_k}{1-\pi_k}\right). \tag{3.3b}$$

An important property of this link function is that it estimates the odds function that was introduced in the previous chapter. Using this link function enables us to estimate the same odds ratio in a case-control design as in a cross-sectional study. Other link functions that are suitable for the binomial distribution do not have this property. The type of model defined in Equation (3.3) is known as a *logistic regression model* (McCullagh & Nelder 1989, page 108).

3.1.2 Functional data analysis

Functional data analysis is a type of data analysis where no underlying theoretical probability distribution is assumed to describe the data. For a simple introduction, assume that data is observed in pairs (ξ_j, t_j), $j = 1, \ldots, n$. As explained by Ramsay & Silverman

(2005, page 38), we interpret ξ_j as a snapshot of a continuous and smooth function $z(t)$, taken at time point t_j. We note that t does not have to represent time, but can be any relevant continuum. By smoothness of z, Ramsay & Silverman (2005, page 38) refer to the existence of a sufficient number of derivatives.

For measurements in general there is often some noise disturbing the observation such that it cannot be assumed that $\xi_j = z(t_j)$ for all observed time points t_j. This is incorporated in the model of ξ_j by

$$\xi_j = z(t_j) + \epsilon_j,$$

where ϵ_j represents noise, and $z(t_j)$ is the value of the function z at time point t_j.

The goal of functional data analysis is according to Ramsay & Silverman (2005, page 38) to estimate the function $z(t)$ and some of its derivatives based on the discrete observations $\xi_1, \dots, \xi_n$. The noise in the data is handled by requiring the estimated function to be smooth, rather than trying to filter the noise from the data to make the observations themselves smooth. An important property of z is periodicity. If z is assumed to have a periodic behaviour, the estimate of z at the beginning of some interval should coincide with the estimate of z at the end of the interval, both with respect to the value of z itself, but also the derivatives. Such periodic functions can be functions that express changes over the four seasons. For non-periodic functions, this criterion does not have to be fulfilled.

Basis functions

The use of basis functions to estimate an unknown function is a tool that extends far beyond the theory of functional data analysis. However, we present basis functions for functional data analysis as we will use basis functions to estimate the function z from a discrete set of observations. Ramsay & Silverman (2005, page) define a system of basis functions as a "set of known functions ϕ_k that are mathematically independent of each other and have the property that we can approximate arbitrary well any function by taking a weighted sum or linear combination of a sufficiently large number K of these functions". In other words, we can approximate the function z by

$$\hat{z}(t) = \sum_{k=1}^{K} c_k \phi_k(t),$$

where c_k are appropriate constants and $\phi_k(t)$ are known basis functions. The choice of K determines the smoothness of the approximation. If we set $K = n$, the observed data $\xi_1, \dots, \xi_n$ will be represented exactly.

We will describe two basis systems that are of interest. These are suggested in the article by Fan et al. (2013) and in the book on functional data analysis by Ramsay & Silverman (2005).

The Fourier basis system

This system of basis functions is appropriate for periodic data without any strong local features (Ramsay & Silverman 2005, page 45). The system is defined as follows;

$$\begin{aligned}\phi_0(t) &= 1,\\ \phi_{2r-1}(t) &= \sin(r\omega t),\\ \phi_{2r}(t) &= \cos(r\omega t).\end{aligned}$$

The period that these functions reflect is given by $2\pi/\omega$ and the periodic behaviour of the data should be used to determine ω.

The B-spline basis system

Before we introduce the B-spine basis system we will give a short description of splines. This description is based on Ramsay & Silverman (2005, pages 46-49).

Consider an interval $[a, b]$ in which the data is observed. If the observed data is given in pairs (ξ_j, t_j) and t represents time, the interval will be a time interval starting at time a and ending at time b. This interval is separated into L subintervals by the values $\tau_1, \ldots, \tau_{L-1}$, called *breakpoints*. The endpoints of the interval are τ_0 and τ_L. In each interval $[\tau_l, \tau_{l+1}]$ a polynomial of order m is fitted to the data. These polynomials go by the name *splines*. We require smooth joining of the splines at the breakpoints such that function values and derivatives up to order $m-2$ must match (matching derivatives is not required for linear functions). The *spline function* estimates the entire function of interest over the interval $[a, b]$, and is the combination of the splines and the breakpoint continuity requirements. A spline function created by splines of order m and $L-1$ interior breakpoints is defined by $m+L-1$ parameters (Ramsay & Silverman 2005, page 49).

Let the B-spline basis system be represented by basis functions $\phi_k(t)$. Let each of these basis functions be spline functions of order m over the interval $[a, b]$ defined by the same breakpoints $\tau_1, \ldots, \tau_{L-1}$. Let each function be positive over no more than m intervals, and require these intervals to be adjacent. In addition, require smooth transitions to the zero regions. This is the so-called compact support property of the B-spline basis system and results in efficient computations (Ramsay & Silverman 2005, page 50). As a result of this property of the basis functions, there will be $m-1$ *knots* at each breakpoint, representing the values of the $m-1$ basis functions that are defined over each breakpoint. If breakpoints are equally spaced, the basis functions that are defined over the middle breakpoints will have the same shapes.

At the boundary points τ_0 and τ_L, $m-1$ knots are placed allowing for estimates of observed boundary behaviour of z, but disregarding smoothness of z outside the interval $[a, b]$. This forces the basis functions that are defined over outermost breakpoints to be defined over less than m breakpoints as well as less smooth transitions to zero at the boundaries. If $L-1$ interior breakpoints are defined, and spline functions of order m are chosen to define the basis functions, there will be a total of $m+L-1$ B-spline basis functions - one for each interior breakpoint, and $m/2$ for each of the two boundary points.

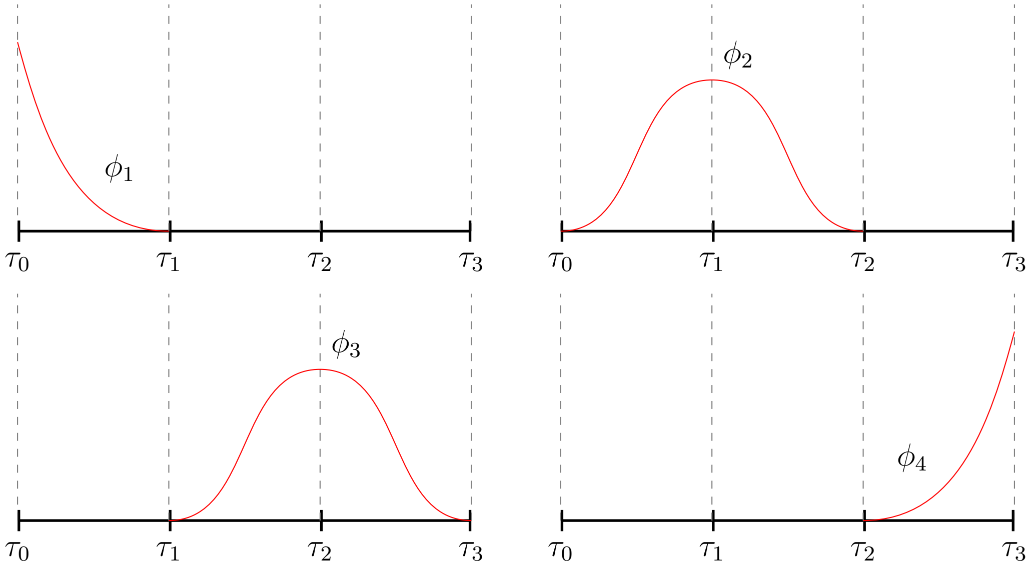


Figure 3.1: Illustration of B-spline basis functions

To illustrate, consider an interval divided into $L = 3$ smaller intervals. This yields $L - 1 = 2$ breakpoints. Assume we want to fit splines of order $m = 2$. This yields $m - 1 = 1$ knot per breakpoint and one knot at each endpoint. A total of $m + L - 1 = 4$ spline functions ($\phi_k, k = 1, \ldots, 4$) are placed in the interval, and at some point t at most two of these are non-zero. We have illustrated the division of the interval by equally spaced breakpoints, and the construction of four basis functions in Figure 3.1. Basis functions ϕ_2 and ϕ_3, which are defined over breakpoints only, have the same shapes. We see that for point $\tau_1 < t < \tau_2$, only basis functions ϕ_2 and ϕ_3 are non-zero.

Now that we have several basis functions defined over the interval $[a, b]$ we want to use these to estimate $z(t)$. We define this estimate in such a way that at some point t, a linear combination of all basis function is taken. We note that at most m of these basis functions are non-zero at point t. This yields

$$\hat{z}(t) = \sum_{k=1}^{m+L-1} c_k \phi_k(t),$$

which is dependent on the choice of breakpoints $\tau_1, \ldots, \tau_{L-1}$.

Determining the coefficients of basis functions

In Chapter 4 in the book by Ramsay & Silverman (2005), the use of least squares is discussed for determining the coefficients $\mathbf{c} = (c_1, \ldots, c_K)^T$ in the basis expansion of $z(t)$. A *simple linear smoother* is an estimation of the coefficients that is found by minimizing the least squares criterion

$$SSE = \sum_{j=1}^{n} \left(\xi_j - \sum_{k=1}^{K} c_k \phi_k(t_j) \right)^2 .$$

Let Φ be a $n \times K$ matrix containing all values $\phi_k(t_j)$ and define $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$. The result of minimizing the sum of squares criterion in order to obtain an estimate for $\mathbf{c}$ is, as stated by Ramsay & Silverman (2005, page 60), given by the well-known formula

$$\hat{\mathbf{c}} = (\Phi^T \Phi)^{-1} \Phi^T \boldsymbol{\xi}. \tag{3.4}$$

Functional linear models

Consider a linear model $Y = \alpha + \boldsymbol{\beta} \mathbf{Z}^T + \epsilon$. If either Y, $\mathbf{Z}$ or both are functional this is an example of a *functional linear model* (Ramsay & Silverman 2005, page 217). For this thesis we will only consider functional linear models where the response Y is scalar and one or more the covariates are functional. As described by Ramsay & Silverman (2005, page 219), such a model with one functional covariate is of the form

$$Y = \alpha + \int_t z(t)\beta(t)dt + \epsilon. \tag{3.5}$$

We note that a scalar response and functional covariate is assumed for all individuals. The intercept and the error term are similar to those of a generalized linear model, while the coefficient β is assumed to be a function of t. The function $\beta(t)$ is by Fan et al. (2013) referred to as the *effect function* of the covariate z as it reflects the effect of the covariate on Y at any point along t.

The covariate z_i is not completely observed for any individual due to the continuity of z. Rather, the observed data can be described by $(Y_{i,j}, \xi_{i,j}, t_{i,j})$ for $j = 1, \dots, n_i$ where n_i is the number of observations of individual i at different "time"-points of t. For a functional linear model it is of interest to estimate both $z(t)$ and $\beta(t)$. For $z(t)$, the method of basis functions can be used based on observations $(\xi_{i,j}, t_{i,j})$ as described above. For $\beta(t)$, the estimation can again be performed by basis functions, but one must take into account that β should reflect the relationship between z and Y.

We have expansions $z_i(t) = \sum_{k=1}^{K} c_{i,k}\phi_k(t) = \mathbf{c}^T \boldsymbol{\phi}(t)$, and $\beta(t) = \sum_{k=1}^{K} b_k \theta_k(t) = \mathbf{b}^T \boldsymbol{\theta}(t)$ for some appropriate systems of basis functions ϕ_k and θ_k. We estimate $c_{i,k}$ by the simple linear smoother as defined in Equation (3.4). Using a vector format we can now write the model as

$$\begin{aligned} Y &= \alpha + \int \hat{\mathbf{c}}^T \boldsymbol{\phi}(t) \boldsymbol{\theta}(t)^T \mathbf{b} dt + \epsilon \\ &= \alpha + \left(\int \hat{\mathbf{c}}^T \boldsymbol{\phi}(t) \boldsymbol{\theta}(t)^T dt \right) \mathbf{b} + \epsilon \\ &= \alpha + \mathbf{W}\mathbf{b} + \epsilon. \end{aligned}$$

This model can be treated as a standard linear regression model, for example such as described in the section on generalized linear models, with unknown coefficients α and $\mathbf{b}$, and variance parameter σ^2.

3.2 Fitting a statistical model

When we fit statistical models to data, our aim is to either estimate effects by estimating the parameters of the model, or to test for associations by ascertaining whether coefficients are non-zero. In this section we will focus on parameter estimation by likelihood theory. In the next section will deal with hypothesis testing.

3.2.1 Likelihood theory

Let the sample $\mathbf{Y}$ consist of *independent random variables* $Y_1, \ldots, Y_n$, with possibly different probability density functions $f_{Y_1}(y_1; \boldsymbol{\theta}), \ldots, f_{Y_n}(y_n; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector consisting of parameters $\theta_1, \ldots, \theta_p$, of which some are unknown. If all probability density functions $f_{Y_1}, \ldots, f_{Y_n}$ are equal, $\mathbf{Y}$ is called a *random sample* (Casella & Berger 2002, page 207), but this is not a necessity for the following results. Let the joint probability density function of $\mathbf{Y}$ be $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$. By independence of the random variables, $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_{Y_i}(y_i; \boldsymbol{\theta})$. Let $\mathbf{Y} = \mathbf{y}$ be an observed sample point.

The *likelihood function* is a function of $\boldsymbol{\theta}$, for a fixed observation $\mathbf{y}$. In other words, $\boldsymbol{\theta}$ is considered as the variable. The likelihood function is defined by

$$L(\boldsymbol{\theta}; \mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}). \tag{3.6a}$$

The likelihood function has the property that for a given $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the function reflects how likely it is to observe $\mathbf{Y} = \mathbf{y}$ under the distribution of $\mathbf{Y}$, $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}_0)$. The likelihood function for observation $\mathbf{y}$ can be expressed as

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i; \boldsymbol{\theta}). \tag{3.6b}$$

It is common to work with the so-called *log likelihood function*, which is defined by taking the natural logarithm of $L(\boldsymbol{\theta}; \mathbf{y})$. The log likelihood function for independent random variables is defined as

$$l(\boldsymbol{\theta}; \mathbf{y}) = \log(L(\boldsymbol{\theta}; \mathbf{y})) = \sum_{i=1}^{n} \log(f_{Y_i}(y_i; \boldsymbol{\theta})). \tag{3.7}$$

Maximum likelihood

The likelihood function is used to find the so-called *maximum likelihood estimates* (MLEs) of the unknown parameters. The parameters, if any, that are known, are simply treated as given.

Informally, to find the MLEs, we choose the parameter vector $\hat{\boldsymbol{\theta}}$, in the parameter space Θ, which makes our observation most likely. Formally, let $\hat{\boldsymbol{\theta}}(\mathbf{y})$ be the vector that satisfies

$$\hat{\boldsymbol{\theta}}(\mathbf{y}) = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{y}),$$

for the observed sample point $\mathbf{y}$. The parameters in $\hat{\boldsymbol{\theta}}(\mathbf{y})$ are the MLEs corresponding to this observation, and will be denoted by $\hat{\theta}_1, \ldots, \hat{\theta}_p$. By Casella & Berger (2002, page 316), we separate between a maximum likelihood *estimator* of $\boldsymbol{\theta}$, which is given by $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ for a sample $\mathbf{Y}$, and the maximum likelihood *estimate* of $\boldsymbol{\theta}$, which is given by $\hat{\boldsymbol{\theta}}(\mathbf{y})$ for the observed sample point $\mathbf{Y} = \mathbf{y}$.

The MLEs of the parameters are found by solving the system of equations $\frac{\partial L(\boldsymbol{\theta};\mathbf{y})}{\partial \theta_j} = 0$, for all $\{\theta_j : j \in \{1, \ldots, p\}$ and θ_j unknown$\}$. This is equivalent to maximizing the log likelihood function and solving

$$\frac{\partial \log L(\boldsymbol{\theta};\mathbf{y})}{\partial \theta_j} = \frac{\partial l(\boldsymbol{\theta};\mathbf{y})}{\partial \theta_j} = 0. \tag{3.8}$$

Likelihood functions for linear models

Consider now the theory of generalized linear models as described previously. Recall that we assumed K possible combinations of covariate levels and corresponding random vector $\mathbf{Y} = (Y_1, \ldots, Y_K)$. In data-analysis we usually have a dataset consisting of $n \gg K$ observations $y_1, \ldots, y_n$. We assume that each of these observations is a realization of one of the K random variables in the random vector $\mathbf{Y}$. In the dataset there will be K groups of observations. The observations in a group will be realizations of the same random variable Y_k. Let the number of observations in each such group be denoted n_k, where $k \in \{1, \ldots, K\}$. Further, let $\mathbf{y}_k$ denote the collections of observations for covariate levels $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{pk})^T$, corresponding to the random variable Y_k.

Consider a linear model $Y = \alpha + \boldsymbol{\beta}\mathbf{Z} + \epsilon$, where ϵ follows a $N(0, \sigma^2)$-distribution. Based on the n observations $\mathbf{y} = (y_1, \ldots, y_n)^T$, we can estimate the unknown parameters α, $\boldsymbol{\beta}$ and σ by maximum likelihood estimation. For the normal distribution, the log likelihood function corresponding to observation $\mathbf{y}$ is given by

$$\begin{aligned} l(\boldsymbol{\mu}, \sigma; \mathbf{y}) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\}\right) \\ &= -n\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2. \end{aligned} \tag{3.9}$$

Here, all n_k observations $y_i \in \mathbf{y}_k$ have corresponding parameters $\mu_i = \mu_k = \alpha + \boldsymbol{\beta}\mathbf{Z}_k$, $i = 1, \ldots, n_k$. For likelihood maximization of the linear model with normal distributed random components, the separation of variables into K groups is superfluous. MLEs are found by the differentiation procedure described in Equation (3.8).

For the vector of observations $\boldsymbol{\gamma}$ from a binomial distribution, the n components are binary values representing success or not success, under different covariate levels. Let y_k denote the number of successes among observations in $\boldsymbol{\gamma}$ with exposures $\mathbf{Z}_k$, $k \in \{1, \ldots, K\}$. We can thus create a vector $\mathbf{y} = (y_1, \ldots, y_K)^T$, that consists of observations from the random vector $\mathbf{Y} = (Y_1, \ldots, Y_K)^T$. Each component in $\mathbf{Y}$ follows a binomial distribution with parameters (π_k, n_k), $k = 1, \ldots, K$. The success probabilities π_k are expressed by Equation (3.3), and estimated by inserting MLEs for α and $\boldsymbol{\beta}$. These MLEs can be found by the Newton-Raphson algorithm (McCullagh & Nelder 1989, page 116), which maximizes the log likelihood function for the logistic regression model. This function is given by

$$
\begin{aligned}
l(\boldsymbol{\pi};\mathbf{y}) &= \sum_{k=1}^{K} \log\left(\binom{n_k}{y_k}\pi_k^{y_k}(1-\pi_k)^{n_k-y_k}\right)\\
&= \sum_{k=1}^{K} y_k\left(\frac{\pi_k}{1-\pi_k}\right) + n_k\log(1-\pi_k) + \log\binom{n_k}{y_k}, \qquad (3.10)
\end{aligned}
$$

where π_k is a function of α and $\boldsymbol{\beta}$, as defined in Equation (3.3).

Likelihood functions for missing data analysis

Missing data analysis is a large area of statistical analysis, some of which is relevant in EPS studies. In these studies, the phenotype is often known for a full sample, while the genotypes can be considered missing for a large portion of this sample. Extreme phenotype studies are nowadays mostly based on a model developed by Huang & Lin (2007) which we refer to as the conditional model. However, Huang & Lin (2007) also investigated a different model based on missing data methods. As we will investigate this model, we will briefly introduce some concepts of missing data analysis.

Ibrahim et al. (2005) present three classifications of missing data; missing completely at random (MCAR); missing at random (MAR) and nonignorable missing data. For our intents and purposes, MAR data are relevant. According to Ibrahim et al. (2005), "data are said to be MAR if, conditional on the observed data, the failure to observe a value does not depend on the data that are unobserved". For example, the response variable Y is completely observed but there are missing values of the covariate $\mathbf{Z}$. If the missingness is due to observed Y and not unobserved $\mathbf{Z}$, the data is MAR.

A complete response analysis is according to Ibrahim et al. (2008) an analysis of data where the response variable is completely observed while some of the covariates are missing at random. In missing data analysis, data is specified somewhat differently from standard regression and likelihood analysis. Instead of considering observations $Y_1, \ldots, Y_n$, where Y_i is a stochastic variable whose distributions depends on values of covariates $\mathbf{Z}_i$, we consider pairs of observations $(Y_i, \mathbf{Z}_i)$ which are both stochastic. Let $\boldsymbol{\alpha}$ be the parameters through which Y depends on $\mathbf{Z}$, let $\boldsymbol{\sigma}$ be nuisance parameters in the distribution of Y, and let $\boldsymbol{\zeta}$ be the parameters that describe the distribution of $\mathbf{Z}$.

The likelihood function based on the complete observations $(y_i, \mathbf{z}_i)$ of the random variables $(Y, \mathbf{Z})$, is defined as

$$
L_{Y,\mathbf{Z}}(\boldsymbol{\alpha},\boldsymbol{\sigma},\boldsymbol{\zeta}; y_1,\ldots,y_n,\mathbf{z}_1,\ldots,\mathbf{z}_n) = \prod_{i=1}^{n} f(y_i \cap \mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\sigma},\boldsymbol{\zeta}) = \prod_{i=1}^{n} f(y_i|\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\sigma}) f(\mathbf{z}_i;\boldsymbol{\zeta}).
$$

However, in complete response analysis the likelihood function is modified to account for the missing observations. Let $\mathcal{M}$ be the set of all individuals with missing covariates. For simplicity we will assume that all individuals are missing the same covariates. We split the variable $\mathbf{Z}_i$ into $\mathbf{Z}_i^{obs}$ for observed covariates, and $\mathbf{Z}^{mis}$ for missing covariates. Because we assume that all individuals are missing the same covariates, $\mathbf{Z}^{mis}$ is equal for all i. By

Ibrahim et al. (2008), the likelihood function for complete response analysis is defined as

$$L_{Y,\mathbf{Z}}(\boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\zeta}; y_1, \dots, y_n, \mathbf{z}_i : i \nsubseteq \mathcal{M}) = \\ \prod_{i:i \nsubseteq \mathcal{M}} f(y_i|\mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\sigma}) f(\mathbf{z}_i; \boldsymbol{\zeta}) \prod_{i:i \subseteq \mathcal{M}} \sum_{\mathbf{z}^{mis}} f(y_i|\mathbf{z}_i^{obs} \cap \mathbf{z}^{mis}; \boldsymbol{\alpha}, \boldsymbol{\sigma}) f(\mathbf{z}_i^{obs} \cap \mathbf{z}^{mis}; \boldsymbol{\zeta}), \tag{3.11}$$

if $\mathbf{Z}^{mis}$ is discrete, and

$$L_{Y,\mathbf{Z}}(\boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\zeta}; y_1, \dots, y_n, \mathbf{z}_i : i \nsubseteq \mathcal{M}) = \\ \prod_{i:i \nsubseteq \mathcal{M}} f(y_i|\mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\sigma}) f(\mathbf{z}_i; \boldsymbol{\zeta}) \prod_{i:i \subseteq \mathcal{M}} \int f(y_i|\mathbf{z}_i^{obs} \cap \mathbf{z}^{mis}; \boldsymbol{\alpha}, \boldsymbol{\sigma}) f(\mathbf{z}_i^{obs} \cap \mathbf{z}^{mis}; \boldsymbol{\zeta}) d\mathbf{z}^{mis},$$

if $\mathbf{Z}^{mis}$ is continuous.

The likelihood function can be maximized to find the maximum likelihood estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\zeta}$. If the distribution of the covariates is known, the parameter $\boldsymbol{\zeta}$ can be inserted.

3.2.2 Using R^2 to evaluate the fit of a multiple linear regression model

For multiple linear regression models, where the response variable is normally distributed, the R^2 value is a simple tool to evaluate the fit of a model to the data. *The coefficient of determination*, R^2, measures the quality of the fit of a linear regression model in that it reflects the "proportion of variability explained by the fitted model" (Walpole et al. 2007, page 407). Let y_i be observed values and $\hat{y}_i$ the estimated values of y_i. The coefficient is defined as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3.12}$$

where SSE is the error sum of squares, or residual sum of squares, and SST is the total corrected sum of squares. If SSE = 0 then all observations are perfectly reproduced and $R^2 = 1$. If SSE = SST then the model does not explain any of the variation in the data and $R^2 = 0$. We note that a high R^2 value can be obtained by overfitting the model to the data, as the value will increase when more variables are included in the model. To account for this, an adjusted R^2 value is often used;

$$R^2_{adj} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}, \tag{3.13}$$

where n is the number of observations and p is the number of variables included in the model.

3.3 Tests for association

Whenever a probability distribution can be assumed for the data, the score test is often a convenient tool for testing whether one or more parameters equal some null value. We will therefore focus on the score test in this thesis, but also consider testing differences between two groups with the χ^2 test, the Hotelling's T^2 test and the Cochran-Armitage trend test, which will be outlined in later sections.

3.3.1 The score test

Consider a sample $\mathbf{Y}$ of independent random variables described by the parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$. The score test is a large-sample test that can be used to test the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$.

The *score vector* for a parameter vector $\boldsymbol{\theta}_0$ is a p-dimensional vector defined by

$$\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y}) = \left.\frac{\partial l(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \tag{3.14}$$

Note that as $\mathbf{S}$ is evaluated in $\mathbf{Y}$, $\mathbf{S}$ must be a random variable. We can therefore approximate the distribution of $\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y})$ by the multivariate central limit theorem, if necessary assumptions are met. Under H_0 we have that

$$\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y}) = \left.\left(\frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{n} \log(f_{Y_i}(Y_i; \boldsymbol{\theta}))\right)\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \sum_{i=1}^{n} \left.\frac{\partial}{\partial \boldsymbol{\theta}} \log(f_{Y_i}(Y_i; \boldsymbol{\theta}))\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

This expression can be rewritten as a sum of independent random vectors $\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y}) = \mathbf{s}_1 + \cdots + \mathbf{s}_n$, where

$$\mathbf{s}_i = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \log(f_{Y_i}(Y_i; \boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \log(f_{Y_i}(Y_i; \boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \end{bmatrix}, \tag{3.15}$$

for $i = 1, \dots, n$. Due to the independence assumption on $\mathbf{Y}$, all random vectors $\mathbf{s}_i$ are independent with expected value $\boldsymbol{\mu}_i$ and covariance matrix Σ_i. By Casella & Berger (2002, page 336) we can write

$$\begin{aligned} \mathrm{E}(\mathbf{s}_{ij}) &= \mathrm{E}\left(\frac{\partial}{\partial \theta_j} \log(f_{Y_i}(Y_i; \boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) \\ &= \mathrm{E}\left(\frac{\frac{\partial}{\partial \theta_j} f_{Y_i}(Y_i; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}}{f_{Y_i}(Y_i; \boldsymbol{\theta}_0)}\right) \\ &= \int_{\mathcal{Y}} \frac{\frac{\partial}{\partial \theta_j} f_{Y_i}(y; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}}{f_{Y_i}(y; \boldsymbol{\theta}_0)} f_{Y_i}(y; \boldsymbol{\theta}_0)\mathrm{d}y \\ &= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta_j} f_{Y_i}(y; \boldsymbol{\theta}_0)\mathrm{d}y \\ &= \frac{\partial}{\partial \theta_j} \int_{\mathcal{Y}} f_{Y_i}(y; \boldsymbol{\theta}_0)\mathrm{d}y \\ &= 0. \end{aligned} \tag{3.16}$$

Thus $\mathrm{E}(\mathbf{s}_i) = \boldsymbol{\mu}_i = \mathbf{0}$, for all $i \in \{1, \dots, n\}$.

The covariance matrix simplifies significantly due to the zero mean, and is given by

$$\Sigma_i = \begin{bmatrix} \mathrm{Var}(\mathbf{s}_{i1}) & \cdots & \mathrm{Cov}(\mathbf{s}_{i1}, \mathbf{s}_{ip}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(\mathbf{s}_{ip}, \mathbf{s}_{i1}) & \cdots & \mathrm{Var}(\mathbf{s}_{ip}) \end{bmatrix} = \begin{bmatrix} \mathrm{E}(\mathbf{s}_{i1}^2) & \cdots & \mathrm{E}(\mathbf{s}_{i1}\mathbf{s}_{ip}) \\ \vdots & \ddots & \vdots \\ \mathrm{E}(\mathbf{s}_{ip}\mathbf{s}_{i1}) & \cdots & \mathrm{E}(\mathbf{s}_{ip}^2) \end{bmatrix}, \tag{3.17a}$$

for $i \in \{1, \ldots, n\}$.

In general we know from Casella & Berger (2002, page 338) that if a probability density function $f(y; \boldsymbol{\theta})$ belongs to an exponential family (normal, exponential, binomial, Poisson or gamma), we have that

$$\mathrm{E}\left(\left(\frac{\partial \log f(Y;\boldsymbol{\theta})}{\partial \theta_i}\frac{\partial \log f(Y;\boldsymbol{\theta})}{\partial \theta_j}\right)\right) = -\,\mathrm{E}\left(\frac{\partial^2 \log f(Y;\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right).$$

Assuming such a probability density function, we can write

$$\Sigma_i = \begin{bmatrix} -\,\mathrm{E}\left(\frac{\partial^2}{\partial \theta_1^2}\log f_{Y_i}(Y_i;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) & \cdots & -\,\mathrm{E}\left(\frac{\partial^2}{\partial \theta_1 \partial \theta_p}\log f_{Y_i}(Y_i;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) \\ \vdots & \ddots & \vdots \\ -\,\mathrm{E}\left(\frac{\partial^2}{\partial \theta_p \partial \theta_1}\log f_{Y_i}(Y_i;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) & \cdots & -\,\mathrm{E}\left(\frac{\partial^2}{\partial \theta_p^2}\log f_{Y_i}(Y_i;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) \end{bmatrix}. \tag{3.17b}$$

Thus, we know that the score vector $\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y})$ is a sum of i independent random variables with mean $\mathbf{0}$ and variances Σ_i. As a consequence, we find the expected value and covariance matrix of the score vector to be given by $\mathrm{E}(\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y})) = \mathbf{0}$, and $\mathrm{Var}(\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y})) = \sum_{i=1}^n \Sigma_i$.

If all Σ_i are equal to some matrix Σ, then the multivariate central limit theorem given in Johnson & Wichern (2007, page 176) is easily applied. By this theorem, the mean of the random vectors $\mathbf{s}_1, \ldots, \mathbf{s}_n$ has a $N(\mathbf{0}, \frac{1}{n}\Sigma)$-distribution. Therefore, the score vector $\mathbf{S}(\boldsymbol{\theta}_\mathbf{0}, \mathbf{Y})$ will be $N(\mathbf{0}, n\Sigma)$-distributed.

However, we have not assumed that the covariance matrices Σ_i are equal. According to Karr (1993, pages 190-196), if the sequence $\mathbf{s}_1 + \cdots + \mathbf{s}_n$ satisfies certain conditions, then by the central limit theorem, $\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y})$ converges in distribution to a $N(\mathbf{0}, \sum_{i=1}^n \Sigma_i)$-distribution. The conditions can be the so-called Lyapunov or Lindeberg conditions. According to Karr (1993, page 193), the Lindeberg condition is for example satisfied for a uniformly bounded sequence.

The *Fisher information matrix* corresponding to $\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y})$ is given by

$$I(\boldsymbol{\theta}_0) = -\begin{bmatrix} \mathrm{E}\left(\sum_{i=1}^n \frac{\partial^2}{\partial \theta_1^2}\log f_{Y_i}(Y_i;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) & \cdots & \mathrm{E}\left(\sum_{i=1}^n \frac{\partial^2}{\partial \theta_1 \partial \theta_p}\log f_{Y_i}(Y_i;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) \\ \vdots & \ddots & \vdots \\ \mathrm{E}\left(\sum_{i=1}^n \frac{\partial^2}{\partial \theta_p \partial \theta_1}\log f_{Y_i}(Y_i;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) & \cdots & \mathrm{E}\left(\sum_{i=1}^n \frac{\partial^2}{\partial \theta_p^2}\log f_{Y_i}(Y_i;\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) \end{bmatrix}$$

which is simply equal to $\sum_{i=1}^n \Sigma_i$. Therefore, by assuming necessary conditions, $\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y})$ has an approximate $N(\mathbf{0}, I(\boldsymbol{\theta}_0))$-distribution.

We are often interested in testing whether a subset of $\boldsymbol{\theta}$ equals some null value. By the notation of Smyth (2003), let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be a partition of $\boldsymbol{\theta}$, where $\boldsymbol{\theta}_2$ are the parameters to be tested, and $\boldsymbol{\theta}_1$ are unknown nuisance parameters. In the score test, nuisance parameters are replaced by their maximum likelihood estimates $\hat{\boldsymbol{\theta}}_1$. We write the null hypothesis as $H_0 : \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{2,0}$ and define $\boldsymbol{\theta}_0 = (\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_{2,0})$. We partition the score vector according to derivatives of nuisance parameters and parameters that are to be tested under H_0. Thereby, let

$$\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y}) = \begin{bmatrix} \mathbf{S}_1(\boldsymbol{\theta}_0, \mathbf{Y}) \\ \mathbf{S}_2(\boldsymbol{\theta}_0, \mathbf{Y}) \end{bmatrix},$$

where $\mathbf{S}_1 = \frac{\partial l}{\partial \boldsymbol{\theta}_1}$ and $\mathbf{S}_2 = \frac{\partial l}{\partial \boldsymbol{\theta}_2}$. We partition the mean $\boldsymbol{\mu}$ and variance $I(\boldsymbol{\theta}_0)$ of $\mathbf{S}(\boldsymbol{\theta}_0, \mathbf{Y})$ according to the partition of $\mathbf{S}$ so that

$$\begin{bmatrix} \mathbf{S}_1(\boldsymbol{\theta}_0, \mathbf{Y}) \\ \mathbf{S}_2(\boldsymbol{\theta}_0, \mathbf{Y}) \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} I(\boldsymbol{\theta}_0)_{11} & I(\boldsymbol{\theta}_0)_{12} \\ I(\boldsymbol{\theta}_0)_{21} & I(\boldsymbol{\theta}_0)_{22} \end{bmatrix} \right).$$

As we have seen, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$.

Because we replace nuisance parameters with maximum likelihood estimates, we have by definition that $\mathbf{S}_1(\boldsymbol{\theta}_0, \mathbf{Y}) = \frac{\partial l}{\partial \boldsymbol{\theta}_1}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0}$. We can now consider the distribution of $\mathbf{S}_2(\boldsymbol{\theta}_0, \mathbf{Y})$ conditioned on $\mathbf{S}_1(\boldsymbol{\theta}_0, \mathbf{Y}) = \mathbf{0}$. Johnson & Wichern (2007, pages 160-161) give a general result for the distribution of such conditional normal distributed random variables. Consequently, we have that the expected value of $\mathbf{S}_2(\boldsymbol{\theta}_0, \mathbf{Y})|(\mathbf{S}_1(\boldsymbol{\theta}_0, \mathbf{Y}) = \mathbf{0})$ is given by

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_2 + I(\boldsymbol{\theta}_0)_{21} I(\boldsymbol{\theta}_0)_{11}^{-1} (\mathbf{S}_1 - \boldsymbol{\mu}_1),$$

which yields $\boldsymbol{\mu}^* = \mathbf{0}$. Further, we have that the variance of $\mathbf{S}_2(\boldsymbol{\theta}_0, \mathbf{Y})|(\mathbf{S}_1(\boldsymbol{\theta}_0, \mathbf{Y}) = \mathbf{0})$ is given by

$$\Sigma^* = I(\boldsymbol{\theta}_0)_{22} - I(\boldsymbol{\theta}_0)_{21} I(\boldsymbol{\theta}_0)_{11}^{-1} I(\boldsymbol{\theta}_0)_{12}. \tag{3.18}$$

For simplicity of notation in future calculations, we denote $\mathbf{S}_2(\boldsymbol{\theta}_0, \mathbf{Y})|(\mathbf{S}_1(\boldsymbol{\theta}_0, \mathbf{Y}) = \mathbf{0})$ by $\mathbf{S}^*$. The *score test statistic* corresponding to $H_0 : \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{2,0}$ is given by

$$T^* = (\mathbf{S}^*)^T (\Sigma^*)^{-1} \mathbf{S}^*. \tag{3.19}$$

For any random variable $\boldsymbol{\xi}$ with a p-dimensional $N(\boldsymbol{\mu}, \Sigma)$-distribution, Johnson & Wichern (2007, page 163) show that the product $(\boldsymbol{\xi} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{\xi} - \boldsymbol{\mu})$ is χ^2-distributed with p degrees of freedom. We have shown that $\mathbf{S}^*$ follows an approximate multivariate normal distribution, and the dimension must be equal to the dimension of $\boldsymbol{\theta}_2$. As $\mathrm{E}(\mathbf{S}^*) = \mathbf{0}$, we see that T^*, as given in Equation (3.19), must be approximately χ^2-distributed with degrees of freedom equal to the dimension of $\boldsymbol{\theta}_2$.

3.3.2 The χ^2 test

The multinomial distribution arises when a series of categorical experiments are repeated n times. Each trial has one outcome from one of K categories. For each trial, the probability for a certain outcome k is given by π_k for $k = 1, \dots, K$, subject to $\sum_{k=1}^{K} p_k = 1$. After n trials have been performed, counts of the number of successes in each category are represented by the discrete random variables $Z_1, \dots, Z_K$ which follow a multinomial distribution. Clearly, $\sum_{k=1}^{K} Z_k = n$. The multinomial density function is according to Walpole et al. (2007, page 149) defined by

$$f_{Z_1,\dots,Z_K}(z_1, \dots, z_K; \pi_1, \dots, \pi_K, n) = \left(\begin{array}{c} n \\ z_1, \dots, z_K \end{array} \right) \pi_1^{z_1} \cdots \pi_K^{z_K},$$

subject to $\sum_{k=1}^{K} \pi_k = 1$ and $\sum_{k=1}^{K} z_k = n$.

The maximum likelihood estimators for the parameters of the multinomial distribution can be found using likelihood theory. The log likelihood function is given given by

$$l(\pi_1, \dots, \pi_K; Z_1, \dots, Z_K) = \log \begin{pmatrix} n \\ Z_1, \dots, Z_K \end{pmatrix} + \sum_{k=1}^{K} Z_k \log(\pi_k),$$

subject to $\sum_{k=1}^{K} \pi_k = 1$. This system can be expressed as one equation using a Lagrange multiplier such that

$$l(\pi_1, \dots, \pi_K; Z_1, \dots, Z_K) = \log \begin{pmatrix} n \\ Z_1, \dots, Z_K \end{pmatrix} + \sum_{k=1}^{K} Z_k \log(\pi_k) + \lambda(1 - \sum_{k=1}^{K} \pi_k),$$

for some $\lambda > 0$. By the procedure described in Equation (3.8), the MLEs based on observations $z_1, \dots, z_k$ are given by $\hat{\pi}_k = \frac{z_k}{\lambda}$. Applying the constraint $\sum_{k=1}^{K} \hat{\pi}_k = 1$ yields $\lambda = n$ and thus $\hat{\pi}_k = \frac{z_k}{n}$.

The χ^2 test can be used to test for homogeneity of two groups such as a case and a control group. Assume as above a series of experiments where one of K possible discrete outcomes are possible. Further, assume that these experiments are performed on subjects in a case and a control group with n_1 and n_2 subjects, respectively. For each group, the outcomes $Z_{1,1}, \dots, Z_{1,K}$ and $Z_{2,1}, \dots, Z_{2,K}$ are multinomially distributed with parameters $\pi_{1,1}, \dots, \pi_{1,K}$ and $\pi_{2,1}, \dots, \pi_{2,K}$. We define $N_k = Z_{1,k} + Z_{2,k}$, $k = 1, \dots, K$.

The null hypothesis in a homogeneity test states that the proportions of subjects in the case group with observed value k, are equal to the proportion in the control group. We can thereby state the null hypothesis as

$$H_0 : \pi_{1,k} = \pi_{2,k} = \pi_k, \text{ for all } k \in \{1, \dots, K\}.$$

Under the null hypothesis the two series of experiments can be considered as one series with $N = n_1 + n_2$ subjects. The maximum likelihood estimator of π_k is given by

$$\hat{\pi}_k = \frac{Z_{1,k} + Z_{2,k}}{N} = \frac{N_k}{N}.$$

As defined by Walpole et al. (2007, page 376), the test statistic for the null hypothesis is defined as

$$X^2 = \sum_{k=1}^{K} \left(\frac{Z_{1,k} - \hat{\pi}_k \cdot n_1}{\hat{\pi}_k \cdot n_1} \right)^2 + \left(\frac{Z_{2,k} - \hat{\pi}_k \cdot n_2}{\hat{\pi}_k \cdot n_2} \right)^2 .$$

Under the null hypothesis, X^2 is approximately χ^2-distributed with $K - 1$ degrees of freedom.

3.3.3 The Hotelling's T^2 test

Consider a genetic association study where the genotypes of M loci are to be compared between a case and control group. As an alternative to performing M separate χ^2 tests, the Hotelling's T^2 test gathers all information in one test. The following procedure is based on the theory in the book by Johnson & Wichern (2002, Chapter 5).

Let the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)^T$ contain the mean values of the M variables $\mathbf{Z} = (Z_1, \ldots, Z_M)$. The covariances of these variables are stored in the covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{Z}_i$ be an M-dimensional vector of observations for a sampled subject. Assume that n independent individuals are sampled, and let $\bar{\mathbf{Z}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i$. By the multivariate central limit theorem (Johnson & Wichern 2002, page 176), $\sqrt{n}(\bar{\mathbf{Z}} - \boldsymbol{\mu})$ is approximately multivariate normal distributed with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}$.

We will derive the Hotelling's T^2 statistic by considering the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$. In the one-dimensional case we can estimate σ^2/n, the variance of $\bar{Z}$, by s^2/n, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$. Under the null hypothesis, the well-known expression

$$t^2 = \frac{(\bar{Z} - \mu_0)^2}{s^2/n},$$

follows an approximate t-distribution with $n-1$ degrees of freedom (Johnson & Wichern 2002, page 211). This theory can be extended to the multivariate case. We define $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{Z}_i - \bar{\mathbf{Z}})(\mathbf{Z}_i - \bar{\mathbf{Z}})^T$ and use $\mathbf{S}/n$ as an estimate for the covariance matrix Σ. The Hotelling's T^2 statistic is defined as the multivariate analogy of the univariate t^2 test statistic. Thus,

$$T^2 = (\bar{\mathbf{Z}} - \boldsymbol{\mu}_0)^T \left(\frac{1}{n}\mathbf{S}\right)^{-1} (\bar{\mathbf{Z}} - \boldsymbol{\mu}_0).$$

According to Johnson & Wichern (2002, page 212), $\frac{n-M}{(n-1)M}T^2$ is, under the null hypothesis, approximately F-distributed with M and $n-M$ degrees of freedom.

We now consider the use of Hotelling's T^2 to test whether the distributions of M variables are equal for a case and control group. Consider samples of sizes n_1 and n_2. Let the two groups have equal variances $\boldsymbol{\Sigma}$ but possibly unequal mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. We want to test the null hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, which is equivalent to testing $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$. We recognize that this is a problem of the same nature as above, for some mean vector $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_0 = \mathbf{0}$.

To elaborate, we define a random variable $\bar{\mathbf{Z}} = \bar{\mathbf{Z}}_1 - \bar{\mathbf{Z}}_2$. By the central limit theorem, $\bar{\mathbf{Z}}_1$ is approximately $N_M(\boldsymbol{\mu}_1, \frac{1}{n_1}\boldsymbol{\Sigma})$-distributed, while $\bar{\mathbf{Z}}_2$ is approximately $N_M(\boldsymbol{\mu}_2, \frac{1}{n_2}\boldsymbol{\Sigma})$-distributed. By independence, the difference $\bar{\mathbf{Z}}_1 - \bar{\mathbf{Z}}_2$ is $N_M(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, (\frac{1}{n_1} + \frac{1}{n_2})\boldsymbol{\Sigma})$-distributed.

We estimate $\boldsymbol{\Sigma}$ separately for the case and control group, yielding $\mathbf{S}_1$ and $\mathbf{S}_2$. The pooled estimate for $\boldsymbol{\Sigma}$ is defined as $\mathbf{S} = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1+n_2-2}$ (Johnson & Wichern 2002, page 284). The T^2 statistic for the null hypothesis is given by

$$T^2 = (\bar{\mathbf{Z}}_1 - \bar{\mathbf{Z}}_2)^T \left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}\right)^{-1} (\bar{\mathbf{Z}}_1 - \bar{\mathbf{Z}}_2).$$

Under the null hypothesis, $\frac{n_1+n_2-M-1}{(n_1+n_2-2)M}T^2$ is approximately F-distributed with M and $n_1 + n_2 - M - 1$ degrees of freedom (Johnson & Wichern 2002, page 285).

3.3.4 The Cochran-Armitage trend test

The Cochran-Armitage trend test (CATT) is a test that allows us to test for an association between a dichotomous variable and an exposure with K levels. For example, the dichotomous variable can represent diseased and not diseased (case and control), while the exposure is a genotype with levels 0, 1 and 2 according to the number of high-risk alleles.

Recalling the definition of penetrance from Chapter 2, we can use the CATT test to test the hypothesis $H_0 : f_0 = f_1 = f_2$ where f represents the penetrance of the disease under the three genotypes. Generally, for K categories of any exposure, we define $f_k = P(D|k)$ and consider the null hypothesis

$$H_0 : f_0 = f_1 = \cdots = f_{K-1}.$$

Let n_1 and n_2 represent the number of subjects in the case and the control group, respectively. Additionally, let m_k denote the number of subjects with exposure level k. Define $m_k = x_k + y_k$, where x_k is the number of diseased subjects with exposure level k and y_k is the number of controls with exposure level k. The alternative hypothesis can be defined specifically for any trend thought appropriate, or generally in the sense that the trend is additive so that the probability of disease increases linearly with increased levels of exposure. For this general alternative hypothesis, we define scores, or weights, $s_1, \dots, s_K$ such that $s_k = \frac{k}{K-1}$, $k = 0, \dots, K-1$. The test statistic is defined as

$$\text{CATT} = \frac{\sum_{k=0}^{K-1} s_k(n_2 x_k - n_1 y_k)}{\sqrt{n_1 n_2 \left(\sum_{k=0}^{K-1} s_k^2 m_k - \frac{1}{n_1+n_2} \left(\sum_{k=0}^{K-1} s_k m_k \right)^2 \right)}}, \tag{3.20}$$

which is asymptotically standard normally distributed under the null hypothesis.

For genetic models we define the scores $(s_0, s_1, s_2) = (0, s, 1)$ with $s = 0, 0.5, 1$ for recessive, additive and dominant models under the alternative hypothesis, respectively (Langaas & Bakke 2013). Langaas & Bakke (2013) showed that if the genetic model under H_1 is unknown, it is good choice to use the scores $(0, 0.5, 1)$ as for an additive genetic model.

3.3.5 Correcting for multiple tests

The Hotelling's T^2 is a so-called multivariate test which considers several variables simultaneously. Thus, one test is used to test an hypothesis on several variables. The χ^2 and CATT tests are singlevariate tests which consider only one variable at a time. If these tests are to be used to test hypothesis on several variables, a correction must be made to account for the repeated tests on the same dataset. This is called multiple testing because there are multiple hypothesis to be tested using the same sample. Let M be the number of hypothesis to be tested. The term *familywise error rate* (FWER) refers to the probability of making at least one type one error among all M tests.

When performing one single test on the dataset, we say that we make a decision based on a significance level α. This means that the probability of making a type one error, i.e. rejecting the null hypothesis when the null hypothesis is true, should be less than or equal to α. The corresponding theory in multiple testing is to use the significance level α for the FWER. To obtain the desired FWER, the significance level for each test can be set to α/M. This so-called *Bonferroni* correction will yield a $(1-\alpha)\cdot 100\%$ confidence level to the overall procedure (Johnson & Wichern 2002, page 232).

Chapter 4

Statistical models and methods for common variant association studies

Genome-wide association (GWA) studies are focused on discovering loci along the genome in which one or more SNPs appear to be associated with a certain disease or phenotype. To reduce the information load, a locus can be represented by a single tag SNP which is in linkage disequilibrium with neighbouring SNPs. Thus, if the tag SNP is found to be associated with the phenotype, it is concluded that one of the SNPs in the locus are associated with and perhaps causal for, the phenotype. As the name suggests, a GWA study includes loci from positions along the entire genome. Once a region is found to be of interest, further studies are performed to ascertain the relationship between the tag SNP and the phenotype. We would like to highlight two particular GWA studies as examples; these are BMI (Body Mass Index) studies by Frayling et al. (2007) and Loos et al. (2008). These studies illustrate the process of a GWA study. Additionally, we will explain how to test for association between SNPs and phenotypes in a case-control study and a cross-sectional study. Furthermore, we will introduce the design known as extreme phenotype sampling (EPS), and explain how association testing can be performed by accounting for this specific sampling procedure.

An example: GWA studies on Body Mass Index

In the article by Frayling et al. (2007), we read that the relationship between variants of the FTO (fat mass and obesity associated) gene and BMI was discovered in a type 2 diabetes study. The study consisted of a genome-wide search for diabetes-associated SNPs in two groups of subjects; diabetes patients and people without diabetes. The SNPs that were found to have different genotype frequencies in the two groups of subjects were considered as associated with type 2 diabetes. Several SNPs along the FTO gene region on chromosome 16 were chosen by this criterion. Through further study, the researches found that the correlation with type 2 diabetes was caused by an underlying relationship between the SNPs and BMI. As there is a high correlation between the SNPs in the FTO region, it was considered sufficient to consider one of them, and Frayling et al. (2007) chose to do further research on the SNP rs9939609 because it had the highest genotyping success rate.

rs9939609 is a SNP located on the FTO gene. The two alleles of the SNP differ in A and T nucleotides, where A constitutes the ancestral version. It is believed that individuals with one or more copies of the A-allele are at higher risk of being overweight, than those with only T-alleles. By *The International HapMap Project* (2002-2009), the A-allele has a MAF of 0.46 in the CEU population (*"Utah residents with Northern and Western European ancestry from the CEPH collection"*) and a MAF of 0.47 in the TSI population (*"Toscans in Italy"*). These MAF estimates are taken from phase 2 of the HapMap project were 113 individuals from the CEU population and 88 individuals from the TSI population were sampled.

Loos et al. (2008) explain that after the discovery of the FTO gene's effect on BMI, a genome-wide study was performed with the aim of discovering other similar relations. Several regions were found to be of interest, but as rare mutations in the MC4R gene on chromosome 18 were known to cause human obesity, this region was considered a good candidate for further study. The particular SNP rs17782313 was chosen as a tag SNP based on similar observations as for rs9939609. rs17782313 is a SNP close to the Melanocortin 4 Receptor (MC4R) gene. The two alleles of the SNP differ in T and C nucleotides. The T-allele constitutes the ancestral gene. For this SNP, the C-allele is the high-risk allele. The MAF of the C-allele is 0.26 in the CEU population and 0.28 in the TSI population (*The International HapMap Project* 2002-2009).

Frayling et al. (2007) report that "individuals homozygous for the A allele at rs9939609 are at substantially increased risk of being overweight (OR = 1.38; 95%CI = 1.26 to 1.52; $p = 4 \cdot 10^{-11}$) or obese (OR = 1.67; 95%CI = 1.47 to 1.89; $p = 1 \cdot 10^{-14}$) compared with those homozygous for the T allele". Loos et al. (2008) report that "rs17782313 was associated with an $\sim$ 8% per-allele increase in the odds of being overweight (BMI $\geq$ 25kg m^{-2}; $\text{OR}_{\text{overweight}} = 1.08(1.05 - 1.11)$; $p = 1.6 \cdot 10^{-9}$) and $\sim$ 12% of being obese (BMI $\geq$ 30kg m^{-2}; $\text{OR}_{\text{obesity}} = 1.12(1.08 - 1.16)$; $p = 5.2 \cdot 10^{-9}$)". Both these studies work with a case-control design. However, we must question the results where they use obese patients as a case group and normal weight patients as a control group, leaving the overweight patients out of the equation. As BMI is a continuous measurement, the sampling of normal weight and obese seems to resemble the EPS design rather than the case-control design. No comments on this practice are reported in the articles.

In Chapter 5, we will present the analysis of a dataset where waist-hip ratio (WHR) is evaluated for association with the tag SNPs rs9939609 and rs17782313. The dataset that is available to us consists of individuals from the upper and lower quartiles of the WHR spectrum in the HUNT3 population. This sampling of individuals for genotyping based on the 25% most extreme phenotypes in opposite ends of the spectrum is an example of the EPS design. We will describe the models and methods that can be used to handle such datasets later in this chapter. First however, we will describe methods of association analysis in case-control and cross-sectional studies with dichotomized and continuous outcomes.

4.1 Association analysis, binary outcome

A binary outcome in genetic association studies is usually a separation between diseased and not diseased. We will consider cross-sectional studies were subjects are randomly sampled independent of disease status and exposures, as well as case-control studies where subjects are sampled independent of exposures, but dependent on disease status.

Estimating effect

Binary data can be modelled by a generalized linear model with binomial distributed random components. The aim is usually to test whether the odds ratio between cases and controls is one. We have seen that Frayling et al. (2007) reported odds ratios between homozygotes for the high-risk allele and homozygotes for the low-risk allele. Loos et al. (2008), on the other hand, reported odds ratios between individuals with j and $j+1$ high-risk alleles, $j = 0, 1$.

If a random sample of sufficient size is drawn from the population, the true disease frequency is reflected in the sample. We denote non-genetic exposures by $\mathbf{X}_i$ and genetic exposures by $\mathbf{G}_i$. The relationship between the success parameter π_i, representing the probability of disease for individual i, and exposures $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{G}_i)$ can be modelled by

$$\operatorname{logit}(\pi_i) = \alpha_0 + \boldsymbol{\alpha}^T\mathbf{X}_i + \boldsymbol{\beta}^T\mathbf{G_i}, \tag{4.1}$$

and maximum likelihood estimators can be found by maximizing the log likelihood function in Equation (3.10). The odds ratio between individuals with genotypes $\mathbf{g}_i$ and $\mathbf{g}_j$ is estimated by $\hat{\text{OR}} = \exp(\hat{\boldsymbol{\beta}}^T(\mathbf{g}_i - \mathbf{g}_j))$.

Recall that we should always include confounders in a regression model. We defined non-confounders are covariates that are neither confounders, colliders or mediators, and have individual effects on the dependent variable. In multiple linear regression models, we include as many non-confounders as possible in order to estimate effects accurately. Concerning logistic regression, Robinson & Jewell (1991) show that including non-confounders reduces precision of estimates. Based on these observations, we exclude any non-confounders from our logistic regression models.

In a case-control study, the disease frequency is much higher in the dataset than in the population. However, we shall see that a logistic regression model fitted to data from a case-control study yields the same parameter $\boldsymbol{\beta}$ and thus the same odds ratio, as in a cross-sectional study. The two models differ only in the intercept parameter α_0.

The logistic regression model applied in a cross-sectional study directly estimates $\operatorname{logit}(\pi_i)$, where π_i is the probability of being diseased under exposure $\mathbf{Z}_i$. For the case-control study, the logistic regression model estimates $\operatorname{logit}(\rho_i)$, where ρ_i is the probability of being diseased under exposure $\mathbf{Z}_i$, given that the individual was sampled for the study. We can find a simple relationship between π_i and ρ_i. Let S denote the event *sampled*, and S^c the event *not sampled*. Further, let D denote the event *diseased* and D^c denote *not diseased*. We define the sampling frequencies of cases and controls by $\tau_0 = P(S|D)$ and $\tau_1 = P(S|D^c)$, respectively. By definition

$$\rho_i = P(D|S; \mathbf{Z}_i = \mathbf{z}_i) = \frac{P(D \cap S; \mathbf{z}_i)}{P(S; \mathbf{z}_i)},$$

where $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{g}_i)$. By applying Bayes' rule we see that

$$\rho_i = \frac{P(S|D; \mathbf{z}_i)P(D; \mathbf{z}_i)}{P(S; \mathbf{z}_i)} = \frac{P(S|D; \mathbf{z}_i)P(D; \mathbf{z}_i)}{P(S|D; \mathbf{z}_i)P(D; \mathbf{z}_i) + P(S|D^c; \mathbf{z}_i)P(D^c; \mathbf{z}_i)}.$$

Note that in a case-control study, the sampling is independent of exposure, as the exposure is unknown at the time of sampling. It follows that

$$\rho_i = \frac{P(S|D)P(D;\mathbf{z}_i)}{P(S|D)P(D;\mathbf{z}_i) + P(S|D^c)P(D^c;\mathbf{z}_i)} = \frac{\tau_0\pi_i}{\tau_0\pi_i + \tau_1(1-\pi_i)}, \tag{4.2}$$

by the definitions of π_i, τ_0 and τ_1. We can replace π_i by its form in Equation (4.1) to get

$$\begin{aligned}\rho_i &= \frac{\tau_0 \exp\left(\alpha_0 + \boldsymbol{\alpha}^T\mathbf{x}_i + \boldsymbol{\beta}^T\mathbf{g}_i\right)/(1+\exp\left(\alpha_0 + \boldsymbol{\alpha}^T\mathbf{x}_i + \boldsymbol{\beta}^T\mathbf{g}_i\right))}{\left(\tau_0 \exp\left(\alpha_0 + \boldsymbol{\alpha}^T\mathbf{x}_i + \boldsymbol{\beta}^T\mathbf{g}_i\right) + \tau_1\right)/(1+\exp\left(\alpha_0 + \boldsymbol{\alpha}^T\mathbf{x}_i + \boldsymbol{\beta}^T\mathbf{g}_i\right))} \\ &= \frac{\exp\left(\alpha_0^* + \boldsymbol{\alpha}^T\mathbf{x}_i + \boldsymbol{\beta}^T\mathbf{g}_i\right)}{1+\exp\left(\alpha_0^* + \boldsymbol{\alpha}^T\mathbf{x}_i + \boldsymbol{\beta}^T\mathbf{g}_i\right)},\end{aligned}$$

where $\alpha_0^* = \alpha_0 + \log(\tau_0/\tau_1)$. Thus we see that the population parameter ρ_i in a retrospective case-control study equals the population parameter π_i of the cross-sectional study, except for the intercept coefficient α_0^*. The odds ratio between individuals with genotypes $\mathbf{g}_i$ and $\mathbf{g}_j$ is given by $\mathrm{OR} = \exp(\boldsymbol{\beta}^T(\mathbf{g}_i - \mathbf{g}_j))$, for both studies. The estimated odds ratio for the case-control study should therefore be similar to that of the cross-sectional study. This illustrates the motivation behind the use of the logit link function in the generalized linear model in Equation (3.3). Other link functions that are suitable for the binomial distribution do not have this property.

The null hypothesis for association can be written as $H_0 : \boldsymbol{\beta} = \mathbf{0}$, corresponding to the model in Equation (4.1). Setting $\boldsymbol{\theta}_1 = (\alpha_0, \boldsymbol{\alpha})$ and $\boldsymbol{\theta}_2 = \boldsymbol{\beta}$, the hypothesis $H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$ can be tested by the score test, using $\boldsymbol{\theta}_0 = ((\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}), \mathbf{0})$. This corresponds to testing $H_0 : \mathrm{OR} = 1$ for all genotypes. Perhaps more common for case-control studies is to compare genotype frequencies between cases and controls. This can be done for one tag SNP at a time by a χ^2 or CATT test, or for several SNPs simultaneously with a Hotelling's T^2 test.

4.2 Association analysis, continuous outcome

Many phenotypes can be measured continuously, such as BMI and WHR. We assume in the following that Y is a continuous phenotype measurement, or an appropriate transformation of the phenotype, such that Y can be assumed to follow a normal distribution in the population. We will consider a model of the form

$$Y = \alpha_0 + \boldsymbol{\alpha}^T\mathbf{X} + \boldsymbol{\beta}^T\mathbf{G} + \epsilon. \tag{4.3}$$

We assume that the covariate vector $\mathbf{X}$ is p_x-dimensional, while $\mathbf{G}$ is p_g-dimensional. As usual, $\mathbf{X}$ contains non-genetic covariates while $\mathbf{G}$ contains genetic covariates. Further, ϵ follows a $N(0,\sigma^2)$-distribution. We will show how to test the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$, i.e. that none of the genetic variants have an effect on phenotype, in different types of studies. Notice that the null hypothesis makes no assumption on $\boldsymbol{\alpha}$. If only a subset of the genetic variants are of interest for the null hypothesis, the rest of the genetic variables could be moved to the $\mathbf{X}$ covariate. If non-genetic variables are of interest in the null hypothesis, similar changes could be made in the other direction. For simplicity we will however focus on the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$.

We will let y_i and $\mathbf{Z}_i$ represent observed phenotypes and covariates for individual i, and $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{G}_i)$.

4.2.1 The cross-sectional design

In the cross-sectional design, the observed data $y_1, \dots, y_n$ represent a realization of a random sample from a population defined by Equation (4.3). The linear regression model we will consider is a generalized linear model with normally distributed random components. The log likelihood function is given by Equation (3.9) and

$$l(\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma; \mathbf{y}) = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{G}_i)^2,$$

where $\mathbf{X}_i$ and $\mathbf{G}_i$ are considered known for all individuals.

All parameters, α_0, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and σ can be estimated using the maximum likelihood procedure based on the likelihood function. The covariate coefficients have simple closed-form expressions;

$$\begin{bmatrix} \hat{\alpha}_0 \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{y}.$$

The maximum likelihood estimate for σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha}_0 - \hat{\boldsymbol{\alpha}}^T \mathbf{X}_i - \hat{\boldsymbol{\beta}}^T \mathbf{G}_i)^2.$$

Let $\hat{\alpha}_0$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}$ be the maximum likelihood estimators of the parameters in the null model

$$Y = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X} + \epsilon.$$

In line with previous notation we define $\boldsymbol{\theta}_1 = (\alpha_0, \boldsymbol{\alpha}, \sigma)$ while $\boldsymbol{\theta}_2 = \boldsymbol{\beta}$ so that $\boldsymbol{\theta}_0 = (\hat{\boldsymbol{\theta}}_1, \mathbf{0})$. Consequently, the score vector $\mathbf{S}^*$ as defined in Section 3.3.1 is given by

$$\mathbf{S}^* = \left. \frac{\partial l}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0} = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (Y_i - \hat{\alpha}_0 - \hat{\boldsymbol{\alpha}}^T \mathbf{X}_i) \mathbf{G}_i.$$

The covariance matrix Σ^* is found by Equation (3.18) which uses the components of the information matrix evaluated in $\boldsymbol{\theta}_0$. In order to express the information matrix we need the second derivatives of the log likelihood function. These are written out in full in Appendix B.1.1. The information matrix is found by taking the negative expectation of these derivatives evaluated in $\boldsymbol{\theta}_0$. We use the fact that $\mathrm{E}(Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i) = 0$, while $\mathrm{E}((Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i)^2) = \sigma^2$ in the null model. The components of the Fisher information matrix are given by

$$\begin{aligned}
I(\boldsymbol{\theta}_0)_{11} &= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} n & n\bar{\mathbf{X}}^T & 0 \\ n\bar{\mathbf{X}} & \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T & \mathbf{0}_{p_x\times 1} \\ 0 & \mathbf{0}_{1\times p_x} & 2n \end{bmatrix}, \\
I(\boldsymbol{\theta}_0)_{21} &= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} n\bar{\mathbf{G}} & \sum_{i=1}^n \mathbf{G}_i\mathbf{X}_i^T & \mathbf{0}_{p_g\times 1} \end{bmatrix}, \\
I(\boldsymbol{\theta}_0)_{12} &= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} n\bar{\mathbf{G}}^T \\ \sum_{i=1}^n \mathbf{X}_i\mathbf{G}_i^T \\ \mathbf{0}_{1\times p_g} \end{bmatrix}, \\
I(\boldsymbol{\theta}_0)_{22} &= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} \sum_{i=1}^n \mathbf{G}_i\mathbf{G}_i^T \end{bmatrix},
\end{aligned}$$

where $\mathbf{0}_{N\times M}$ is a $N \times M$ matrix with all entries equal to zero. The vectors $\bar{\mathbf{X}}$ and $\bar{\mathbf{G}}$ are averages computed across all subjects in the sample.

The above components are used to find the covariance matrix Σ^* according to Equation (3.18). Finally the score test statistic T^* is found by the expression in Equation (3.19); $T^* = (\mathbf{S}^*)^T(\Sigma^*)^{-1}\mathbf{S}^*$, and is asymptotically χ^2-distributed with p_g degrees of freedom.

4.2.2 Extreme phenotype sampling

We have seen studies where individuals selected according to a continuous phenotype measurements were sampled based on high or low measured phenotype values which were later dichotomized so as to fit the case-control format. In our previous example, Frayling et al. (2007) and Loos et al. (2008) considered individuals with a BMI classification of obese as the case group and those with the classification normal weight as the control group, neglecting the in-between group of overweight individuals. In a case-control study where obese patients are cases, the control group should ideally be a random sample from all individuals with BMI level below obese. In Appendix A, we give a theoretical explanation to why a case-control scenario is not appropriate in EPS studies where the aim is to estimate the odds ratio. The estimated ratio is not the same as the odds ratio in a cross-sectional study, and interpreting the result is difficult. Yet, it is always valid to use a test for homogeneity to test for a potential difference between the two BMI groups. However, extreme phenotype sampling is by definition dependent on a continuous outcome and so it seems inefficient to dichotomize the observations.

As discussed in Chapter 2, Huang & Lin (2007) introduced a model for a design which they referred to as selective genotyping, which in later research has been referred to as extreme phenotype sampling. The idea is roughly to sample from the extreme ends of the phenotype spectrum in order to increase the frequency of causal variants in the sample. Huang & Lin (2007) worked with common variants with MAFs of 0.05 and 0.2. The novelty of the work by Huang & Lin (2007) was to derive the distribution of the extreme phenotypes, given lower and upper cut-off values c_l and c_u. We will refer to the model proposed by Huang & Lin (2007) as the *conditional model* as it conditions on the sampling procedure. We will however, extend this model by allowing for non-genetic effects to be included as nuisance variables.

In many EPS studies, the phenotypes and non-genetic covariates are known for the full sample that contains phenotype values from the whole spectrum, not only the extreme ends. Huang & Lin (2007) proposed, in addition to the conditional model, a model that incorporates known phenotypes for a large number of individuals whose genotypes are treated as missing data. Based on simulation studies, Huang & Lin (2007) concluded that this *missing genotype model* has approximately the same power and bias as the conditional model, when only phenotype and one genotype are considered. However, we aim to consider a situation where not only the phenotype is measured for the full sample, but also non-genetic covariates. In a simulation setting where data are perfectly normally distributed, the missing genotype model might have similar power as the conditional model, as shown by Huang & Lin (2007). However, we shall see that when real data is analysed, it is often preferable to use models that include as much of the available information as possible.

The conditional model

The probability distribution of phenotypes among individuals with phenotypes $Y < c_l$ or $Y > c_u$ will be referred to as the *conditional distribution.* Let the set $\mathcal{C}$ contain all extreme phenotype values. All subjects in a population with phenotypes that lie in the set $\mathcal{C}$ could be sampled for genotyping. In order to find the distribution of phenotypes in $\mathcal{C}$, we condition on the event ”individual i could be sampled”. Huang & Lin (2007) discussed the conditional model for only one genetic covariate, but we can easily extend the theory to a more general model.

The conditional probability distribution is given by

$$F_{Y|Y\in\mathcal{C};\mathbf{X},\mathbf{G}}(y;\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta},\sigma) = P(Y \le y | Y \in \mathcal{C}; \mathbf{X}, \mathbf{G}),$$

for all $y \in \mathcal{C}$, while $F_{Y|Y\in\mathcal{C};\mathbf{X},\mathbf{G}}(y;\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta},\sigma) = 0$ for all $y \notin \mathcal{C}$. Consider some $y \in \mathcal{C}$. The probability distribution can be rewritten as

$$\begin{aligned} F_{Y|Y\in\mathcal{C};\mathbf{X},\mathbf{G}}(y;\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta},\sigma) &= \frac{P(Y \le y \cap Y \in \mathcal{C}; \mathbf{X}, \mathbf{G})}{P(Y \in \mathcal{C}; \mathbf{X}, \mathbf{G})} \\ &= \frac{P(Y \le y; \mathbf{X}, \mathbf{G})}{P(Y \in \mathcal{C}; \mathbf{X}, \mathbf{G})} \\ &= \frac{P(Y \le y; \mathbf{X}, \mathbf{G})}{1 - P(c_l < Y < c_u; \mathbf{X}, \mathbf{G})} \\ &= \frac{\Phi\left(\frac{y-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}-\boldsymbol{\beta}^T\mathbf{G}}{\sigma}\right)}{1-\Phi\left(\frac{c_u-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}-\boldsymbol{\beta}^T\mathbf{G}}{\sigma}\right)+\Phi\left(\frac{c_l-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}-\boldsymbol{\beta}^T\mathbf{G}}{\sigma}\right)}, \end{aligned}$$

where $\Phi()$ represents the probability distribution of the standard normal distribution. Let ϕ denote the density function of the standard normal distribution. The conditional probability density function for any $y \in \mathcal{C}$ is given by

$$\begin{aligned} f_{Y|Y\in\mathcal{C};\mathbf{X},\mathbf{G}}(y;\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta},\sigma) &= \frac{d}{dy}F_{Y|Y\in\mathcal{C};\mathbf{X},\mathbf{G}}(y;\sigma,\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta}) \\ &= \frac{\frac{1}{\sigma}\phi\left(\frac{y-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}-\boldsymbol{\beta}^T\mathbf{G}}{\sigma}\right)}{1-\Phi\left(\frac{c_u-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}-\boldsymbol{\beta}^T\mathbf{G}}{\sigma}\right)+\Phi\left(\frac{c_l-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}-\boldsymbol{\beta}^T\mathbf{G}}{\sigma}\right)}. \end{aligned}$$

Maximum likelihood estimators can be computed based on knowledge of the probability distribution of the data, and in particular the likelihood. Consider a sample of n individuals with phenotypes Y_i and covariates $\mathbf{X}_i$ and $\mathbf{G}_i$, sampled from the set $\mathcal{C}$. The conditional likelihood is defined as

$$\begin{aligned} &L_c(\sigma,\alpha_0,\boldsymbol{\alpha},\boldsymbol{\beta};Y_1,...,Y_n,\mathbf{X}_1,...,\mathbf{X}_n,\mathbf{G}_1,...,\mathbf{G}_n) = \\ &\prod_{i=1}^{n}\frac{\frac{1}{\sigma}\phi\left(\frac{Y_i-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}_i-\boldsymbol{\beta}^T\mathbf{G}_i}{\sigma}\right)}{1-\Phi\left(\frac{c_u-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}_i-\boldsymbol{\beta}^T\mathbf{G}_i}{\sigma}\right)+\Phi\left(\frac{c_l-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}_i-\boldsymbol{\beta}^T\mathbf{G}_i}{\sigma}\right)}. \end{aligned}$$

For simplicity, we introduce the following notation to represent the likelihood function;

$$L_c = \prod_{i=1}^{n}\frac{\frac{1}{\sigma}\phi_i}{1-\Phi_{u,i}+\Phi_{l,i}}.$$

The conditional log likelihood can be written as

$$l_c = -n\log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n}\frac{1}{2\sigma^2}(Y_i-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}_i-\boldsymbol{\beta}^T\mathbf{G}_i)^2 - \sum_{i=1}^{n}\log(1-\Phi_{u,i}+\Phi_{l,i}).$$

All parameters, α_0, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and σ can be estimated by maximizing the conditional log likelihood function. However, no analytic expressions can be found for these parameters. We recommend using optimization software to find the parameters that maximize the log likelihood.

In order to test for association, $H_0 : \boldsymbol{\beta} = \mathbf{0}$, we use the score test with α_0, $\boldsymbol{\alpha}$ and σ as nuisance parameters. In the simple case where the model $Y_i = \alpha_0 + \beta G_i + \epsilon$ is considered, Tang (2010) proved that the score test based on the conditional model is equivalent to the score test based on normally distributed phenotypes. In other words, although the data are conditionally distributed, one can assume a normal distribution and obtain an unbiased test statistic. In Appendix B.2, we provide the proof that this also holds for a model with several genetic covariates $\mathbf{G}_i$ and none non-genetic covariates $\mathbf{X}_i$, with corresponding null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$.

If the null hypothesis does not contain all covariate coefficients of the linear model, this result does not hold. In such a case, the score test statistic must be obtained by first assuming a conditional distribution of the data. We will outline the expression for the score test statistic, as was done for the linear regression model.

Assume that $\hat{\alpha}_0$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}$ are the maximum likelihood estimators of the parameters in the conditional model, under the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$. We write $\boldsymbol{\theta}_0 = (\hat{\boldsymbol{\theta}}_1, \mathbf{0})$ where $\hat{\boldsymbol{\theta}}_1 = (\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}, \hat{\sigma})$. Before proceeding with the score vector and information matrix we define some functions that simplify notation. The following functions are inspired by Tang (2010);

$$h_{ji} = \frac{-\phi_{u,i} \cdot \left(\frac{c_u - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{G}_i}{\sigma}\right)^j + \phi_{l,i} \cdot \left(\frac{c_l - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{G}_i}{\sigma}\right)^j}{1 - \Phi_{u,i} + \Phi_{l,i}},$$

for $j = 0, 1, 2, 3$. Under H_0 we insert $\boldsymbol{\theta}_0$ and refer to these expressions as h^*_{0i}, h^*_{1i}, h^*_{2i} and h^*_{3i}.

The score vector is given by

$$\begin{aligned}
\mathbf{S}^* &= \left.\frac{\partial l_c}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \left.\left(\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \alpha_0 - \boldsymbol{\alpha}\mathbf{X}_i - \boldsymbol{\beta}\mathbf{G}_i)\mathbf{G}_i - \frac{1}{\sigma}\sum_{i=1}^n \frac{(-\phi_{u,i} + \phi_{l,i})\mathbf{G}_i}{1 - \Phi_{u,i} + \Phi_{l,i}}\right)\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \frac{1}{\hat{\sigma}^2}\sum_{i=1}^n (Y_i - \hat{\alpha}_0 - \hat{\boldsymbol{\alpha}}^T\mathbf{X}_i + \hat{\sigma} h^*_{0i})\mathbf{G}_i.
\end{aligned}$$

The information matrix is found by taking the negative expectation of the second derivatives of the conditional log likelihood, $\boldsymbol{\theta}_0$ inserted. The second derivates are written out in Appendix B.1.2. The elements of the information matrix are given by

$$\begin{aligned}
I(\boldsymbol{\theta}_0)_{11} &= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} \sum_{i=1}^n a_i & \sum_{i=1}^n a_i \mathbf{X}_i^T & \sum_{i=1}^n b_i \\ \sum_{i=1}^n a_i \mathbf{X}_i & \sum_{i=1}^n a_i \mathbf{X}_i \mathbf{X}_i^T & \sum_{i=1}^n b_i \mathbf{X}_i \\ \sum_{i=1}^n b_i & \sum_{i=1}^n b_i \mathbf{X}_i^T & \sum_{i=1}^n c_i \end{bmatrix}, \\
I(\boldsymbol{\theta}_0)_{21} &= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} \sum_{i=1}^n a_i \mathbf{G}_i & \sum_{i=1}^n a_i \mathbf{G}_i \mathbf{X}_i^T & \sum_{i=1}^n b_i \mathbf{G}_i \end{bmatrix}, \\
I(\boldsymbol{\theta}_0)_{12} &= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} \sum_{i=1}^n a_i \mathbf{G}_i^T \\ \sum_{i=1}^n a_i \mathbf{X}_i \mathbf{G}_i^T \\ \sum_{i=1}^n b_i \mathbf{G}_i^T \end{bmatrix}, \\
I(\boldsymbol{\theta}_0)_{22} &= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} \sum_{i=1}^n a_i \mathbf{G}_i \mathbf{G}_i^T \end{bmatrix},
\end{aligned}$$

where, for simplicity, we have used the notation

$$\begin{aligned}
a_i &= 1 - h^*_{1i} - h^{*2}_{0i}, \\
b_i &= 2h^*_{0i} - h^*_{2i} - h^*_{0i}h^*_{1i}, \\
c_i &= 2 + 3h^*_{1i} - h^*_{3i} - h^{*2}_{1i}.
\end{aligned}$$

The information matrix can be used to find the covariance matrix Σ^*, according to Equation (3.18). Finally the score test statistic T^* is found by Equation (3.19). The test statistic is asymptotically χ^2 distributed with degrees of freedom equal to the dimension of $\mathbf{G}$.

The missing genotype model

The extreme phenotype design can be considered a missing at random design, where a subset $\mathcal{M}$ of individuals have missing genotype covariates. Moreover, the distribution of a genotype G for one SNP is known if the MAF q is known, and given by;

$$P(G = g) = \begin{cases} (1-q)^2, & \text{if } g = 0, \\ 2q(1-q), & \text{if } g = 1, \\ q^2, & \text{if } g = 2. \end{cases}$$

If all SNPs in a model can be assumed to have independent genotypes, i.e. that the SNPs are not co-inherited, the probability distribution of a set of genotypes is simply the product of the individual distributions. We can thus define the probability distribution of a set of genotypes $\mathbf{G}$ as

$$P(\mathbf{G} = \mathbf{g}) = \prod_{j=1}^{p_g} P(G_j = g_j),$$

using the individual probability distributions of each SNP as defined above. For a certain number p_g of SNPs, there will be 3^{p_g} different combinations $\mathbf{g}$ of possible genotypes. Let $\mathcal{G}$ denote the set of all these $\mathbf{g}$ for a given number of SNPs.

Let $\mathbf{q}$ be the collection of MAFs, and let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ represent the parameters through which non-genetic covariates $\mathbf{X}$, and genetic covariates $\mathbf{G}$, are related to the phenotype Y. Let $\boldsymbol{\zeta}$ be the parameters that describe the distribution of $\mathbf{X}$. The nuisance parameter is simply σ. We order the data such that $i = 1, \dots, n$ represent completely observed individuals, and $i = n+1, \dots, N$ represent individuals with missing genotypes. We assume that non-genetic and genetic covariates are independent so that $f(\mathbf{x} \cap \mathbf{g}) = f(\mathbf{x})f(\mathbf{g})$.

Based on the likelihood function in Equation (3.11) we can write the likelihood

$$\begin{aligned} L_m &= \prod_{i=1}^{n} f(Y_i|\mathbf{x}_i \cap \mathbf{g}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) f(\mathbf{x}_i \cap \mathbf{g}_i; \boldsymbol{\zeta}, \mathbf{q}) \prod_{i=n+1}^{N} \sum_{\mathbf{g} \in \mathcal{G}} f(Y_i|\mathbf{x}_i \cap \mathbf{g}; \boldsymbol{\alpha}\,\boldsymbol{\beta}, \sigma) f(\mathbf{x}_i \cap \mathbf{g}; \boldsymbol{\zeta}, \mathbf{q}) \\ &= \prod_{i=1}^{n} f(Y_i|\mathbf{x}_i \cap \mathbf{g}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) f(\mathbf{x}_i; \boldsymbol{\zeta}) f(\mathbf{g}_i; \mathbf{q}) \prod_{i=n+1}^{N} \sum_{\mathbf{g} \in \mathcal{G}} f(Y_i|\mathbf{x}_i \cap \mathbf{g}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) f(\mathbf{x}_i; \boldsymbol{\zeta}) f(\mathbf{g}; \mathbf{q}) \\ &= \prod_{i=1}^{n} f(Y_i|\mathbf{x}_i \cap \mathbf{g}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) f(\mathbf{g}_i; \mathbf{q}) \left(\prod_{i=n+1}^{N} \sum_{\mathbf{g} \in \mathcal{G}} f(Y_i|\mathbf{x}_i \cap \mathbf{g}; \boldsymbol{\alpha}\,\boldsymbol{\beta}, \sigma) f(\mathbf{g}; \mathbf{q}) \right) \prod_{i=1}^{N} f(\mathbf{x}_i; \boldsymbol{\zeta}) \end{aligned}$$

For maximum likelihood purposes with respect to the linear regression model, the estimation of $\boldsymbol{\zeta}$ is irrelevant, and we assume that $\mathbf{q}$ is known. The log likelihood function for parameters α_0, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and σ is given by

$$\begin{aligned} l_m &= \sum_{i=1}^{n} f(y_i|\mathbf{x}_i \cap \mathbf{g}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) f(\mathbf{g}_i; \mathbf{q}) \sum_{i=n+1}^{N} \sum_{\mathbf{g}\in\mathcal{G}} f(y_i|\mathbf{x}_i \cap \mathbf{g}; \boldsymbol{\alpha\beta}, \sigma) f(\mathbf{g}; \mathbf{q}) \\ &= -n \log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{G}_i)^2 \\ &- \sum_{i=n+1}^{N} \log \left(\sum_{\mathbf{g}\in\mathcal{G}} \phi \left(\frac{Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{g}}{\sigma} \right) P(\mathbf{G} = \mathbf{g}) \right), \end{aligned}$$

where $P(\mathbf{G} = \mathbf{g})$ is as defined above.

In order to test for association, $H_0 : \boldsymbol{\beta} = 0$, we use the score test with α_0, $\boldsymbol{\alpha}$ and σ as nuisance parameters. Assume that $\hat{\alpha}_0$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}$ are the maximum likelihood estimators of the parameters in the conditional model under the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ and let $\boldsymbol{\theta}_0 = (\hat{\boldsymbol{\theta}}_1, \mathbf{0})$. We define the functions

$$\begin{aligned} k_{ji} &= \frac{\sum_{\mathbf{g}} (Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{g})^j \phi_i(\mathbf{g}) P(\mathbf{G} = \mathbf{g})}{\sum_{\mathbf{g}} \phi_i(\mathbf{g}) P(\mathbf{G} = \mathbf{g})}, \\ k_{ji}(\mathbf{g}) &= \frac{\sum_{\mathbf{g}} (Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{g})^j \phi_i(\mathbf{g}) P(\mathbf{G} = \mathbf{g})\mathbf{g}}{\sum_{\mathbf{g}} \phi_i(\mathbf{g}) P(\mathbf{G} = \mathbf{g})}, \end{aligned}$$

for $j = 0, 1, 2, 3, 4$, where

$$\phi_i(\mathbf{g}) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left(\frac{1}{2\sigma^2} (Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{g})^2 \right).$$

Under H_0, $k_{ij} = (Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i)^j$ and $k_{ij}(\mathbf{g}) = (Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i)^j \,\mathrm{E}(\mathbf{G})$, as $\sum_{\mathbf{g}} P(\mathbf{G} = \mathbf{g}) = 1$ and $\sum_{\mathbf{g}} P(\mathbf{G} = \mathbf{g})\mathbf{g} = \mathrm{E}(\mathbf{G})$.

The score vector is given by

$$\mathbf{S}^* = \frac{\partial l_m}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0} = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (Y_i - \hat{\alpha}_0 - \hat{\boldsymbol{\alpha}}^T \mathbf{X}_i) \mathbf{G}_i + \frac{1}{\hat{\sigma}^2} \sum_{i=n+1}^{N} (Y_i - \hat{\alpha}_0 - \hat{\boldsymbol{\alpha}}^T \mathbf{X}_i) \,\mathrm{E}(\mathbf{G}).$$

The information matrix is found by taking the negative expectation of the second derivatives of the log likelihood, $\boldsymbol{\theta}_0$ inserted. The second derivatives are written out in Appendix B.1.3. We note that under H_0, expressions such as $k_{2i} - k_{1i}^2$ disappear. The elements of the information matrix are given by

$$
\begin{aligned}
I(\boldsymbol{\theta}_0)_{11} &= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} N & N\bar{\mathbf{X}}^T & 0 \\ N\bar{\mathbf{X}} & \sum_{i=1}^N \mathbf{X}_i\mathbf{X}_i^T & \mathbf{0}_{p_x\times 1} \\ 0 & \mathbf{0}_{1\times p_x} & 2N \end{bmatrix}, \\
I(\boldsymbol{\theta}_0)_{21} &= \frac{1}{\hat{\sigma}^2}\left[\sum_{i=1}^n \mathbf{G}_i + \sum_{i=n+1}^N \mathrm{E}(\mathbf{G}) \quad \sum_{i=1}^n \mathbf{G}_i\mathbf{X}_i^T + \sum_{i=n+1}^N E(\mathbf{G})\mathbf{X}_i^T \quad \mathbf{0}_{p_g\times 1}\right], \\
I(\boldsymbol{\theta}_0)_{12} &= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} \sum_{i=1}^n \mathbf{G}_i^T + \sum_{i=n+1}^N \mathrm{E}(\mathbf{G})^T \\ \sum_{i=1}^n \mathbf{X}_i\mathbf{G}_i^T + \sum_{i=n+1}^N \mathbf{X}_i E(\mathbf{G})^T \\ \mathbf{0}_{1\times p_g} \end{bmatrix}, \\
I(\boldsymbol{\theta}_0)_{22} &= \frac{1}{\hat{\sigma}^2}\left[\sum_{i=1}^n \mathbf{G}_i\mathbf{G}_i^T + \sum_{i=n+1}^N \mathrm{E}(\mathbf{G}\mathbf{G}^T)\right], \qquad (4.4)
\end{aligned}
$$

where $E(\mathbf{G}\mathbf{G}^T) = \sum_{\mathbf{g}} \mathbf{g}\mathbf{g}^T P(\mathbf{G} = \mathbf{g})$. These expressions are used to find the covariance matrix Σ^* according to Equation (3.18), and the score test statistic T^* according to Equation (3.19). The test statistic is asymptotically χ^2 distributed with p_g degrees of freedom.

In the above we assumed that the MAFs $\mathbf{q}$ were known or externally estimated. It is possible to include $\mathbf{q}$ as a nuisance parameter in the model but this will cause the derivation of the score test statistic to be much more complicated. For our intents and purposes it is sufficient to assume that the MAFs are known. We have also assumed that genotypes and non-genetic covariate levels occur independent of each other which allowed us to exclude the parameter $\boldsymbol{\zeta}$ from the likelihoods. This means that we cannot use the score test as defined above in models where non-genetic covariates are mediators between genetic variables and the phenotype.

Chapter 5

An extreme phenotype sampling common variant association study

In this chapter we will apply the common variant association methods presented in the previous chapter to a real dataset. This dataset stems from the HUNT project as introduced in Chapter 1. The dataset has known non-genetic variables for the full sample, while two genetic variables (genotypes of the SNPs rs9939609 and rs17782313) are only known for individuals with extreme measurements of waist-hip ratio (WHR). The methods presented in the previous chapter are all appropriate in ideal situations where data fulfil assumptions of independence, normality and constant variance. The HUNT dataset allows us to investigate the performance of these methods in a real sample where assumptions are not necessarily satisfied. Although this analysis is concerned with common variants, some of the lessons learned from this study could be relevant for further work with rare variant association studies. The genotyping of extreme individuals in the WHR study headed by Ingrid Mostad, sparked our interest in the statistical theory of extreme phenotype sampling. The analysis performed in this chapter is only part of a larger study in which we aim to quantify the effects of the SNPs rs9939609 and rs17782313 on WHR in the HUNT population. In this chapter we focus on how extreme phenotype sampling methods for common variants can be applied in a real dataset, and the challenges that are revealed in the process.

5.1 Descriptive analysis

The SNP dataset was acquired from the population of Nord-Trøndelag as illustrated in Figure 5.1. Most individuals in Nord-Trøndelag were invited to participate in HUNT3 and thus the HUNT3 population is practically identically to the population of Nord-Trøndelag. Those that decided to participate in the study constitute the HUNT3 sample. In the analysis by Mostad et al. (2014), this sample was somewhat reduced to exclude individuals with some missing critical values, particularly waist and hip circumference measurements. The WHR dataset consists of 50336 subjects; 27459 women and 22877 men. The SNP dataset consists of individuals that were selected for genotyping of the SNPs rs9939609 and rs17782313 by an extreme phenotype sampling criterion. This results in 24916 genotyped individuals; 13598 women and 11318 men. Extreme phenotypes were defined to be WHR values in the upper and lower WHR quartiles based on the WHR

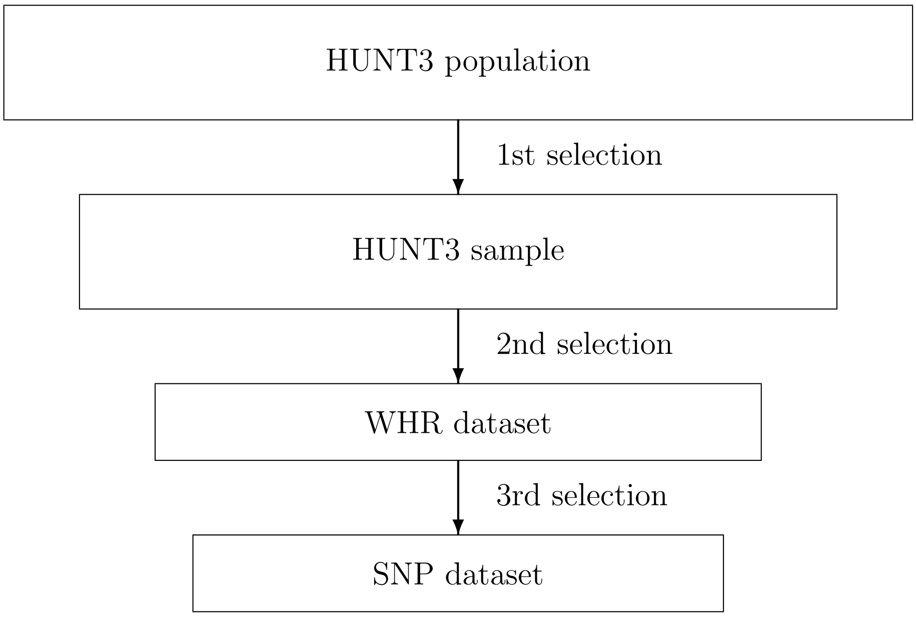


Figure 5.1: Acquiring the datasets for this study

dataset, calculated for men and women separately. Thus, the SNP dataset was sampled based on the criteria WHR $> c_u^f$ or WHR $< c_l^f$ for women, and WHR $> c_u^m$ or WHR $< c_l^m$ for men, where

$$\begin{aligned} c_l^f &= 0.817, \quad c_u^f = 0.917, \\ c_l^m &= 0.895, \quad c_u^m = 0.981. \end{aligned}$$

As sampling of extreme phenotypes is performed separately for the two genders, we will always perform analysis for these two datasets separately. There are more women than men who were genotyped for the SNPs. This reflects the degree of participation to the HUNT3 sample. There is an uneven amount of genotyped individuals for the two SNPs which we assume is caused by some occasional unsuccessful genotyping in the lab.

The response variable

We consider waist-hip ratio (WHR) as the response or dependent variable in this study. Waist and hip circumferences may in general be approximately normally distributed in a population as they are physical measurements. A ratio of the two is on the other hand, not necessarily normally distribution. For example, it is widely accepted that BMI, which is a ratio of measurements of height and weight, is generally log-transformed before assuming a normal distributed. To obtain an approximately normally distributed response variable, a transformation of WHR is needed. The use of WHR is not as widespread as BMI, but the log-transform does appear reasonable, see Appendix C.2 for details. It is clear that even after the transformation of WHR, the data are not perfectly normally distributed. We must be aware of this when fitting models to the data.

The average WHR in the WHR dataset is 0.8685 among women and 0.9387 for men. The corresponding standard deviation is 0.0741 for women and 0.0659 for men. Summary statistics of WHR in separate WHR quartiles are presented in Table 5.1. We see that women have on average lower WHR values than men, but higher variation in these measurements.

Gender	WHR1	WHR2	WHR3	WHR4	WHR
Women	6844	6874	6813	6928	27459
Men	5668	5764	5692	5753	22877

(a) Number of individuals in each WHR quartile

Gender	WHR1	WHR2	WHR3	WHR4	WHR
Women	0.7751	0.8438	0.8915	0.9626	0.8685
Men	0.8571	0.9163	0.9578	1.0227	0.9387

(b) Average WHR in each WHR quartile

Gender	WHR1	WHR2	WHR3	WHR4	WHR
Women	0.0328	0.0146	0.0136	0.0410	0.0741
Men	0.0314	0.0120	0.0126	0.0387	0.0659

(c) Standard deviation of WHR in each WHR quartile

Table 5.1: Waist-hip ratio summary statistics.

Non-genetic covariates

The HUNT study provides information on several non-genetic variables for each individual. Investigating dietary factors that could influence WHR in the HUNT population is not the focus of this thesis, but rather the work of Mostad et al. (2014). In this thesis we focus on the inclusion of genetic variables in a regression model with non-genetic covariates present. The genetic data is only available in the SNP dataset, but other covariates are available for the WHR dataset. Concerning non-genetic covariates, we focus on age groups, smoking habits and exercise frequency (see Appendix C.1 for definitions).

Age is by Mostad et al. (2014) treated as a categorical variable with the levels ($20 \leq \text{age} < 30$, $30 \leq \text{age} < 40$, $40 \leq \text{age} < 50$, $50 \leq \text{age} < 60$, $60 \leq \text{age} < 70$, $70 \leq \text{age} < 80$, $\text{age} \geq 80$). These seven categories will be referred to as $a1$, $a2$, $a3$, $a4$, $a5$, $a6$ and $a7$. The number of individuals in the WHR dataset that have valid age group values are presented in Table 5.2, together with the counts of those individuals with valid age group, rs9939609 and rs17782313 recordings. We see that age groups $a3$, $a4$ and $a5$ are well represented, while age groups $a1$, $a2$ and $a6$ less so. There are especially few individuals in $a7$, which is the group consisting of the oldest individuals. The age covariate will be included in our models as a non-confounder. The distributions of log(WHR) within each age group are illustrated in Figure 5.2. This figure also illustrates the uneven number of subjects in each age group. The vertical lines are the logarithms of the upper and lower cut-off values for genotyping. We see that the curves for age groups 3, 4 and 5 peak close to the center of the two vertical lines. Age groups 1 and 2 have their peaks in the lower end while age groups 6 and 7 peak close to the upper cut-off. These figures clearly illustrate that the different age groups have different centres of their respective WHR distributions.

		WHR dataset		SNP dataset	
Category		Women	Men	Women	Men
Age	$a1$	2544	1838	1510	1316
	$a2$	3941	2839	1974	1469
	$a3$	5433	4548	2559	2050
	$a4$	5980	5395	2703	2456
	$a5$	5104	4644	2521	2200
	$a6$	3066	2639	1586	1337
	$a7$	1391	973	745	490
Smoke	$s0$	12093	9248	6159	4526
	$s1$	2022	1896	995	1001
	$s2$	5372	4578	2522	2170
	$s3$	1396	2459	679	1252
	$s4$	3751	2081	1798	1031
	$s5$	1785	1814	917	937
Exercise	$e0$	1075	1500	584	816
	$e1$	3585	4553	1845	2261
	$e2$	5595	4930	2727	2379
	$e3$	11146	7712	5367	3816
	$e4$	5260	3606	2645	1766

Table 5.2: Counts of individuals with valid non-genetic variables in the WHR dataset and the SNP dataset.

The curves in Figure 5.2 suggest that the subgroups created by age categories have different variances in log(WHR). We have analysed this more carefully in Appendix C.2. If the variance of log(WHR) is not constant after we have corrected for genetic and non-genetic covariates, one assumption of the continuous phenotype regression models described in Chapter 4, is violated.

Ideally, we would analyse the effects of the two SNPs on WHR by correcting for population stratification. We know that in the population of Nord-Trøndelag, many people are related, which introduces confounding in our studies. A family's genetic heritage can be a cause for genotype, while a family's cultural heritage can be a cause for high or low WHR values. Unfortunately, we cannot correct for population stratification directly due to lack of information. However, we may assume that smoking habits and exercise frequencies are variables that to a certain degree reflects different families' cultural heritage. Thus, we can include these variables in our models as potential confounders. Additionally, we expect to see an association between these variables and WHR, separate of the confounding effects.

We have created a smoking variable (see Appendix C.1) which is a categorical variable with six levels ranging from 0 to 5. These categories will be referred to as $s0$, $s1$, $s2$, $s3$, $s4$ and $s5$. The corresponding counts of subjects with valid replies in the HUNT3 study are presented in Table 5.2. Similarly, an exercise variable with five categories was created based on frequency of exercise (see Appendix C.1). These levels will be referred to as $e0$, $e1$, $e2$, $e3$ and $e4$. The corresponding counts are presented in Table 5.2.

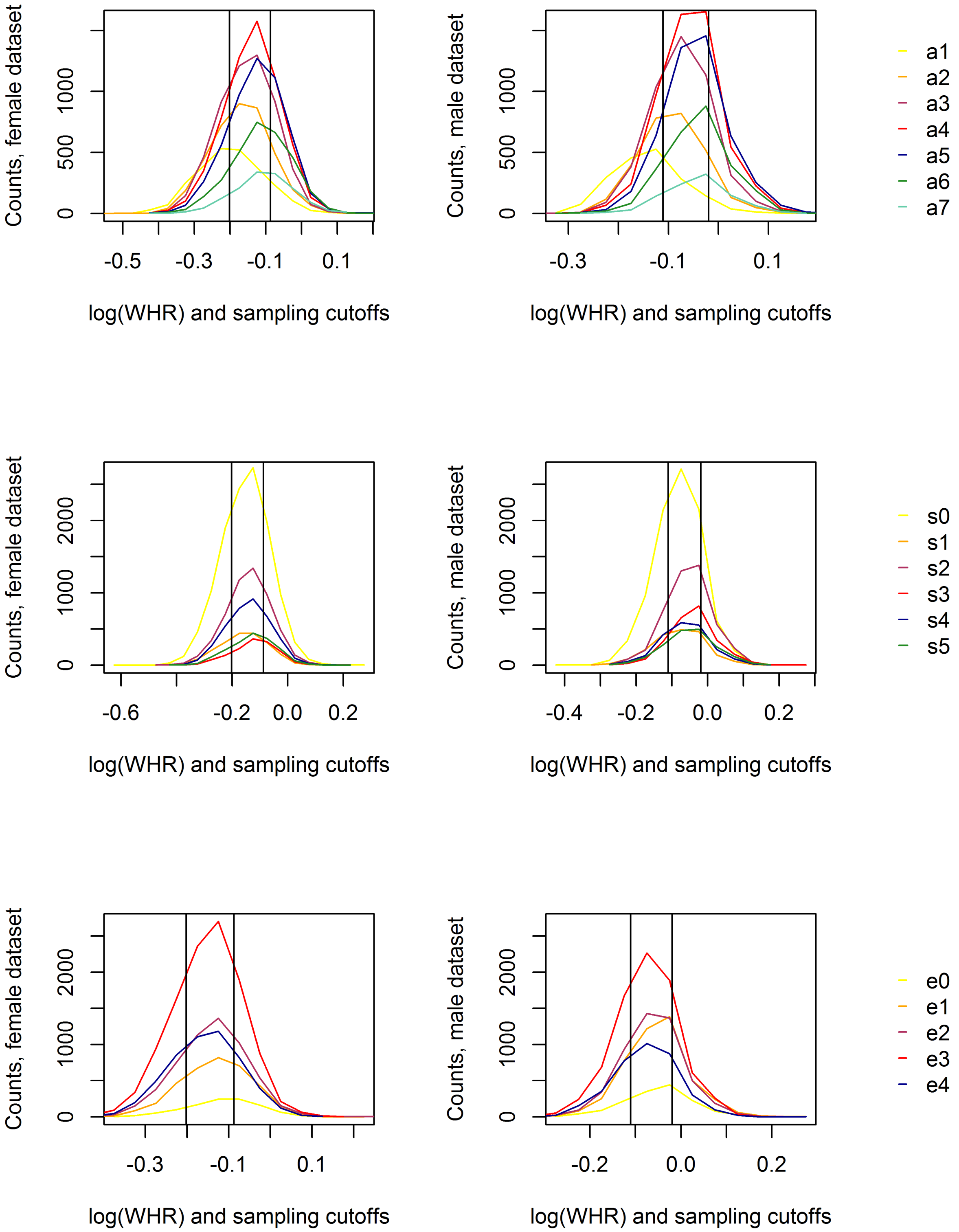


Figure 5.2: Histograms of the seven age groups, six smoke levels, and five exercise frequencies, with cut-off values. Results for women are in the left column, and results for men are in the right column.

Figure 5.2 also illustrates distributions of log(WHR) among the different smoke and exercise groups. The differences between groups are not as clear for these variables as for age groups. However, it does seem as if some smoke levels and exercise frequencies are associated with higher WHR values. Again we see a suggestion of different variances in the different subgroups. We must be cautious of these assumption violations in the following.

Genetic covariates

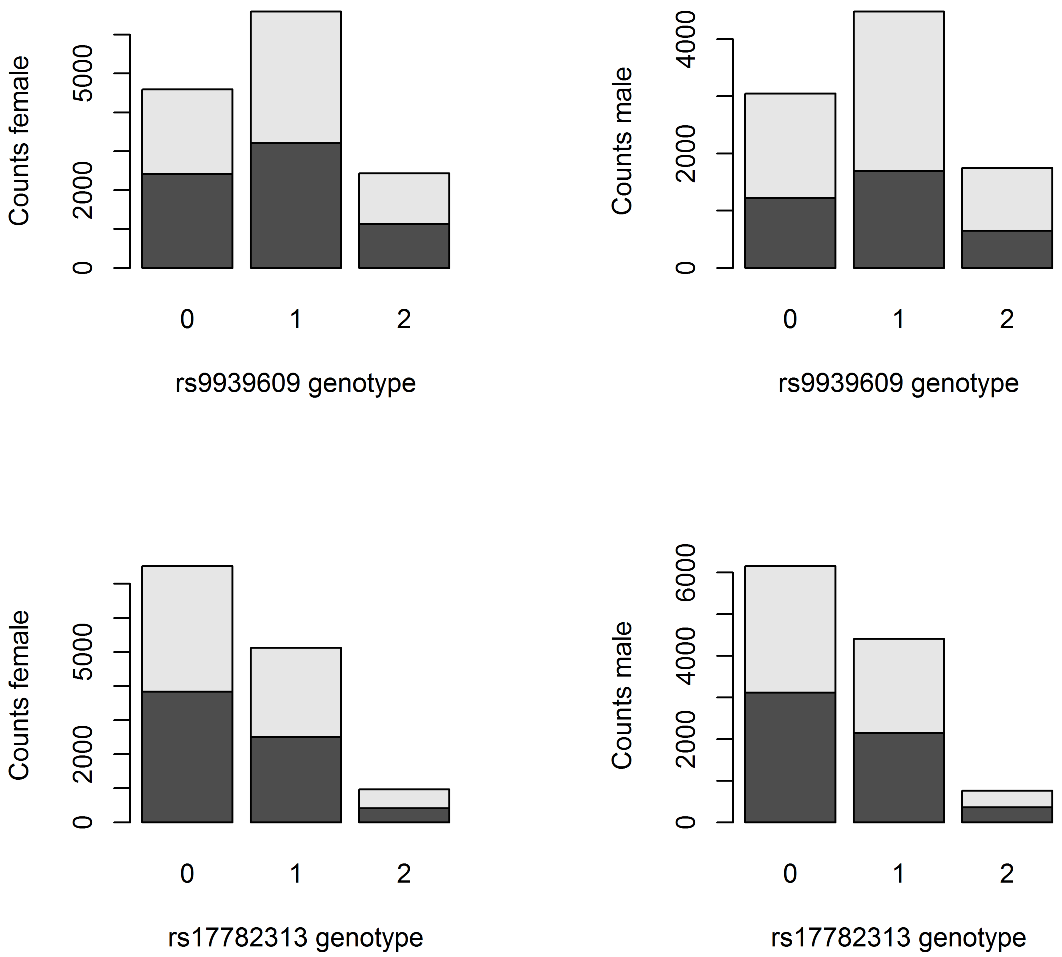


Figure 5.3: Histograms of the genotypes of the two SNPs in WHR1 (dark grey) and WHR4 (light grey). Results for women are in the left column, and results for men are in the right column.

For known MAFs of the SNPs, frequencies of the three different genotypes can be calculated as in Equation (2.2), where q is the minor allele frequency. Additionally, we can estimate genotype frequencies and MAFs from our dataset. Due to the extreme sampling, we do not expect frequencies in the SNP dataset to equal the reference values. Under the assumption of a monotone genetic model, we would expect to see higher frequencies of the homozygous high-risk genotypes in WHR4 and lower frequencies in WHR1, compared to reference frequencies. The genotype frequencies for men and women in the SNP dataset are presented in Table 5.3. Based on the SNP rs9939609 and the assumption of a monotone genetic model, we would expect the population MAF for Nord-Trøndelag to lie between 0.4 and 0.44. For the SNP rs17782313, we would expect the population MAF to lie between 0.25 an 0.27. Consequently, the population of Nord-Trøndelag seems to reflect the rs17782313 CEU population MAF of 0.26 and possibly the TSI population MAF of 0.27, but the sample MAF of rs9939609 is not coherent with neither the CEU nor the TSI population, with reference MAFs of 0.46 and 0.47, respectively (*The International HapMap Project* 2002-2009). Our estimates are based on a much larger sample than the HapMap estimates. Although we have estimated MAFs based on extreme phenotype individuals we will, in accordance with a monotone genetic model, assume that the MAFs lie in between estimated MAFs of the WHR1 and WHR4 groups. We will therefore discard the HapMap MAFs for the SNP rs9939609, and use $q_{rs9939609} = 0.42$ and $q_{rs17782313} = 0.26$ as MAF estimates for the population of Nord-Trøndelag.

In Figure 5.3 we illustrate the different distributions of genotypes in WHR quartiles 1 and 4. It is difficult to see any distinct differences between WHR1 and WHR4 although we notice that among men, the distribution of rs9939609 genotypes is clearly different in the two WHR groups. It appears that in general, the proportion of WHR4 individuals increases while the relative proportion of WHR1 individuals decreases, when the number of high-risk alleles increase.

Gender	SNP	WHR	g_0	g_1	g_2	MAF	CATT p-value
Male	rs9939609	1	0.3500	0.4988	0.1512	0.4006	$1.913 \cdot 10^{-8}$
		4	0.3198	0.4876	0.1926	0.4364	
	rs17782313	1	0.5546	0.3818	0.0636	0.2545	$6.261 \cdot 10^{-3}$
		4	0.5332	0.3959	0.0709	0.2689	
Female	rs9939609	1	0.3582	0.4750	0.1667	0.4043	$7.027 \cdot 10^{-7}$
		4	0.3161	0.4938	0.1901	0.4370	
	rs17782313	1	0.5681	0.3717	0.0601	0.2460	$2.988 \cdot 10^{-8}$
		4	0.5372	0.3816	0.0812	0.2720	

Table 5.3: Observed allele frequencies and MAFs in the SNP dataset

5.2 Case-control analysis

Before we proceed with our analysis of WHR as a continuous phenotype, we will (incorrectly) perform a case-control analysis where we consider WHR4 individuals as cases and WHR1 individuals as controls. This definition is not consistent with the definition of a case-control analysis because the controls are not randomly sampled from individuals outside the case group. We do this analysis in order to compare our data with the results of the BMI studies by Frayling et al. (2007) and Loos et al. (2008). First, we will (correctly) use the CATT test defined in Equation (3.20) to test for equality of the genotype frequencies between the WHR1 and WHR4 groups. As alternative hypothesis, we use an additive genetic model. The resulting p-values are presented in Table 5.3. We see that the null hypothesis is rejected on a 0.05 significance level (0.025 if a Bonferroni correction is applied) for all gender and SNP groups, corresponding to an increase in WHR group with an increase in high-risk alleles for the two SNPs.

We use a logistic regression model to estimate the odds ratio. As our focus is on estimating precise effects, we use the results by Robinson & Jewell (1991) and include only confounders, and not non-confounders, in our models. The SNPs we are investigating can be assumed non-confounders as they lie on different chromosomes. The effect of these must therefore be estimated in two separate models. Age is a nonconfounder and is excluded from the model. We include smoking and exercise frequency because these can be considered as possible confounders for WHR and genotype, albeit indirectly. The coefficient estimates for genetic variables among women are

$$\hat{\beta}_{rs9939609} = 0.1257, \quad \hat{\beta}_{rs17782313} = 0.1291,$$

and

$$\hat{\beta}_{rs9939609} = 0.1565, \quad \hat{\beta}_{rs17782313} = 0.0895$$

among men.

Frayling et al. (2007) reported an odds ratio between homozygotes of the rs9939609 SNP of 1.38 for being overweight, and 1.67 for being obese, compared with normal weight individuals. Between individuals in WHR1 and WHR4, we estimate corresponding odds ratios of 1.29 among women, and 1.38 among men. These ratios compare homozygotes for the low risk and high risk alleles of rs9939609, i.e. the risk of increased WHR between individuals with no high-risk alleles and individuals with two high-risk alleles.

Loos et al. (2008) reported an per-allele odds ratio of the rs17782313 SNP of 1.05 for being overweight, and 1.08 for being obese compared with normal weight individuals. Between individuals in WHR1 and WHR4, we estimate corresponding odds ratios of 1.14 among women, and 1.09 among men. These ratios reflect the risk of increased WHR between individuals with j and $j+1$ high risk alleles, $j = 0, 1$.

Recall that the interpretation of these ratios, as with the odds ratio between obese and normal weight people, is not interpretable as the standard odds ratio we know from theory. This is due to the lack of random sampling of the so-called controls in these studies. However, the results of both the CATT test and the logistic regression analysis suggest that there are genetic differences between the WHR1 and WHR4 groups.

5.3 Continuous phenotype association analysis

Our aim with the following association analysis is to apply the methods described in the previous chapter to a real dataset. That is, we will analyse a cross-sectional sample with a linear regression model, as well as an extreme phenotype sample, using the conditional and missing genotype models. We perform the analysis of the genetic effects on WHR with different models in order to investigate differences between them when applied to real data.

Datasets

We define the *complete sample* as the subset of the WHR dataset for which all individuals have valid WHR, age, smoke and exercise values. The complete sample consists of 26108 women and 21954 men. Only half of the subjects in the complete sample have valid genotype values and we have excluded those with only one valid genotype. We can consider the complete sample including or excluding SNP data. We define the *extreme sample* as the subset of the SNP dataset for which all individuals have valid WHR, age, smoke, exercise and genetic variables. This results in 12893 women and 10960 men in the extreme sample. We will consider the extreme sample both with genotypes included and excluded. We also consider a *small sample*, which is a random sample of half of the subjects of the complete sample with genotypes excluded. This sample represents a different sampling design than the extreme sampling but with the same sample size, as opposed to the complete sample where the sample size is twice as large.

Models

A multiple linear regression (MLR) model can be fitted to the complete sample if genotypes are excluded. This model provides reference values for the coefficients of age, smoke and exercise. The MLR model can also be fitted to the small sample. The missing genotype model can be fitted to the complete sample including SNP data so as the estimate genetic effects. The non-genetic parameters of the missing genotype model can be compared to the corresponding parameters of the MLR model fitted to the complete sample. The conditional model can be fitted to the extreme sample with both genotypes excluded and included. When SNP data are excluded, the parameters of non-genetic covariates can be compared to that of the MLR model fitted to the small sample, which is of the same sample size. This comparison can aid us in evaluating whether a random or extreme sample is more appropriate for estimating effects accurately. In the event where N individuals are sampled and non-genetic variables recorded, and $N/2$ extremes are thereafter genotyped, we can assess the performance of the conditional and missing genotype models by fitting these to the extreme and complete samples with SNP data included. If data were normally distributed, independent, and had constant variance, these models should in theory yield similar results.

In order to keep track of the different datasets and regression models, we have illustrated which models that will be fitted to which samples in Table 5.4.

Sample	Model
Small sample, excluding SNP data Complete sample, excluding SNP data	MLR model
Complete sample, including SNP data	Missing genotype model
Extreme sample, excluding SNP data Extreme sample, including SNP data	Conditional model

Table 5.4: An overview over the different samples and which models that apply to them.

A note on the missing genotype model

For the missing genotype regression model it is possible to estimate the MAFs in the model, or insert externally estimated parameters. The score test that is defined in Chapter 4.2.2 uses externally estimated MAFs. (For completeness, we implemented a model where MAFs were estimated internally and compared parameters to those of our chosen model. The results were similar.) We use externally estimated MAFs of 0.42 for rs9939609 and 0.26 for rs17782313. As discussed, the MAFs are on the extreme sample and therefore difficult to estimate accurately. However, we believe that these estimated MAFs are more appropriate than the HapMap reference MAFs, for our dataset.

Tests

We will consider five tests for main effects and seven tests for interaction effects, in each model. Thus, by the Bonferroni correction, an overall significance level of 0.05, translates to 0.004 for each test.

Sample	Model	Gender	Covariate	p-value
Complete	Missing genotype	Women	rs9939609	$1.5986 \cdot 10^{-12}$
			rs17782313	$1.2203 \cdot 10^{-13}$
		Men	rs9939609	$1.6258 \cdot 10^{-7}$
			rs17782313	$1.3248 \cdot 10^{-2}$
Extreme	Conditional	Women	rs9939609	$2.0222 \cdot 10^{-6}$
			rs17782313	$6.5116 \cdot 10^{-8}$
			rs9939609:rs17782313	$2.4696 \cdot 10^{-1}$
			rs9939609:age	$1.5008 \cdot 10^{-1}$
			rs17782313:age	$2.2354 \cdot 10^{-1}$
			rs9939609:smoke	$8.9816 \cdot 10^{-1}$
			rs17782313:smoke	$4.4982 \cdot 10^{-1}$
			rs9939609:exercise	$1.5960 \cdot 10^{-1}$
			rs17782313:exercise	$9.5124 \cdot 10^{-1}$
		Men	rs9939609	$1.5494 \cdot 10^{-8}$
			rs17782313	$1.7185 \cdot 10^{-2}$
			rs9939609:rs17782313	$7.7818 \cdot 10^{-1}$
			rs9939609:age	$1.2561 \cdot 10^{-1}$
			rs17782313:age	$2.4027 \cdot 10^{-1}$
			rs9939609:smoke	$9.4043 \cdot 10^{-1}$
			rs17782313:smoke	$9.0247 \cdot 10^{-2}$
			rs9939609:exercise	$5.0038 \cdot 10^{-1}$
			rs17782313:exercise	$5.3558 \cdot 10^{-1}$

Table 5.5: Score test p-values for the genetic covariates, main effects and interactions

We first tested the three hypothesis of whether age, smoke or exercise coefficients are zero in the MLR model and conditional model fitted to the complete, small and extreme samples with genetic variables excluded. The variables were treated as dummy variables where the lowest levels ($a1$, $s0$ and $e0$) were included in the intercept α_0. For the age variables, the null hypothesis is defined as $H_0 : \alpha_{a2} = \ldots = \alpha_{a7} = 0$. With all other factors held constant, this translates to $H_0 : \alpha_{a1} = \alpha_{a2} = \ldots = \alpha_{a7}$, i.e. that the effect of age is the same for all age groups and therefore not an age specific effect. For smoking, the null hypothesis is $H_0 : \alpha_{s1} = \ldots = \alpha_{s5} = 0$, while for exercise frequencies we have $H_0 : \alpha_{e1} = \ldots = \alpha_{e4} = 0$. In each test, remaining covariates were treated as nuisance variables. For the MLR model fitted to the complete and small samples, the score test statistics are defined in Section 4.2.1, while for the extreme sample, the conditional model is assumed and the score test statistic is given in Section 4.2.2. All tests for the effect of non-genetic variables yielded p-values lower than 10^{-16} and we conclude that age, smoke and exercise are factors that separately affect the waist-hip ratio of men and women, and that these should be included in our models.

We used the missing genotype model fitted to the complete sample including SNP data, and the conditional model fitted to the extreme sample including SNP data, to test for the significance of genotypes with respect to WHR. For the conditional model, we tested the two SNPs separately while including the other SNP in the model as a nuisance variable. This is not possible by our definition of the missing genotype model, and we therefore tested each SNP separately with the other SNP excluded from the analysis. The results are presented in Table 5.5. After a Bonferroni correction, none of the models yielded a significant effect of rs17782313 among men. Both SNPs were found to be significant for WHR in women, and rs9939609 seems to be associated with WHR in men.

Using the conditional model and corresponding score test, we tested whether there was evidence in the extreme dataset for interaction effects between genotypes and other covariates. The results are presented in Table 5.5. We see that none of these effects are significant. By our definition of the missing genotype model we are not able to test for specific genetic interaction effects while simultaneously including main genetic effects in the model. We also note that based on our definition of the missing genotype likelihood, we must assume that the SNPs rs9939609 and rs17782313 do not influence smoke or exercise habits.

Effect estimates

In summary, we fitted the MLR model to the complete and small samples without genotypes, the conditional model to the extreme sample without genotypes, and the MGM and CM to the complete and extreme samples with genotypes included. The parameters of the fitted models are presented in Tables 5.6 (women) and 5.7 (men). These parameters represent coefficients of non-genetic and genetic covariates in the different models, using log(WHR) as the dependent variable. In order to obtain the effects of different covariates on WHR, we take the exponential of these estimates. For example, the effect of being in age group 2 for women, as estimated by the MLR model fitted to the complete dataset, is given by $\exp(2.66444 \cdot 10^{-2}) = 1.027$. Compared to the WHR value $\hat{Y}_{a1}$ predicted for age group 1, an increase in age group corresponds to predicted WHR $\hat{Y}_{a2} = 1.027\hat{Y}_{a1}$.

Genetic effects are estimated by the missing genotype model and conditional model fitted to the complete and extreme datasets with SNP data included. These estimated coefficients for these models with log(WHR) as the dependent variable are presented in columns two and five of Tables 5.6 and 5.7. The missing genotype model estimates genetic effects of 1.001 ($= \exp(9.1207 \cdot 10^{-4})$) for the SNP rs9939609, and 1.004 ($= \exp(3.6082 \cdot 10^{-3})$) for the SNP rs17782313, among women. Similarly, estimated genetic effects are 1.035 and 1.025 among men. This means that if a person will increase her (his) WHR by a multiple of 1.001 (1.035) if the rs9939609 genotype is increased with one high-risk allele, and by a multiple of 1.004 (1.025) if the rs17782313 genotype is increased with one high-risk allele. The conditional model estimates genetic effects of 1.006 (rs9939609) and 1.008 (rs17782313) among women. Similarly, estimated genetic effects are 1.006 and 1.002 among men.

We illustrate these effect estimates by the following examples. A woman with waist circumference 80 cm and hip circumference of 100 cm will have a waist-hip ratio of 0.8. A 1.001 genetic effect corresponds to an increased WHR of 0.8008. If this increase was due to central fat only, this corresponds to a 0.08 cm increase in waist circumference. This increase is clearly clinically insignificant. A man with waist circumference 81 cm and hip circumference 90 cm will have a waist-hip ratio of 0.9. A genetic effect of 1.035 corresponds to an increased WHR of 0.9315 which is a 2.835 cm increase in waist circumference.

Sample	Complete		Small	Extreme	
Model	MLR	Missing	MLR	Conditional	Conditional
Intercept	-1.8219e-01	-1.8333e-01	-1.8744e-01	-2.0350e-01	-2.1200e-01
$a2$	2.6444e-02	2.6276e-02	2.6087e-02	3.0727e-02	3.0493e-02
$a3$	4.1659e-02	4.1711e-02	4.2698e-02	5.6481e-02	5.6743e-02
$a4$	5.6048e-02	5.5952e-02	5.7492e-02	8.2339e-02	8.2175e-02
$a5$	6.9408e-02	6.9441e-02	6.8023e-02	1.0838e-01	1.0820e-01
$a6$	8.3784e-02	8.3644e-02	8.3057e-02	1.3333e-01	1.3284e-01
$a7$	9.0100e-02	8.9744e-02	9.3682e-02	1.4216e-01	1.4176e-01
$s1$	1.6263e-03	1.8815e-03	2.2706e-03	1.0905e-03	1.1209e-03
$s2$	7.6442e-03	7.7900e-03	7.7614e-03	1.4731e-02	1.4626e-02
$s3$	2.8852e-02	2.8665e-02	2.6990e-02	4.8314e-02	4.7907e-02
$s4$	8.8743e-03	8.9446e-03	9.4017e-03	1.5504e-02	1.5184e-02
$s5$	2.2483e-02	2.2115e-02	2.2762e-02	3.9681e-02	3.9278e-02
$e1$	-5.5619e-03	-5.8926e-03	-3.8855e-03	-4.4613e-03	-5.1652e-03
$e2$	-1.3148e-02	-1.3259e-02	-9.9883e-03	-1.6646e-02	-1.7508e-02
$e3$	-2.3514e-02	-2.3553e-02	-1.8114e-02	-3.5884e-02	-3.6589e-02
$e4$	-3.3326e-02	-3.3348e-02	-2.8563e-02	-5.1103e-02	-5.1780e-02
rs9939609	-	9.1207e-04	-	-	6.2919e-03
rs17782313	-	3.6082e-03	-	-	8.0234e-03
σ	8.0599e-02	8.0360e-02	8.0907e-02	1.0543e-01	1.0522e-01

Table 5.6: Regression coefficients, female population

Sample	Complete		Small	Extreme	
Model	MLR	Missing	MLR	Conditional	Conditional
Intercept	-1.2772e-01	-1.8187e-01	-1.2773e-01	-1.4564e-01	-1.5221e-01
$a2$	4.4467e-02	3.6146e-02	4.3821e-02	4.7403e-02	4.7125e-02
$a3$	6.1652e-02	5.5745e-02	6.2471e-02	7.2838e-02	7.2481e-02
$a4$	7.8680e-02	7.7524e-02	7.7872e-02	1.0462e-01	1.0442e-01
$a5$	9.1163e-02	9.5699e-02	9.0217e-02	1.2825e-01	1.2790e-01
$a6$	9.6211e-02	1.0575e-01	9.6931e-02	1.3283e-01	1.3242e-01
$a7$	9.6389e-02	1.1015e-01	9.8141e-02	1.3417e-01	1.3404e-01
$s1$	7.9770e-03	8.5792e-03	7.7164e-03	1.2653e-02	1.2816e-02
$s2$	1.1942e-02	1.5943e-02	1.0549e-02	2.1022e-02	2.1069e-02
$s3$	2.3655e-02	4.3921e-02	2.5637e-02	3.9394e-02	3.9220e-02
$s4$	4.6428e-03	7.5284e-03	5.3806e-03	8.0356e-03	7.7803e-03
$s5$	1.7141e-02	2.9338e-02	1.3580e-02	2.8901e-02	2.9071e-02
$e1$	-1.3133e-03	-7.2479e-04	-4.4794e-04	1.6743e-03	1.8236e-03
$e2$	-1.0820e-02	-1.6889e-02	-9.9697e-03	-1.3395e-02	-1.3442e-02
$e3$	-2.4030e-02	-3.3968e-02	-2.3357e-02	-3.5721e-02	-3.5674e-02
$e4$	-3.0762e-02	-4.4040e-02	-2.9906e-02	-4.4905e-02	-4.5109e-02
rs9939609	-	3.4808e-02	-	-	6.3071e-03
rs17782313	-	*2.5228e-02*	-	-	*2.9600e-03*
σ	6.2883e-02	8.4008e-02	6.2974e-02	8.0345e-02	8.0206e-02

Table 5.7: Regression coefficients, male population

We now use the results in Tables 5.6 and 5.7 to compare the performance of the different models. The two first columns of these tables contain estimates based on the complete sample including and excluding SNP data. Although the MLR model and missing genotype model are different, it is not surprising that the non-genetic coefficients of these models are very similar as these models are fitted to the same non-genetic data.

The small and extreme samples without genetic covariates represent different sampling strategies that can be used if only half of a population can be analysed. The MLR model and and the conditional model have been fitted to these samples. The estimated parameters can be compared to the parameters of the ”full” population, namely those of the complete sample without genotypes. We see that the results based on the small sample are much more accurate than those of the extreme sample. It could be argued that extreme sampling design is intended for estimating effects of genotypes and not age, smoke and exercise variables. However, if all assumptions were met, the third and fourth columns of Tables 5.6 and 5.7 should be approximately equal. We notice that the coefficients of old age and heavy smoking are higher in the conditional model, while the coefficients of frequent exercise are smaller. It therefore seems as if the conditional model based on the extreme sample has overestimated effects of causal variables, compared to the MLR model fitted to the small sample, as well as the MLR model fitted to the complete sample.

Genetic effects are estimated by the missing genotype model and conditional model fitted to the complete and extreme samples with SNP data included. In Table 5.7, effects that were not found to be significant in association tests are in italic. In the female population we see similar results for genetic covariates as for non-genetic covariates; while rs9939609 and rs17782313 have small effects in the missing genotype model, the estimated coefficients are higher in the conditional model. In the male population however, we see the opposite; the genetic effects are estimated to be larger in the missing genotype model compared to the conditional model. Based on these observations we cannot conclude that in our datasets, the conditional model consistently overestimates effect sizes.

We illustrate the effects of covariates in general, and genetic covariates in particular in Figures 5.4 to 5.9. All figures present estimated WHR for different levels of age, smoke and exercise as estimated by the MLR model fitted to the complete sample excluding SNP data. These estimates are included as a measure of comparison with estimated non-genetic effects of the conditional model and the missing genotype model. The figures present estimated WHR for different genotypes of rs9939609 and rs17782313 for different levels of age, smoke and exercise, based on estimates from the conditional model or the missing genotype model.

Consider Figure 5.4 which presents estimated WHR for women for different age groups and different genotypes. We see that compared to the MLR model fitted to the complete sample excluding genetic variables, the effects of old and young age are overestimated by the conditional model fitted to the extreme sample including genetic variables. The missing genotype model fitted to the complete sample with SNP data included, appears to estimate age effects more accurately. We see that the genetic effects estimated by the missing genotype model are perhaps statistically significant, but clinically insignificant. In Figure 5.5 we see the same results for the male population. Again, we see that the conditional model overestimates age effects compared to the MLR model fitted to the complete sample. The missing genotype model estimates larger genetic effects for men.

Figures 5.6 and 5.7 show corresponding results for different smoke levels, while Figures 5.8 and 5.9 illustrate exercise effects.

Evaluating the fit of the models

It is interesting to note the R^2 and R^2_{adj} values, as defined in Equations (3.12) and (3.13), of the fitted models. Models fitted to the complete sample including and excluding SNP data, and the the small sample excluding SNP data, have R^2 and R^2_{adj} values of about 0.11 among women. Among men, the values are approximately 0.19 for the MLR model fitted to the complete and small samples excluding SNP data, and 0.26 for the missing genotype model fitted to the complete sample including SNP data. Models fitted to the extreme sample including genetic variables have R^2 and R^2_{adj} values of about 0.18 for women and 0.29 for men.

The R^2 values indicate that the conditional model explains a larger portion of the variation in the data than the MLR model. This observation could help explain the over- and underestimation of parameters in the conditional regression models. In the extreme sample, the variation is focused in the tails of the distributions. The MLR model "overlooks" tail irregularities in the presence of the bulk of observations in the centre of the WHR scale for the complete and small samples. The conditional model, on the other hand, fits a model that reflects a normal distribution based on what it sees; namely

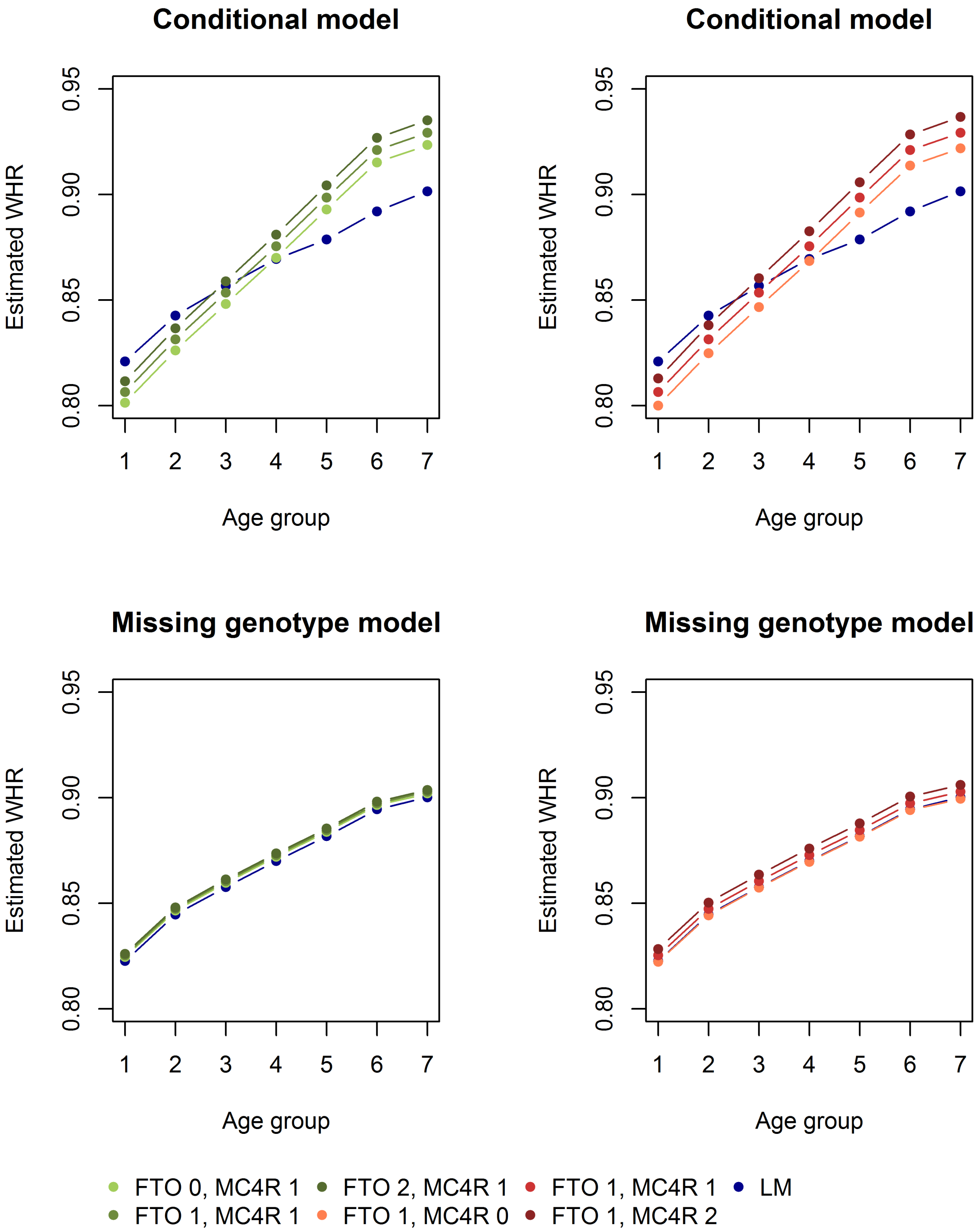


Figure 5.4: Estimated WHR for women, smoke level 0 and exercise frequency 3 held constant. FTO 0 represents the genotype 0 (no high-risk alleles) for the SNP rs9939609, FTO 1 represents the genotype 1 (one high-risk allele) for the SNP rs9939609, FTO 2 represents the genotype 2 (two high-risk alleles) for the SNP rs9939609, MC4R 0 represents the genotype 0 (no high-risk alleles) for the SNP rs17782313, MC4R 1 represents the genotype 1 (one high-risk allele) for the SNP rs17782313, MC4R 2 represents the genotype 2 (two high-risk alleles) for the SNP rs17782313.

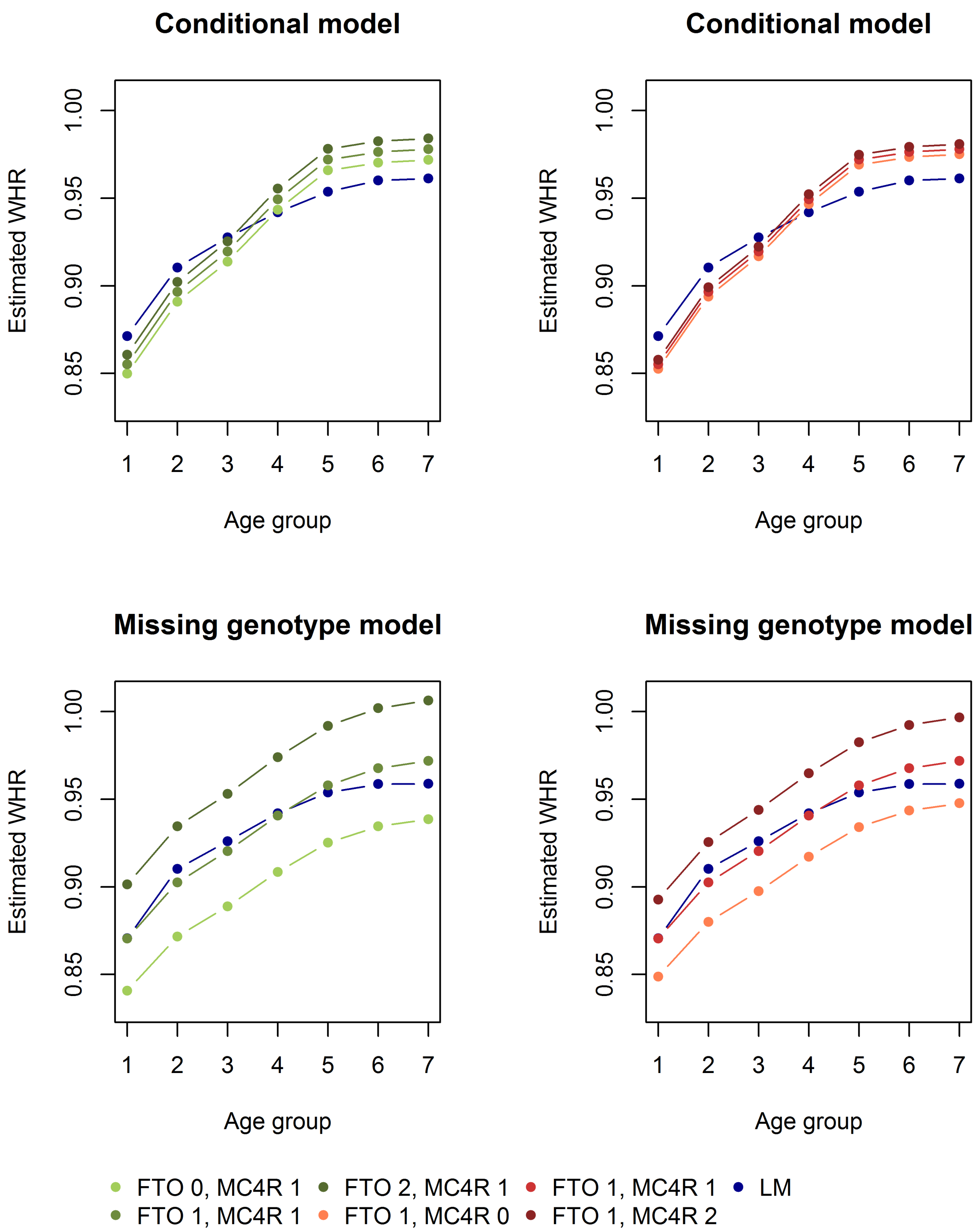


Figure 5.5: Estimated WHR for men, smoke level 0 and exercise frequency 3 held constant. FTO 0 represents the genotype 0 (no high-risk alleles) for the SNP rs9939609, FTO 1 represents the genotype 1 (one high-risk allele) for the SNP rs9939609, FTO 2 represents the genotype 2 (two high-risk alleles) for the SNP rs9939609, MC4R 0 represents the genotype 0 (no high-risk alleles) for the SNP rs17782313, MC4R 1 represents the genotype 1 (one high-risk allele) for the SNP rs17782313, MC4R 2 represents the genotype 2 (two high-risk alleles) for the SNP rs17782313.

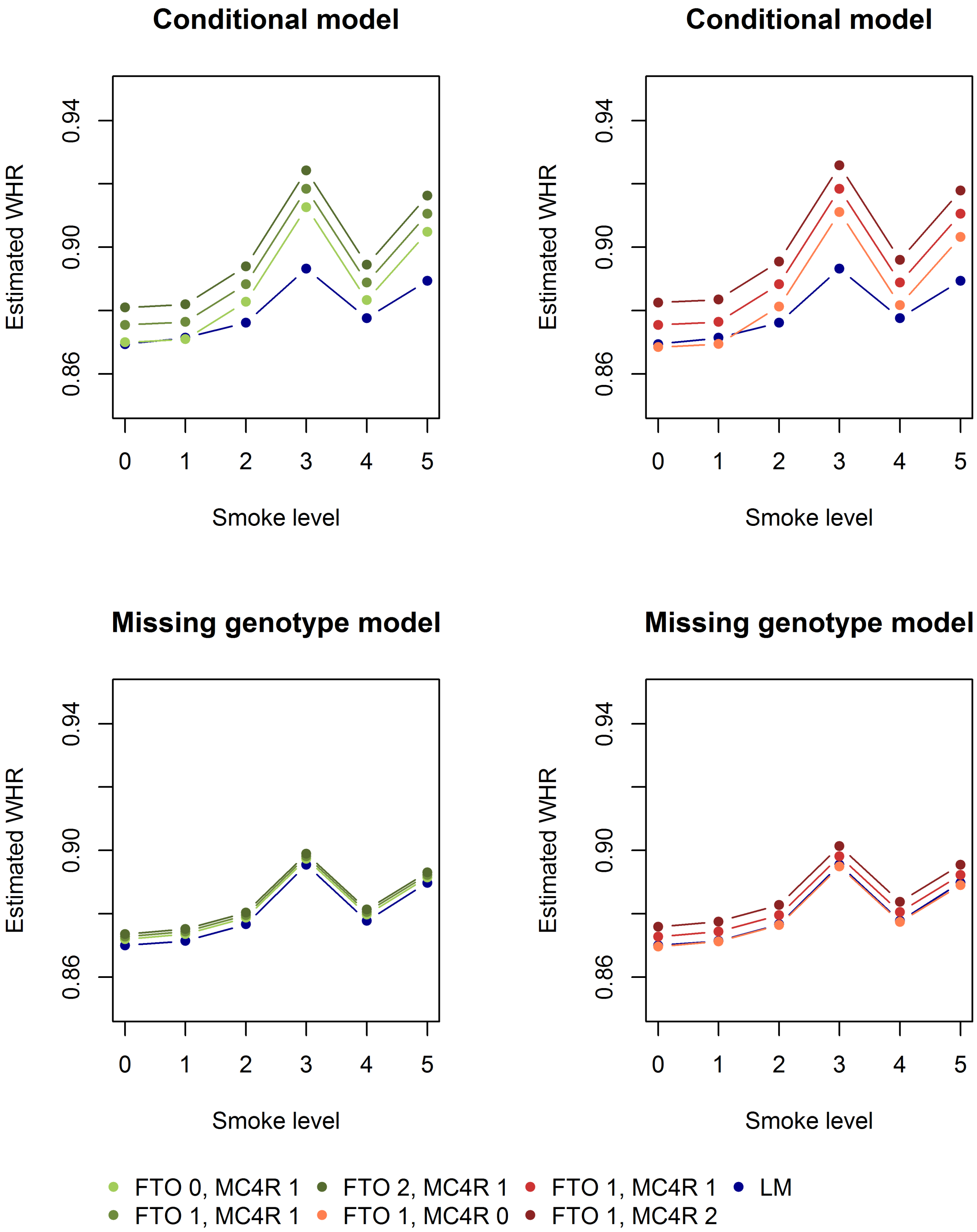


Figure 5.6: Estimated WHR for women, age group 4 and exercise frequency 3 held constant. FTO 0 represents the genotype 0 (no high-risk alleles) for the SNP rs9939609, FTO 1 represents the genotype 1 (one high-risk allele) for the SNP rs9939609, FTO 2 represents the genotype 2 (two high-risk alleles) for the SNP rs9939609, MC4R 0 represents the genotype 0 (no high-risk alleles) for the SNP rs17782313, MC4R 1 represents the genotype 1 (one high-risk allele) for the SNP rs17782313, MC4R 2 represents the genotype 2 (two high-risk alleles) for the SNP rs17782313.

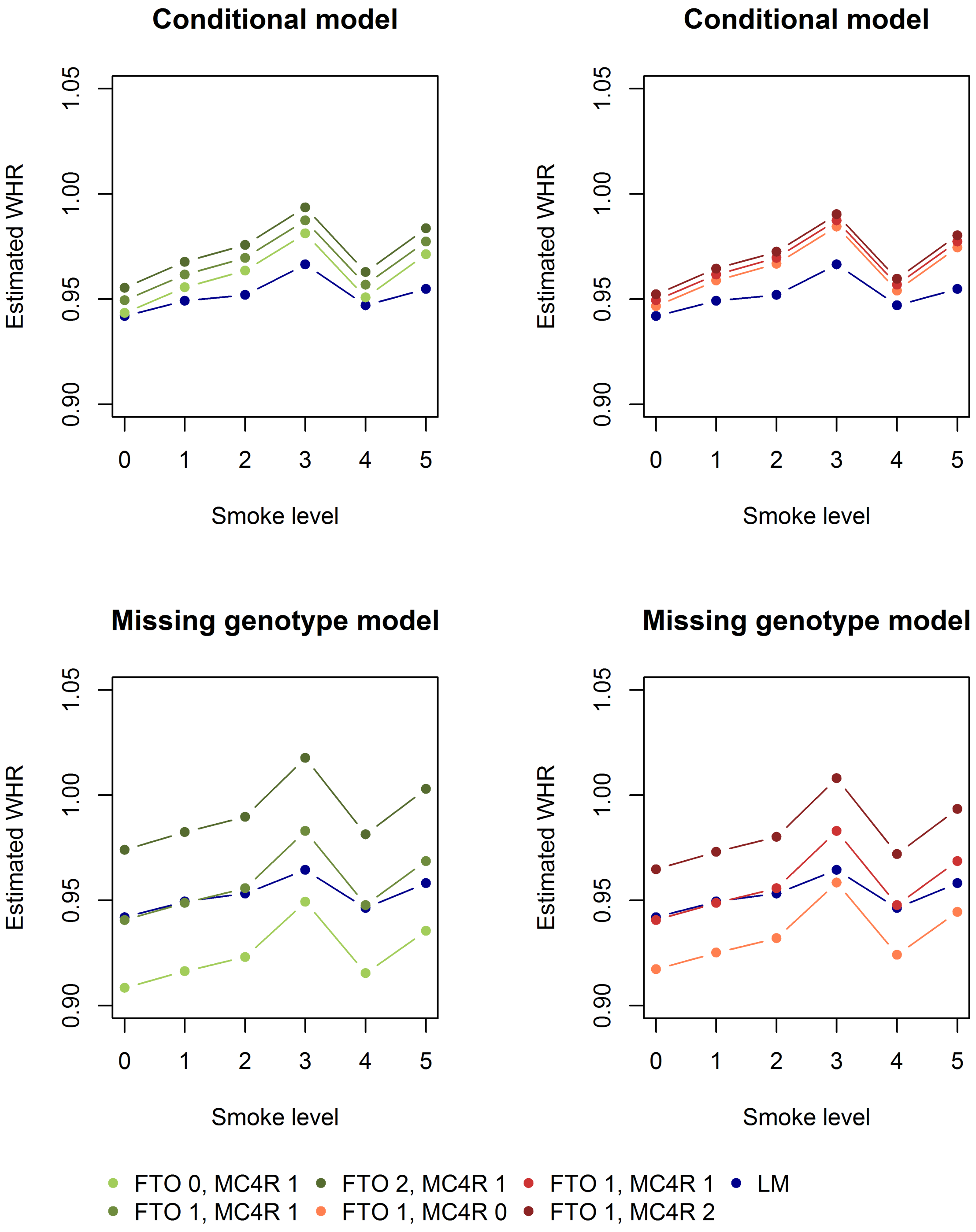


Figure 5.7: Estimated WHR for men, age group 4 and exercise frequency 3 held constant. FTO 0 represents the genotype 0 (no high-risk alleles) for the SNP rs9939609, FTO 1 represents the genotype 1 (one high-risk allele) for the SNP rs9939609, FTO 2 represents the genotype 2 (two high-risk alleles) for the SNP rs9939609, MC4R 0 represents the genotype 0 (no high-risk alleles) for the SNP rs17782313, MC4R 1 represents the genotype 1 (one high-risk allele) for the SNP rs17782313, MC4R 2 represents the genotype 2 (two high-risk alleles) for the SNP rs17782313.

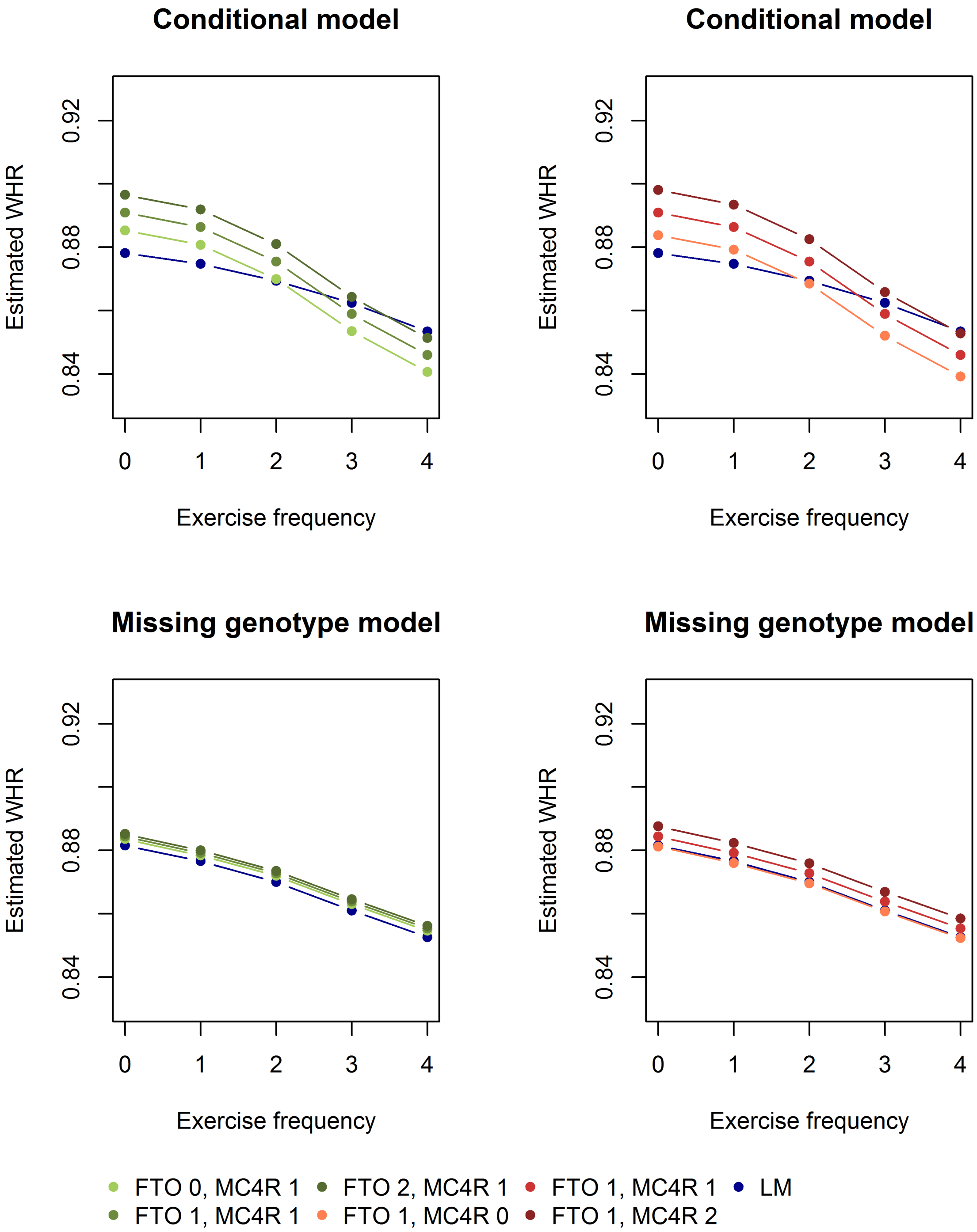


Figure 5.8: Estimated WHR for women, age group 4 and smoke level 0 held constant. FTO 0 represents the genotype 0 (no high-risk alleles) for the SNP rs9939609, FTO 1 represents the genotype 1 (one high-risk allele) for the SNP rs9939609, FTO 2 represents the genotype 2 (two high-risk alleles) for the SNP rs9939609, MC4R 0 represents the genotype 0 (no high-risk alleles) for the SNP rs17782313, MC4R 1 represents the genotype 1 (one high-risk allele) for the SNP rs17782313, MC4R 2 represents the genotype 2 (two high-risk alleles) for the SNP rs17782313.

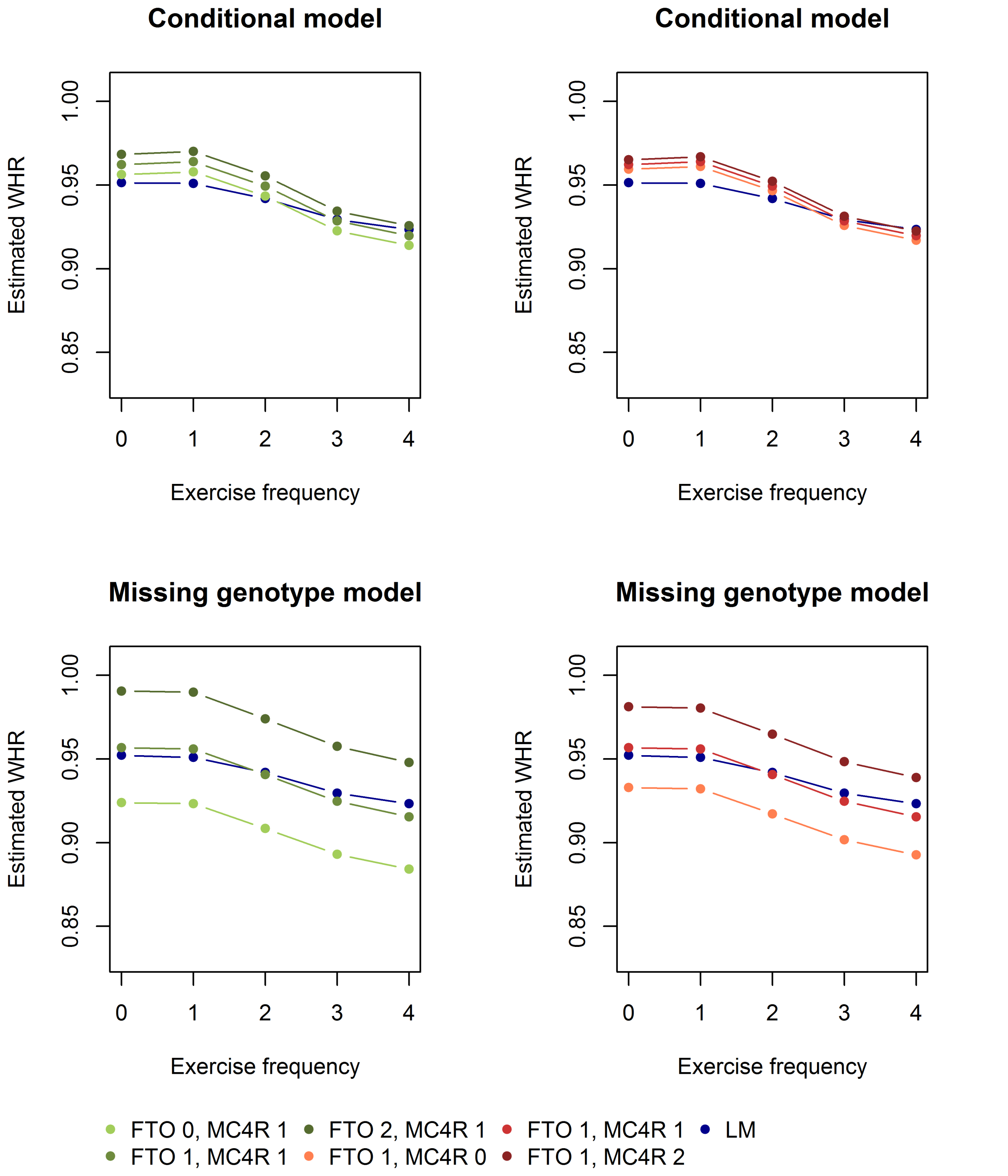


Figure 5.9: Estimated WHR for men, age group 4 and smoke level 0 held constant. FTO 0 represents the genotype 0 (no high-risk alleles) for the SNP rs9939609, FTO 1 represents the genotype 1 (one high-risk allele) for the SNP rs9939609, FTO 2 represents the genotype 2 (two high-risk alleles) for the SNP rs9939609, MC4R 0 represents the genotype 0 (no high-risk alleles) for the SNP rs17782313, MC4R 1 represents the genotype 1 (one high-risk allele) for the SNP rs17782313, MC4R 2 represents the genotype 2 (two high-risk alleles) for the SNP rs17782313.

the tails of the distribution. Consequently, when the underlying data are not perfectly normally distributed in the tails, the different models assume different underlying normal distributions. This does not occur when the data are perfectly normally distributed - in which case all the models would give similar estimates. The R^2 values cannot be calculated precisely for the missing genotype model as phenotype cannot be predicted in the half of the dataset that lacks genotype values. However, we have estimated the values by predicting phenotype based on non-genetic variables were genetic variables are missing, and based on non-genetic and genetic variables were both are present.

For women, the missing genotype model does not explain more variation in the complete sample than the MLR model. For men, the missing genotype model explains more variation than the MLR model. We have seen that in the female samples, the missing genotype model does not find any clinically significant effect of genotypes, while in the male samples, the genetic effects estimated by the missing genotype model are relatively large. This is reflected in the observed R^2 values.

Residual plots are presented for completeness in Appendix C.3. With such large numbers of subjects it is not surprising that the residuals appear as blocks of random observations. No trends are seen in any of the plots.

Discussion and conclusion

In our final models, we included age, smoke, exercise and both genotypes as covariates for the modelling of log(WHR). The SNP rs17782313 was not found to be significant among men, but from an epidemiological point of view, it is important to include this variable in the models. The parameters of the models are presented in Tables 5.6 and 5.7. We have also created figures that illustrate effects of age, smoke, exercise and genotypes as estimated in the missing genotype and conditional models, see Figures 5.4 to 5.9.

As mentioned, we consider the multiple linear regression model based on the complete dataset excluding SNP data, as the most accurate with respect to the estimates of age, smoke and exercise coefficients. This is due to the large sample size of the complete sample. As no interaction effects were found between genetic and non-genetic variables, the effect of genotypes should be a constant shift in log(WHR), and correspondingly an exponential shift on the WHR scale. Models that include genetic variables should ideally estimate parameters for non-genetic variables similarly to the multiple linear regression model. We see that the missing genotype model achieves this, while the conditional model seems to over- and underestimate effect sizes. It appears as if the extreme phenotype sampling design is less robust than random sampling designs against violations of assumptions such as normality, independent observations and constant variance. We cannot reveal which of these violations the conditional model is more sensitive to, based on this study. As further work in this field, we hope to test the different designs and models in a GWAS study where it is possible to correct for population stratification and where genotypes are known for a complete sample. Additionally, simulation studies could be used to investigate the effects of model violations.

It is not surprising that the missing genotype model, which uses a lot more of the available information than the conditional model, resembles the MLR model fitted to the complete sample. If some covariates are known in a full sample, while genotypes are only known for an extreme phenotype sample, the missing genotype model is preferable in any event where one cannot be certain that all assumptions of the conditional model are satisfied. We recognize that for this dataset where only two SNPs were investigated, the missing genotype model was relatively easy to implement and computationally fast to use. However, due to the iterations over all possible combinations of genotypes, the method can become more time-consuming as the number of genetic variables increases. We have also had the liberty in this study to assume known distributions of the genetic variables, and that these variables are independent. If this was not the case, the score test for the missing genotype model would be more complicated due to the unknown parameters of the genetic distributions, as well as the inclusion of joint distributions in the likelihood function.

In our collaborative project with Ingrid Mostad we aim to estimate the effects of genotypes in the HUNT population. Due to the issues discussed here of population stratification and differences in variance across subgroups, we must consider other methods to assess the risk of high WHR as a result of genotype. We will use epidemiological methods in this analysis and consider subgroups separately in order to identify risk among healthy and less healthy individuals. We will also investigate the use of z-scores for each subgroup so as to create variables that are centred and scaled in order to correct for different subgroup variances. However, this analysis is outside the realms of this thesis. Here, we have used the HUNT dataset as an example of the extreme phenotype sampling design, and used it to test different EPS methods. Although the EPS design can theoretically be more powerful than random sampling to detect associations between genotypes and phenotypes, the EPS design appears to be sensitive to violations of assumptions for which the random sampling design is more robust. For studies such as the one we have analysed here, we would recommend randomly sampling individuals for genotyping if for some reason one does not have the opportunity to genotype all subjects.

Chapter 6

Statistical models and methods for rare variant association studies

Rare variant association studies are based on the common disease rare variant hypothesis, although recent methods aim to incorporate a combination of both the CDRV and the CDCV hypothesis. Lee et al. (2012) explain that "since standard individual variant tests, typically used for analysis of SNPs, are underpowered to detect rare variant effects due to the low allele frequencies and the large number of rare variants in the genome, region-based analysis has become the standard approach for analysing rare variants in sequencing studies". Such regions could be a whole chromosome, a gene or subregions of a gene. We will refer to these regions as loci.

The main goal of current rare variant association methods is to discover loci that are associated with a disease or phenotype. By the CDRV hypothesis, the goal is to discover loci that were not already found in GWA studies. The focus is not so much to quantify the effect of carrying rare variants in the genome. Because of high allelic heterogeneity, it is the region as a unit that is of interest and not the individual variants. Many variants in a locus are thought to have an effect on phenotype, and one tag SNP is not assumed to be sufficient to describe the effect of all potentially causal rare variants.

The methods used for testing for an association between a locus and a phenotype in this chapter, build in the score test statistics that were developed for the multiple linear regression model, and conditional model, in Chapter 4. This chapter is therefore a rare-variant extension of Chapter 4.

We will often refer to genetic variables G in the following models. If G represents genotypes of particular positions, the coding of G is as mentioned in Chapter 2, dependent of the genetic models of the different variants. That is, for a monotone genetic model, G takes values 0, 1, and 2 according to the number of high-risk alleles which in this chapter corresponds to the number rare variants at a certain position. For recessive models, G takes the values 0, 0, and 1, while for dominant models, G takes the values 0, 1, and 1.

6.1 Burden methods

Burden methods are quite straight-forward methods that collapse information across a locus in order to reduce the number of variables and thereby simplify test procedures. In the following, we will introduce the burden methods proposed by Li & Leal (2008).

These were developed for the case-control design, but we will adapt them for use in cross-sectional studies with a continuous outcome. Other well-known burden methods are the Cohort Allelic Sums Test (CAST) (Morgenthaler & Thilly 2006) and the Weighted Sum Test (WST) (Madsen & Browning 2009).

Li & Leal (2008) propose two burden methods; the *collapsing method* and the *Combined Multivariate and Collapsing* (CMC) method. In the literature, the CMC method is often confused with the more simple collapsing method. The distinction is important as the CMC method clearly outperforms the collapsing method in complex studies.

In a case-control study, the aim is to test whether there is a difference in genotype frequencies between the two groups. For GWA problems, the methods used to go about testing this would be a single-variate test for each position with a correction for multiple tests, or one multivariate test that includes all positions in the locus. Li & Leal (2008) propose the χ^2 and the Hotelling's T^2 tests for these methods, but claim that both will be low powered in a rare variant association study. For single variant testing, a large number of tests must be performed and the correction for multiple testing causes a loss of power. If a multivariate test is used, the number of degrees of freedom of the test statistic becomes very large and this in turn causes a loss of power. To increase power, Li & Leal (2008) propose the collapsing method and the CMC method.

The collapsing method

Consider a case and a control group that are genotyped for M variants in a specific locus. The collapsing method summarizes the genetic information across this locus into *one* variable defined by

$$G = \begin{cases} 1, & \text{if one or more rare variants are present, and} \\ 0, & \text{otherwise.} \end{cases} \tag{6.1}$$

Li & Leal (2008) explain that as it is uncommon to carry two rare variants at the same locus, the collapsing method is valid for loci consisting of only rare variants. The underlying assumption must be that all rare variants in the locus have the same effect on the disease.

The null hypothesis is that the distribution of G is equal in the case and control group. This hypothesis can be tested by the χ^2 test as defined in Equation (3.3.2). If several separate loci are to be tested, a Bonferroni correction for multiple testing can be used. The number of multiple tests will be relatively low, and the correction will not cause a significant loss of power.

The Combined Multivariate and Collapsing method

The collapsing method is a crude method as it collapses all variants in a locus. A lot of relevant information will be lost in the presence of one or more common variants. Although the definition of a common variant is MAF ≥ 0.05, we have for example seen that the MAF of the SNP rs9939609 is close to 0.5, which means that the probability that subjects carry one or more copies of the high-risk allele, is large. Using the collapsing method on a locus with this SNP and other rare variants present would effectively serve to estimate the effect of rs9939609 only.

The CMC method collapses subregions of the locus, defined for example as all variants with MAFs lower than some threshold value. Thus, $m \ll M$ subregions of the loci are collapsed to form the variables $G_1, \ldots, G_m$. A common variant can remain uncollapsed and the corresponding genetic variable can retain the values corresponding to the chosen genetic model. The collapsed genetic variables are defined as in Equation (6.1). Li & Leal (2008) propose to use the Hotelling's T^2 test to test whether the distributions of $G_1, \ldots, G_m$ are equal among cases and controls.

The collapsing and CMC methods adapted to the cross-sectional design with a continuous phenotype

For a cross-sectional study with a continuous phenotype we assume a linear regression model with a normally distributed outcome Y that is either the phenotype itself, or some appropriate transformation of it. We let $\mathbf{X}$ denote any non-genetic covariates. For the collapsing method we assume a linear model

$$Y = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X} + \beta G + \epsilon,$$

where G is the collapsed random variable as defined in Equation (6.1) and ϵ is $N(0, \sigma^2)$-distributed. The inclusion of non-genetic variables $\mathbf{X}$ in the model means that we are able to control for population stratification and other confounders. The null hypothesis $H_0 : \beta = 0$ can be tested by the score test, as defined for the multivariate linear regression model in Section 4.2.1. The model described here has previously been used by Li et al. (2011).

For the CMC method, genetic information from M variants is collapsed into $m << M$ genetic variables $G_1, \ldots, G_m$. We let these be elements of a vector $\mathbf{G}$. The linear model is given by

$$Y = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X} + \boldsymbol{\beta}^T \mathbf{G} + \epsilon.$$

Again, the score test can be used to test the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$, i.e. that there is no association between the locus and the phenotype.

Criticism of burden methods

Li & Leal (2008) discuss misclassification and its effect on burden tests. Misclassification is defined as the inclusion of biologically nonfunctional variants in the model, or the exclusion of biologically functional variants. The CMC method is only able to distinguish between functional and nonfunctional variants if the collapsing criteria is to collapse these types separately. However, this information is rarely known beforehand. Li & Leal (2008) provide evidence that the collapsing method is always more powerful than the χ^2 test and the Hotelling's T^2 test, but recognize that the multivariate Hotelling's T^2 test is more robust under misclassification. This issue provides motivation for the CMC method, which is shown to be both powerful and robust against misclassification. Yet, Lee et al. (2012) claim that all burden tests suffer under misclassification as region sizes increase.

A criticism of the CMC method is that it requires threshold values for collapsing different variants and that there exist no natural such threshold values.

Wu et al. (2011) emphasize that burden tests assume that rare variants are either all deleterious or all protective, i.e. that they affect a phenotype in the same direction. For the CMC method, if the nature of the variants are known, this can be incorporated in the model by separately collapsing variants with effects in the opposite directions. However, this information is rarely available beforehand.

6.2 Kernel-based methods

The Sequence Kernel Association Test (SKAT) was proposed by Wu et al. (2011). The test improves upon the issues with burden tests in that it allows for effects of variants in either direction, and different magnitudes of effect. It allows for the inclusion of other covariates and thus the opportunity to control for population stratification. The SKAT method is developed for dichotomous and continuous phenotypes. The kernel-based methods are not the focus of this thesis. However, they are important to discuss as they form the link between the well-known burden tests and the recent functional linear model approach. We will therefore give a brief overview of the SKAT methods. For in-depth descriptions, see Wu et al. (2011) and Lee et al. (2012).

The SKAT method for a continuous normally distributed outcome assumes a linear model

$$Y = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X} + \boldsymbol{\beta}^T \mathbf{G} + \epsilon,$$

where $\mathbf{G}$ contains genotypes of all genotyped variants in the locus. The coefficients $\beta_1, \ldots, \beta_M$ are assumed to follow arbitrary distributions with mean zero and variance $w_j \tau$, where w_j are chosen weights for each variant, $j = 1, \ldots, M$. The null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ can thus be rewritten as $H_0 : \tau = 0$, which reduces the dimension of the problem from M to one test. According to Wu et al. (2011), this null hypothesis can be tested by a so-called variance-component score test. The test statistic

$$Q = (\mathbf{Y} - \hat{\mathbf{Y}}_0)^T \mathbf{K} (\mathbf{Y} - \hat{\mathbf{Y}}_0),$$

follows a mixture of χ^2 distributions which is approximated in the SKAT R-program by the so-called Davies method. Here $\hat{\mathbf{Y}}_0$ represents estimated phenotype under the null hypothesis. The matrix $\mathbf{K}$ contains genetic information and its elements are defined by the kernel function $K(\mathbf{G}_i, \mathbf{G}_{i'}) = \sum_{j=1}^{M} w_j G_{ij} G_{i'j}$, where w_j are the chosen weights. If the weights are all set to one, all variants would be assumed to have similar effect sizes on phenotype, as in burden methods. Wu et al. (2011) however, define the weights by a Beta-distribution, such that $w_j = Beta(\mathrm{MAF}_j; 1, 25)$. This definition increases the weight of rare variants and decreases weight of common variants.

An optimal SKAT method (SKAT-O) was proposed by Lee et al. (2012). This method incorporates a correlation effect between variants and was shown through a simulation study to perform equal to or better than the SKAT method in various scenarios.

The kernel-based methods are recently developed, yet popular in the literature. The methods are implemented in the R library SKAT and are simple to use. The SKAT methods were shown to perform better than several burden tests by Wu et al. (2011) and Lee et al. (2012) in power simulation studies. Based on these observations, Fan et al. (2013) chose to focus on comparing the functional linear model approach only with kernel-based approaches. However, we have not seen kernel or functional linear models compared to the CMC method. Fan et al. (2013) found through the same simulation studies as performed by Lee et al. (2012) that the functional linear model approaches had higher power than the kernel-based approaches.

6.3 Functional data methods

The use of functional linear models for association analysis was recently proposed by Fan et al. (2013). This approach is appropriate for scenarios including both causal rare and causal common variants. In addition, linkage information is utilized in the analysis, information which is according to Fan et al. (2013), not sufficiently taken into account by burden or kernel-based approaches.

Fan et al. (2013) consider approaches for association analysis based on the functional linear model

$$Y = \alpha_0 + \boldsymbol{\alpha}^T\mathbf{X} + \int_0^1 G(t)\beta(t)dt + \epsilon.$$

As is our chosen notation, Y represents a normally distributed phenotype and $\mathbf{X}_i$ represents non-genetic covariates. The inclusion of these variables allow for correction for population stratification and other confounders. We assume that the non-genetic covariates are non-functional while the genetic covariates are functional. Here, functionality of covariates refers to the mathematical principle as described in Section 3.1.2, and not the biological principle of functional genetic variants. The genetic information is assumed to be contained in the function $G(t)$, with $\beta(t)$ as the genetic effect function. Fan et al. (2013) consider t to be physical positions along the genome such that $G(t)$ is the genotype at position t. This is unnatural in the sense that the genome is not continuous, but t is. However, we will see in the following that we can relax this condition somewhat. The region of interest is scaled down to unit length such that $t \in [0, 1]$. Note that it is necessary to require that the *position* of each variant that is genotyped is known. This is generally unproblematic in genetic association studies.

Fan et al. (2013) consider three approaches for the modelling of genetic data; a standard approach as was described in Chapter 3, functional principal component analysis (FPCA), and a beta-smooth only approach. As the first and latter approaches prove to be more powerful than the FPCA approach, we will focus on these in the following.

Assume that for positions $t_1, \ldots, t_M$, the genotype $g(t_j)$ is known for each individual and that $g(t_j)$ takes on values 0, 1 or 2 according to the number of high-risk alleles at position t_j. For an additive genetic model we simply set $G(t_j) = g(t_j)$ to be the discrete observations of the function $G(t)$. For other genetic models, we define $G(t_j)$ as described previously.

Following the theory of functional linear models, the function $G(t)$ can be estimated from the discrete data $G(t_j)$ by a basis function and the simple linear smoother as defined in Equation (3.4). Further, $\beta(t)$ could be discretized by a basis function so that a linear regression model as in Equation (3.5) with additional terms for non-genetic covariates, is assumed. Fan et al. (2013) proposes this model and the use of the B-spline basis or Fourier basis. However, as there is no assumption of periodicity in the genetic data, it is unclear why the Fourier basis is considered as a relevant basis.

An alternative method proposed by Fan et al. (2013) is the *beta-smooth only* (β-SO) method. This model avoids the assumption of a continuous genotype function $G(t)$ by letting G be a non-functional variable. The model does however assume a continuous genetic effect function $\beta(t)$. This model is of the form

$$Y = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X} + \sum_{j=1}^{M} G(t_j)\beta(t_j) + \epsilon.$$

Again, the function $\beta(t)$ is expanded by a B-spline or Fourier basis. According to Fan et al. (2013), the beta-smooth only model performs similar to the full functional linear model as described above. This conclusion is based on extensive simulation studies. As this model is somewhat less technical and has an easier interpretation, we will use the β-SO method in our simulation studies.

Let $\theta_k(t)$ be the B-spline basis functions used for the expansion of $\beta(t)$; $\beta(t) = \sum_{k=1}^{K} b_k \theta_k(t)$. For M variants in a region, the linear model will be given as

$$\begin{aligned} Y &= \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X} + \sum_{j=1}^{M} G(t_j) \sum_{k=1}^{K} b_k \theta_k(t_j) + \epsilon \\ &= \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X} + \sum_{k=1}^{K} b_k \sum_{j=1}^{M} G(t_j)\theta_k(t_j) + \epsilon \\ &= \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X} + \mathbf{b}^T \mathbf{W} + \epsilon, \end{aligned}$$

where $\mathbf{b}$ and $\mathbf{W}$ are K-dimensional vectors, and the k'th element of $\mathbf{W}$ is defined by $\sum_{j=1}^{M} G(t_j)\theta_k(t_j)$.

The main goal of the article by Fan et al. (2013) is to test the null hypothesis of no association between phenotype and genotype with the highest possible power. Fan et al. (2013) focus on three scenarios. These scenarios are (1) all causal variants have a positive effect, (2) 20%/80% causal variants have a negative/positive effect, and (3) 50%/50% causal variants have a negative/positive effect. All three scenarios were tested for a mixture of causal rare and common variants in addition to a case with only causal rare variants.

The null hypothesis is given as $H_0 : \beta(t) = 0$ which is approximated by testing $b_1 = b_2 = \ldots = b_K = 0$, where b_k are the coefficients of the basis expansion and K is the number of basis functions used in the expansion of $\beta(t)$. We propose that by assuming a normal distribution of the phenotype, the score test as defined in Section 4.2.1, is a relevant test for this null hypothesis. Fan et al. (2013) use an F-test for the null hypothesis but refers to an article where the score test was used without any additional covariates. As described previously, we have defined the score test for use with additional non-genetic covariates.

6.4 Extreme phenotype sampling

According to Barnett et al. (2013), the presence of rare causal variants is enriched in the phenotypic extremes of a population. Consequently, extreme sampling can yield higher power to detect association compared with random sampling, in rare variant association studies.

If a phenotype is assumed to be normally distributed in a population, and an extreme sample is drawn, the conditional distribution describes the sampled data. Due to the large number of genetic variables in the locus, the conditional model as described in Section 4.2.2, is not appropriate. The rare variant methods as described above can be modified to account for the extreme selection criterion. We have not seen the missing genotype model that was proposed by Huang & Lin (2007) used in rare variant association studies. Often, simulation studies are used to compare performance of models, and when data are simulated as independent and normally distributed, the conditional model yields results as unbiased and powerful as the missing genotype model (Huang & Lin 2007). The missing genotype model assumes that the distributions of the genetic variables are known. These can be difficult to estimate when the genetic variable is a collapsed variable of a functional variable. Additionally, the expected genotype of a rare variant is close to zero which could cause numerical issues. We will continue the current work of using rare variant methods based on the conditional model. In future real data rare variant EPS studies, we hope to develop and investigate the relative performance of the missing genotype model.

The collapsing method (without covariates) for the EPS design was used by Li et al. (2011), who worked with the linear model

$$Y = \alpha_0 + \beta G + \epsilon,$$

where G is a collapsed variable as defined in Equation (6.1) and ϵ follows a $N(0, \sigma^2)$ distribution. Knowing the sampling cut-off values c_l and c_u, Li et al. (2011) assumed that the sampled data followed a conditional distribution with density function

$$f_{Y|Y \in \mathcal{C};G}(y; \alpha_0, \beta, \sigma) = \frac{\frac{1}{\sigma}\phi\left(\frac{y-\alpha_0-\beta G}{\sigma}\right)}{1 - \Phi\left(\frac{c_u-\alpha_0-\beta G}{\sigma}\right) + \Phi\left(\frac{c_l-\alpha_0-\beta G}{\sigma}\right)}.$$

The null hypothesis $H_0 : \beta = 0$ can be tested with the score test for the conditional model, as defined in Section 4.2.2. Li et al. (2011) showed that the EPS design yielded higher power to detect association compared to a random sampling design with the same sample size.

The extension of the CMC method into the EPS format has not been attempted. This extension is simple as we can assume a conditional model with genotype variables represented by collapsed variables. Inclusion of other covariates is unproblematic and the score test can be used to test for association between a locus and phenotype.

The SKAT and SKAT-O models were adapted for use in EPS studies by Barnett et al. (2013). The extension to the EPS framework was based on the conditional distribution of the sampled data. The authors provide an R program for the use of the SKAT methods in EPS studies. For extreme sampled continuous data, Barnett et al. (2013) compare the collapsing method, the SKAT method and the SKAT-O method. Through a simulation study Barnett et al. (2013) show that the collapsing method and the SKAT-O method have similar power to detect association when all variants have the same direction of effects. When, on the other hand, the variants have effects in opposite directions, all methods lose power, but SKAT-O is more robust than the collapsing method.

The functional linear model as proposed by Fan et al. (2013), has not yet been applied in an EPS study. The β-SO approach can be adjusted by assuming that the phenotype Y in the linear model

$$Y = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X} + \mathbf{b}^T \mathbf{W} + \epsilon,$$

follows the conditional distribution in the extreme sample, with density function

$$f_{Y|Y \in \mathcal{C};G}(y; \alpha_0, \boldsymbol{\alpha}, \mathbf{b}, \sigma) = \frac{\frac{1}{\sigma}\phi\left(\frac{y - \alpha_0 - \boldsymbol{\alpha}^T\mathbf{X} - \mathbf{b}^T\mathbf{W}}{\sigma}\right)}{1 - \Phi\left(\frac{c_u - \alpha_0 - \boldsymbol{\alpha}^T\mathbf{X} - \mathbf{b}^T\mathbf{W}}{\sigma}\right) + \Phi\left(\frac{c_l - \alpha_0 - \boldsymbol{\alpha}^T\mathbf{X} - \mathbf{b}^T\mathbf{W}}{\sigma}\right)}.$$

The score test for the conditional distribution can be used to test the null hypothesis $H_0 : \mathbf{b} = \mathbf{0}$. We will investigate the power of this method in the next chapter.

Chapter 7

A rare variant simulation study

In this chapter we will evaluate rare variant association methods both in cross-sectional and EPS studies. First we will compare the collapsing method, the CMC method, the SKAT and SKAT-O methods and the β-SO method for a cross-sectional design. Fan et al. (2013) showed that the β-SO method has a higher power than the kernel-based methods, while Lee et al. (2012) showed that the kernel-based methods were often more powerful than the collapsing method. The novelty of our study will be to include the CMC method and compare all methods in the same study. Following the design of Fan et al. (2013), we will consider cases where all causal variants are causal in the same directions, as well as opposite directions. Lee et al. (2012) worked only with causal rare variants, but we will consider regions with both causal rare and causal common variants.

Secondly, we will evaluate the five rare variant methods under the EPS design. We will compare the power of the methods to detect associations, as well as their power to detect associations in an EPS sample as opposed to a random sample of the same size. The novelty of this study is to test the EPS versions of the CMC and β-SO method in a simulation study, and compare them to kernel based methods and the collapsing method.

7.1 The simulated datasets

Simulation of genetic datasets for a given population can be performed by coalescent simulator tools. We have used the COSI simulator, developed by Schaffner et al. (2005). The novelty of the COSI program is that it comes with an out-of-Africa implementation that simulates the genetic development in humans from one source population (Africans) into European, Asian and West African populations. Genotypes are simulated for a 1 MB region which is referred to as a chromosome by Schaffner et al. (2005). This simulation is therefore the counterpart of genotyping *one* of the 23 pairs of chromosomes in a group of individuals. Motivated by the choice of simulation tool used by Lee et al. (2012) and Fan et al. (2013), we used the COSI simulator to simulate genotypes in the so-called European population. We simulated regions of length 1MB (mega base). The simulation program provides the positions of all mutations, the frequencies of the ancestral and mutations alleles at these sites. The program can also use these frequencies to simulate genotypes (0 or 1) in the specific regions for a chosen number of subjects. We simulated genotypes of 1MB regions in N European individuals by simulating genotypes (0 or 1) for $2N$ regions and simply adding two and two regions together. Thus, we have genotypes coded as 0, 1 and 2 for N individuals. As the SKAT models assume, we always set the most rare allele to be causal so that if the mutation allele is more common than the ancestral allele, the ancestral allele is seen as potentially causal for a phenotype. Thus, the MAF of any SNP was set to be the minimum of the frequency of the ancestral and mutation alleles. We created datasets with 500, 1000 and 2000 individuals.

The dataset with 500 individuals consists of genotype information on a 1MB region in which there are 9622 mutations. Of these mutations, 2535 are common variants, i.e. MAF ≥ 0.05. In the 1000 individuals dataset, there are 12412 mutations of which 2169 are common variants. In the 2000 individuals dataset, there are 14895 mutations of which 1993 are common variants. These datasets are simulated separately and are not subsets of each other.

7.1.1 Simulation of phenotypes

We followed the methods of Lee et al. (2012) and Fan et al. (2013) for creating non-genetic covariates, choosing causal loci and creating phenotypes.

Under the null hypothesis, phenotypes were generated as

$$Y = 0.5X_1 + 0.5X_2 + \epsilon,$$

where X_1 takes the values 0 and 1 with probability 0.5, while X_2 and ϵ are drawn from a $N(0,1)$ distribution. These simulated phenotypes allow us to test the type I error of the various models.

Under the alternative hypothesis, a region of length 3kB was randomly chosen and a certain fraction of the mutations in this region were chosen as causal for the phenotype. The fraction of causal mutations were set to 10%, 20% and 50%. We simulated three scenarios; one where all causal variants were causal in the same direction; one where 20% of causal variants were negatively associated with phenotype; and one scenario where 50% of causal variants were negatively associated with phenotype.

Let $t_1, \ldots, t_m$ be the positions of the causal mutations and let $g(t_j)$ denote the genotype at a position. As mentioned, we have simulated genotypes with coding 0, 1 and 2, corresponding to monotone genetic models. Phenotypes were generated under the alternative hypothesis as

$$Y = 0.5X_1 + 0.5X_2 + \beta_1 g(t_1) + \cdots + \beta_m g(t_m) + \epsilon,$$

where $|\beta_j|$ was defined as $c \cdot |\log(\text{MAF}_j)|/2$ with $c = \log(7), \log(5), \log(2)$ when 10%, 20% and 50% of mutations were causal, respectively. MAF_j refers to the MAF of the causal allele at position t_j. As explained, we varied between all β positive and certain mixtures of positive and negative effects.

For the EPS design, we simulated phenotypes under H_0 and H_1 as above and thereafter sampled the 25% lowest and highest phenotypes with corresponding covariates. In order to compare EPS to random sampling, we also randomly sampled 50% of subjects from the generated datasets. This yielded datasets of sizes 250, 500 and 1000 for the EPS and random sampling (RS) designs.

7.2 Methods

As methods for testing for an association between a phenotype and a loci, we used the models as presented in the previous chapter and thereafter the score tests for cross-sectional and conditional models, as presented in Chapter 4. We used our own implementations of the collapsing method, the CMC method and the β-SO method with corresponding score tests. Our implementation of the β-SO was based on the script by Fan et al. (2013) that is available at http://www.nichd.nih.gov/about/org/diphr/bbb/software/ under the heading "Functional Linear Models for Association Analysis of Quantitative Traits, by R. Fan and Y. Wang". We also used the kernel-based methods SKAT and SKAT-O which for a cross-sectional design are downloadable as the CRAN package SKAT. For the EPS design, the SKAT and SKAT-O methods implemented in R are part of the package CEPSKAT that we downloaded from http://www.hsph.harvard.edu/xlin/software.html. These methods return a p-value for the association test between genotypes and phenotypes. All implementations and use of methods are presented in Appendix D.

Parameters

Some parameters of the models were set before using the methods. For the CMC method, we used a threshold MAF of 0.05 below which all variants were collapsed and above which all variants were kept separate. For the β-SO model we used 15 basis functions of order four, as implemented by Fan et al. (2013).

7.3 Estimation of power and type I error rate

For the assessment of type I error rate, we simulated phenotypes as described above and repeated the simulations for 10000 different loci. A locus was defined as a randomly sampled 3 kB region which was chosen to be causal under H_1. For each of these iterations, a 3kB region was chosen as causal under H_1. For each model, we tested the null hypothesis of no association between the locus and the phenotype, and obtained a p-value by the score test for the burden models and functional linear model. The SKAT and SKAT-O models come with pre-programmed tests. The p-values are distributed on the interval $(0, 1)$. We define the outcome as a success if the p-value is below some significance level α such that after 10000 iterations we have a binomially distributed variable with parameter α and $n = 10000$, if tests preserve the type I error at level α. We estimate α by the MLE such that $\hat{\alpha} = n_{\text{successes}}/n$. The estimated parameter $\hat{\alpha}$ is the type I error frequency of the simulation. As the number of iterations increases, $\hat{\alpha}$ converges to α. However, for a limited number of iterations, we consider confidence intervals to asses the precision of the estimates. By the well-known normal approximation of the binomial, the confidence interval for α (Walpole et al. 2007, page 300) is given by

$$\left[\hat{\alpha} - z_{\alpha/2}\sqrt{\hat{\alpha}(1-\hat{\alpha})/n}, \hat{\alpha} - z_{\alpha/2}\sqrt{\hat{\alpha}(1-\hat{\alpha})/n}\right], \tag{7.1}$$

where $z_{\alpha/2}$ is the z-value from the standard normal distribution that leaves an area of $\alpha/2$ to the right. If the confidence interval for a given method, sample size and a chosen significance level α has a lower bound that exceeds α, the method is considered to not preserve type I error.

The power of a method is estimated by the fraction of p-values less than some significance level α over a total of $n = 2000$ simulated loci, as used by Fan et al. (2013), given that the there is an association between the phenotype and genotypes. In other words, the power of a method reflects the fraction of rejections of the null hypothesis, when the alternative is true. Following the simulation experiments of Fan et al. (2013), we simulated power for significance levels 0.05, 0.01 and 0.001.

7.4 Results

7.4.1 Type I error

Cross-sectional sample

For each complete dataset ($N = 500, 1000, 2000$) we simulated phenotypes under the null hypothesis and applied the following five models to test for an association between genotype and phenotype; (1) collapsing; (2) CMC; (3) SKAT; (4) SKAT-O; and (5) β-SO. For the collapsing, CMC and β-SO methods, we used the score test to find p-values. The SKAT and SKAT-O models come with pre-programmed tests and return corresponding p-values. We performed these simulations and tests a total of 10000 times. For significance levels $\alpha = 0.05, 0.01, 0.001$, we estimated type I error frequencies as defined above.

The resulting type I error rates for the cross-sectional study are presented in Table 7.1a. Confidence intervals for each estimate was calculated by Equation (7.1). Results in italic are those whose confidence interval's lower bound exceeded the significance level α. We see that type I error is preserved in all cases except for the SKAT-O method with significance level $\alpha = 0.05$ and sample size $N = 500$.

Extreme phenotype sample

For each of the datasets ($N = 500, 1000, 2000$) we simulated phenotypes under the null hypothesis and thereafter extracted the subjects with phenotypes in the lower or upper quartiles. The five models adapted for the EPS design were used to test for association between a 3 kB region and the phenotype. Type I error was estimated as described above.

The resulting type I error rates for the extreme phenotype samples are presented in Table 7.1b. Results in italic are those whose confidence interval's lower bound exceeded the significance level α. We see that the collapsing method does not preserve type I error for any sample size or significance level. Additionally, the kernel-based methods have problems with preserving type I error for low sample sizes. We find that the CMC and β-SO methods consistently preserve type I error in these simulations.

Extreme phenotype sampling versus random sampling

If only half of a group of individuals can be genotyped, one can either choose to sample the extreme ends of the phenotype spectrum, or to sample randomly. We will compare these two methods and therefore present type I error rates for random sampling from the datasets of sizes 500, 1000 and 2000 that were simulated under H_1. The tests used are the same as those used in a cross-sectional study. The type I error rates for the five models are presented in Table 7.1c. All methods and sample sizes preserve type I error for the given

		Model				
α	N	Collapse	CMC	SKAT	SKAT-O	β-SO
	500	0.0416	0.0304	0.0500	*0.0558*	0.0405
0.05	1000	0.0434	0.0324	0.0528	0.0538	0.0524
	2000	0.0444	0.0392	0.0503	0.0528	0.0512
	500	0.0072	0.0070	0.0098	0.0118	0.0072
0.01	1000	0.0087	0.0062	0.0093	0.0112	0.0092
	2000	0.0076	0.0070	0.0096	0.0115	0.0107
	500	0.0006	0.0008	0.0009	0.0009	0.0007
0.001	1000	0.0008	0.0004	0.0007	0.0015	0.0006
	2000	0.0010	0.0006	0.0010	0.0012	0.0005

(a) Estimated type I error rates for cross-sectional studies.

		Model				
α	$N/2$	Collapse	CMC	SKAT	SKAT-O	β-SO
0.05	250	*0.1397*	0.0287	*0.0547*	*0.0612*	0.0369
	500	*0.0898*	0.0294	*0.0561*	*0.0589*	0.0464
	1000	*0.0993*	0.0360	0.0542	*0.0591*	0.0453
0.01	250	*0.1114*	0.0045	0.0121	*0.0143*	0.0074
	500	*0.0533*	0.0047	0.0105	0.0125	0.0075
	1000	*0.0632*	0.0070	0.0105	0.0113	0.0084
0.001	250	*0.1040*	0.0005	0.0006	0.0008	0.0006
	500	*0.0462*	0.0007	0.0014	0.0014	0.0004
	1000	*0.0556*	0.0005	0.0011	0.0013	0.0004

(b) Estimated type I error rates for EPS studies.

		Model				
α	$N/2$	Collapse	CMC	SKAT	SKAT-O	β-SO
0.05	250	0.0411	0.0263	0.0501	0.0502	0.0352
	500	0.0489	0.0305	0.0480	0.0494	0.0450
	1000	0.0439	0.0365	0.0500	0.0525	0.0481
0.01	250	0.0081	0.0046	0.0088	0.0102	0.0052
	500	0.0095	0.0065	0.0088	0.0101	0.0071
	1000	0.0090	0.0075	0.0100	0.0101	0.0099
0.001	250	0.0007	0.0005	0.0009	0.0008	0.0003
	500	0.0009	0.0004	0.0009	0.0011	0.0007
	1000	0.0006	0.0004	0.0009	0.0010	0.0005

(c) Estimated type I error rates for random sampling studies.

Table 7.1: Estimated type I error rates

significance levels. Thus it seems as if the random sampling method gives better results than the extreme phenotype sampling method, when we consider the collapsing method and the kernel-based methods. The CMC and β-SO methods have similar performance for both the cross-sectional and the EPS design.

7.4.2 Power estimates under the alternative hypothesis

For power simulations we replicated the design of Lee et al. (2012) and Fan et al. (2013) and drew 2000 regions of 3 kB length in which some variants were chosen as causal. The phenotypes were simulated as described above. We let the regions contain 10%, 20% and 50% causal variants of which none, 20% or 50% were negatively causal (protective) for the phenotype.

Cross-sectional sample

For each of the datasets ($N = 500, 1000, 2000$) we simulated phenotypes under H_1 and used the five methods to test for an association. Power was estimated as the fraction of p-values that fell below the threshold α, i.e. the fraction of associations found when an association was always present. Estimated power for the five methods and for all scenarios, are presented in Tables 7.2, 7.3 and 7.4. Some selected results are also presented in Figure 7.1.

We see that the collapsing method consistently displays low power to detect the association. When few of the variants (10%) are causal, the β-SO method outperforms the CMC and the kernel-based methods. These differences even out as sample size increases and the number of causal variants increase. For the burden methods (the collapsing and CMC method) this is in line with the problem of misclassification as discussed by Lee et al. (2012). However, as argued by Li & Leal (2008), it is clear that the CMC method is more robust towards misclassification than the collapsing method.

We expect to see that the burden methods suffer the most when variants can be causal in both directions, and this is in fact reflected by the simulations. The collapsing method is clearly not appropriate for use in such instances, but the CMC method can not be written off. It is also clear that although the kernel-based method were developed partly to solve this issue, they are not as robust as the β-SO method towards causality on both directions.

As an example, consider sample size $N = 1000$ and significance level $\alpha = 0.001$. For only positively causal variants and $(10\%, 20\%, 50\%)$ causal variants, the power of the CMC method is estimated as $(0.72, 0.88, 0.91)$, while for the β-SO method the power is $(0.90, 0.98, 0.97)$. These observations illustrate the robustness of the β-SO method. The corresponding power estimations when 20% of causal variants are negatively causal are exactly the same for both of these methods. When 50% are negatively causal, the power estimates are $(0.62, 0.80, 0.76)$ for the CMC method and $(0.86, 0.97, 0.92)$ for the β-SO method. Here we see a general loss of power due to more negative variants. It appears as if the loss of power due to many negatively causal variants is stronger than the gain in power due to less misclassification when many variants are in fact causal.

Extreme phenotype sample

In order to test the EPS design, we simulated phenotypes for the full samples ($N = 500, 1000, 2000$) and extracted the upper and lower quartiles ($N/2 = 250, 500, 1000$). Tests were performed on these extreme datasets and power was estimated as described above. The results are presented in Tables 7.5, 7.6 and 7.7. Some results are illustrated in Figure 7.2 as well.

We examine the results further by considering a similar example as above. Consider a sample size of $N/2 = 500$, and $\alpha = 0.001$. If none of the variants are negatively causal, the power of the CMC method is estimated as $(0.73, 0.92, 0.96)$, while for the β-SO method power is $(0.87, 0.97, 0.96)$, for $(10\%, 20\%, 50\%)$ causal variants in the locus. We see that the methods are comparable when the degree of misclassification decreases. The corresponding power estimates are $(0.67, 0.88, 0.91)$ and $(0.80, 0.94, 0.89)$ for the CMC and β-SO methods when 20% of causal variants are negatively causal, and $(0.64, 0.84, 0.82)$ and $(0.78, 0.92, 0.82)$, when 50% of causal variants are negatively causal. Again, we see that when all variants are causal in the same direction, power is highest when many variants (50%) in the loci are causal. When some variants are causal in the opposite direction, we see that the methods are most powerful when only 20% of variants are causal.

Extreme phenotype sampling versus random sampling

As well as extracting the upper and lower quartiles of the full samples for the EPS design, we randomly sampled half of the full samples for a smaller cross-sectional design. For significance level $\alpha = 0.01$, the results from the EPS and the random sampling (RS) designs are presented in Tables 7.8, 7.9 and 7.10. Some of these results are also illustrated in Figure 7.3.

When all causal variants are causal in the same direction, the extreme phenotype sampling method has a higher power than the random sampling for all methods. This is in line with the theory that more information on causal rare variants is found in the extreme phenotype subjects. When 20% of causal variants are negatively causal, the differences are not as striking. It seems that when there are few causal variants, i.e. a high degree of misclassification, and the sample sizes are low, random samples have similar or higher power to detect the association than the EPS design. We observe similar results when 50% of causal variants are negatively associated with the phenotype.

7.5 Discussion and conclusion

We observe that the β-SO is robust for all variations of sampling and causality and consistently more powerful than the four other methods. Lee et al. (2012) claimed that the kernel-based methods outperformed burden methods, but we see that the CMC method is often more powerful than both the SKAT and SKAT-O methods and in some cases the CMC and β-SO methods have very similar power. We included the collapsing method for reference to more simple burden methods and it is clear that this method is too crude to be useful in these studies. Surprisingly, we saw that the collapsing method in the EPS design found false associations when there were none in the type I error rate simulations, but that this methods failed to discover enough associations when an association was present. We have used a simple implementation of the CMC method where we have set a threshold to 0.05. This method could be improved upon by tuning this threshold based on known information on variants in specific loci. We note that the CMC method requires known MAFs, or at least whether variants are common or rare, for all genotyped positions in the loci. The β-SO method assumes that the positions of the variants are known. The SKAT methods do not require any other input than the genotypes themselves. The SKAT methods are therefore widely applicable, but this could also cause these methods to be less powerful than other methods that use additional information.

Due to the power and simple implementation of the β-SO method, we would recommend the use of this method in both cross-sectional and extreme phenotype studies. However, as we learnt in Chapter 5, although the EPS method are powerful in theory, extreme samples can be inaccurate when all assumptions are not met, assumptions for which random samples are more robust. We saw that the missing genotype model was preferable in such instances. As further work on this topic we suggest to look into an adaptation of the missing genotype method to rare variant EPS studies. For the missing genotype model to be used with the β-SO approach, the distribution of the variable $\mathbf{W}$, defined in Chapter 6.3, must be known. The parameters of this distribution do not have to be known as they can be estimated by maximum likelihood. The score test statistic could thereafter be found using similar procedures as described previously.

α	Sample size	% causal	Collapse	CMC	SKAT	SKAT-O	β-SO
0.05	500	10	0.31	0.82	0.84	0.83	0.95
		20	0.43	0.94	0.95	0.95	0.99
		50	0.50	0.97	0.91	0.95	0.98
	1000	10	0.51	0.81	0.90	0.90	0.95
		20	0.62	0.93	0.98	0.98	0.99
		50	0.65	0.97	0.98	0.98	0.99
	2000	10	0.48	0.82	0.94	0.94	0.97
		20	0.66	0.96	0.99	0.99	1.00
		50	0.77	1.00	0.99	1.00	1.00
0.01	500	10	0.22	0.77	0.74	0.74	0.91
		20	0.33	0.92	0.88	0.89	0.98
		50	0.41	0.95	0.81	0.90	0.95
	1000	10	0.43	0.76	0.84	0.84	0.93
		20	0.55	0.91	0.95	0.94	0.99
		50	0.57	0.95	0.94	0.95	0.98
	2000	10	0.41	0.77	0.88	0.88	0.95
		20	0.60	0.95	0.97	0.98	0.99
		50	0.71	1.00	0.98	0.99	1.00
0.001	500	10	0.15	0.72	0.62	0.61	0.86
		20	0.25	0.89	0.79	0.80	0.96
		50	0.30	0.87	0.64	0.78	0.87
	1000	10	0.36	0.72	0.76	0.76	0.90
		20	0.47	0.88	0.90	0.90	0.98
		50	0.49	0.91	0.87	0.90	0.97
	2000	10	0.35	0.73	0.80	0.80	0.92
		20	0.54	0.92	0.94	0.95	0.98
		50	0.64	0.99	0.96	0.99	1.00

Table 7.2: Estimated power under the cross-sectional design when all causal variants are causal in the positive direction, causal variants are both rare and common.

α	Sample size	% causal	Collapse	CMC	SKAT	SKAT-O	β-SO
0.05	500	10	0.30	0.80	0.85	0.83	0.93
		20	0.35	0.91	0.94	0.93	0.99
		50	0.35	0.93	0.89	0.89	0.97
	1000	10	0.51	0.81	0.90	0.90	0.95
		20	0.62	0.93	0.98	0.98	0.99
		50	0.65	0.97	0.98	0.98	0.99
	2000	10	0.43	0.78	0.94	0.93	0.95
		20	0.57	0.94	0.99	0.99	0.99
		50	0.62	0.98	0.99	0.99	1.00
0.01	500	10	0.22	0.76	0.74	0.74	0.90
		20	0.26	0.88	0.87	0.86	0.97
		50	0.25	0.86	0.77	0.79	0.93
	1000	10	0.43	0.76	0.84	0.84	0.93
		20	0.55	0.91	0.95	0.94	0.99
		50	0.57	0.95	0.94	0.95	0.98
	2000	10	0.37	0.73	0.88	0.88	0.93
		20	0.48	0.91	0.97	0.97	0.99
		50	0.53	0.97	0.97	0.98	1.00
0.001	500	10	0.15	0.71	0.64	0.63	0.84
		20	0.18	0.84	0.77	0.76	0.95
		50	0.16	0.74	0.58	0.63	0.81
	1000	10	0.36	0.72	0.76	0.76	0.90
		20	0.47	0.88	0.90	0.90	0.98
		50	0.49	0.91	0.87	0.90	0.97
	2000	10	0.31	0.69	0.81	0.81	0.90
		20	0.42	0.88	0.93	0.93	0.98
		50	0.45	0.95	0.94	0.95	0.99

Table 7.3: Estimated power under the cross-sectional design when 80%/20% of causal variants are causal in the positive/negative direction, causal variants are both rare and common.

α	Sample size	% causal	Collapse	CMC	SKAT	SKAT-O	β-SO
0.05	500	10	0.20	0.75	0.82	0.78	0.92
		20	0.21	0.84	0.93	0.90	0.98
		50	0.15	0.84	0.85	0.82	0.95
	1000	10	0.38	0.70	0.89	0.87	0.93
		20	0.47	0.86	0.97	0.97	0.99
		50	0.38	0.89	0.96	0.95	0.98
	2000	10	0.35	0.70	0.93	0.91	0.94
		20	0.48	0.86	0.99	0.99	0.99
		50	0.42	0.95	0.99	0.99	1.00
0.01	500	10	0.12	0.70	0.71	0.69	0.88
		20	0.14	0.81	0.85	0.83	0.96
		50	0.08	0.75	0.72	0.67	0.87
	1000	10	0.30	0.66	0.81	0.80	0.90
		20	0.40	0.83	0.94	0.93	0.98
		50	0.28	0.83	0.91	0.90	0.97
	2000	10	0.29	0.66	0.87	0.85	0.91
		20	0.41	0.84	0.97	0.96	0.98
		50	0.31	0.93	0.97	0.96	0.99
0.001	500	10	0.07	0.66	0.60	0.59	0.82
		20	0.08	0.76	0.74	0.73	0.93
		50	0.04	0.62	0.53	0.49	0.75
	1000	10	0.24	0.62	0.72	0.72	0.86
		20	0.31	0.80	0.89	0.88	0.97
		50	0.18	0.76	0.83	0.81	0.92
	2000	10	0.23	0.63	0.78	0.77	0.88
		20	0.33	0.81	0.93	0.92	0.97
		50	0.22	0.90	0.93	0.92	0.99

Table 7.4: Estimated power under the cross-sectional design when 50%/50% of causal variants are causal in the positive/negative direction, causal variants are both rare and common.

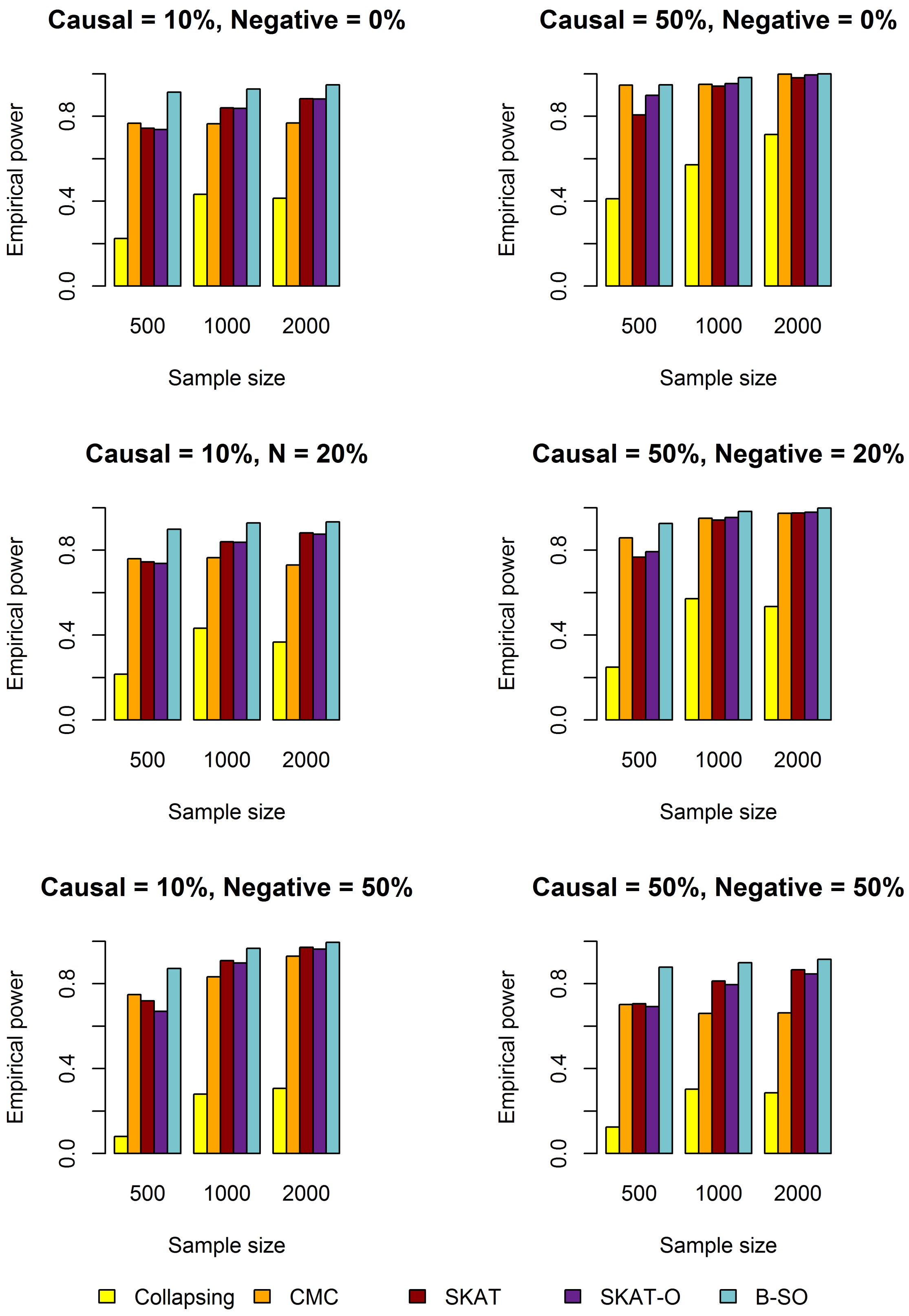


Figure 7.1: Power estimates for a cross-sectional study with $\alpha = 0.01$.

α	Sample size	% causal	Collapse	CMC	SKAT	SKAT-O	β-SO
0.05	250	10	0.38	0.81	0.81	0.80	0.92
		20	0.51	0.94	0.92	0.92	0.98
		50	0.56	0.94	0.84	0.91	0.93
	500	10	0.58	0.83	0.91	0.91	0.95
		20	0.74	0.97	0.98	0.98	0.99
		50	0.83	0.99	0.98	0.99	1.00
	1000	10	0.52	0.82	0.95	0.94	0.97
		20	0.70	0.97	0.99	0.99	1.00
		50	0.79	1.00	0.99	1.00	1.00
0.01	250	10	0.30	0.75	0.68	0.66	0.87
		20	0.41	0.91	0.82	0.83	0.95
		50	0.46	0.88	0.68	0.82	0.86
	500	10	0.52	0.78	0.83	0.83	0.92
		20	0.68	0.94	0.95	0.95	0.98
		50	0.77	0.99	0.94	0.97	0.99
	1000	10	0.45	0.78	0.89	0.89	0.94
		20	0.64	0.95	0.97	0.98	0.99
		50	0.73	1.00	0.97	0.99	1.00
0.001	250	10	0.23	0.68	0.52	0.51	0.79
		20	0.32	0.85	0.68	0.69	0.91
		50	0.35	0.77	0.46	0.64	0.73
	500	10	0.45	0.73	0.73	0.74	0.87
		20	0.61	0.92	0.89	0.91	0.97
		50	0.70	0.96	0.86	0.94	0.96
	1000	10	0.38	0.73	0.80	0.80	0.91
		20	0.57	0.93	0.94	0.95	0.98
		50	0.65	0.99	0.93	0.98	0.99

Table 7.5: Estimated power under the EPS design when all causal variants are causal in the positive direction, causal variants are both rare and common.

α	Sample size	% causal	Collapse	CMC	SKAT	SKAT-O	β-SO
0.05	250	10	0.38	0.75	0.67	0.65	0.84
		20	0.47	0.91	0.82	0.82	0.94
		50	0.52	0.88	0.72	0.85	0.86
	500	10	0.60	0.81	0.81	0.81	0.92
		20	0.73	0.96	0.94	0.95	0.98
		50	0.81	0.98	0.93	0.98	0.98
	1000	10	0.53	0.85	0.85	0.86	0.95
		20	0.70	0.97	0.97	0.97	0.99
		50	0.77	1.00	0.96	0.99	1.00
0.01	250	10	0.29	0.65	0.51	0.51	0.75
		20	0.36	0.84	0.68	0.71	0.90
		50	0.42	0.78	0.52	0.71	0.74
	500	10	0.50	0.74	0.71	0.70	0.87
		20	0.66	0.93	0.88	0.89	0.97
		50	0.74	0.95	0.86	0.95	0.95
	1000	10	0.44	0.80	0.76	0.76	0.91
		20	0.62	0.95	0.93	0.94	0.98
		50	0.71	0.99	0.92	0.98	0.99
0.001	250	10	0.22	0.53	0.33	0.34	0.61
		20	0.28	0.71	0.50	0.54	0.79
		50	0.30	0.62	0.34	0.51	0.58
	500	10	0.42	0.67	0.59	0.60	0.80
		20	0.58	0.88	0.78	0.82	0.94
		50	0.65	0.91	0.72	0.87	0.89
	1000	10	0.36	0.74	0.65	0.66	0.85
		20	0.55	0.92	0.85	0.88	0.97
		50	0.61	0.97	0.85	0.94	0.97

Table 7.6: Estimated power under the EPS design when 80%/20% of causal variants are causal in the positive/negative direction, causal variants are both rare and common.

α	Sample size	% causal	Collapse	CMC	SKAT	SKAT-O	β-SO
0.05	250	10	0.33	0.72	0.65	0.64	0.83
		20	0.41	0.87	0.79	0.79	0.92
		50	0.42	0.79	0.61	0.73	0.78
	500	10	0.51	0.79	0.81	0.79	0.91
		20	0.65	0.93	0.92	0.93	0.98
		50	0.73	0.95	0.89	0.94	0.96
	1000	10	0.47	0.83	0.84	0.83	0.94
		20	0.62	0.96	0.96	0.96	0.99
		50	0.68	0.98	0.95	0.98	0.99
0.01	250	10	0.23	0.61	0.50	0.48	0.73
		20	0.31	0.78	0.66	0.67	0.86
		50	0.32	0.65	0.42	0.57	0.63
	500	10	0.43	0.72	0.70	0.70	0.85
		20	0.58	0.90	0.87	0.88	0.96
		50	0.64	0.90	0.79	0.88	0.91
	1000	10	0.38	0.78	0.75	0.74	0.89
		20	0.53	0.94	0.91	0.92	0.98
		50	0.58	0.96	0.88	0.95	0.97
0.001	250	10	0.18	0.49	0.33	0.32	0.59
		20	0.23	0.65	0.48	0.51	0.74
		50	0.22	0.46	0.26	0.37	0.44
	500	10	0.34	0.64	0.57	0.57	0.78
		20	0.48	0.84	0.78	0.80	0.92
		50	0.52	0.82	0.65	0.76	0.82
	1000	10	0.30	0.71	0.63	0.63	0.85
		20	0.44	0.90	0.83	0.85	0.96
		50	0.47	0.92	0.77	0.88	0.94

Table 7.7: Estimated power under the EPS design when 50%/50% of causal variants are causal in the positive/negative direction, causal variants are both rare and common.

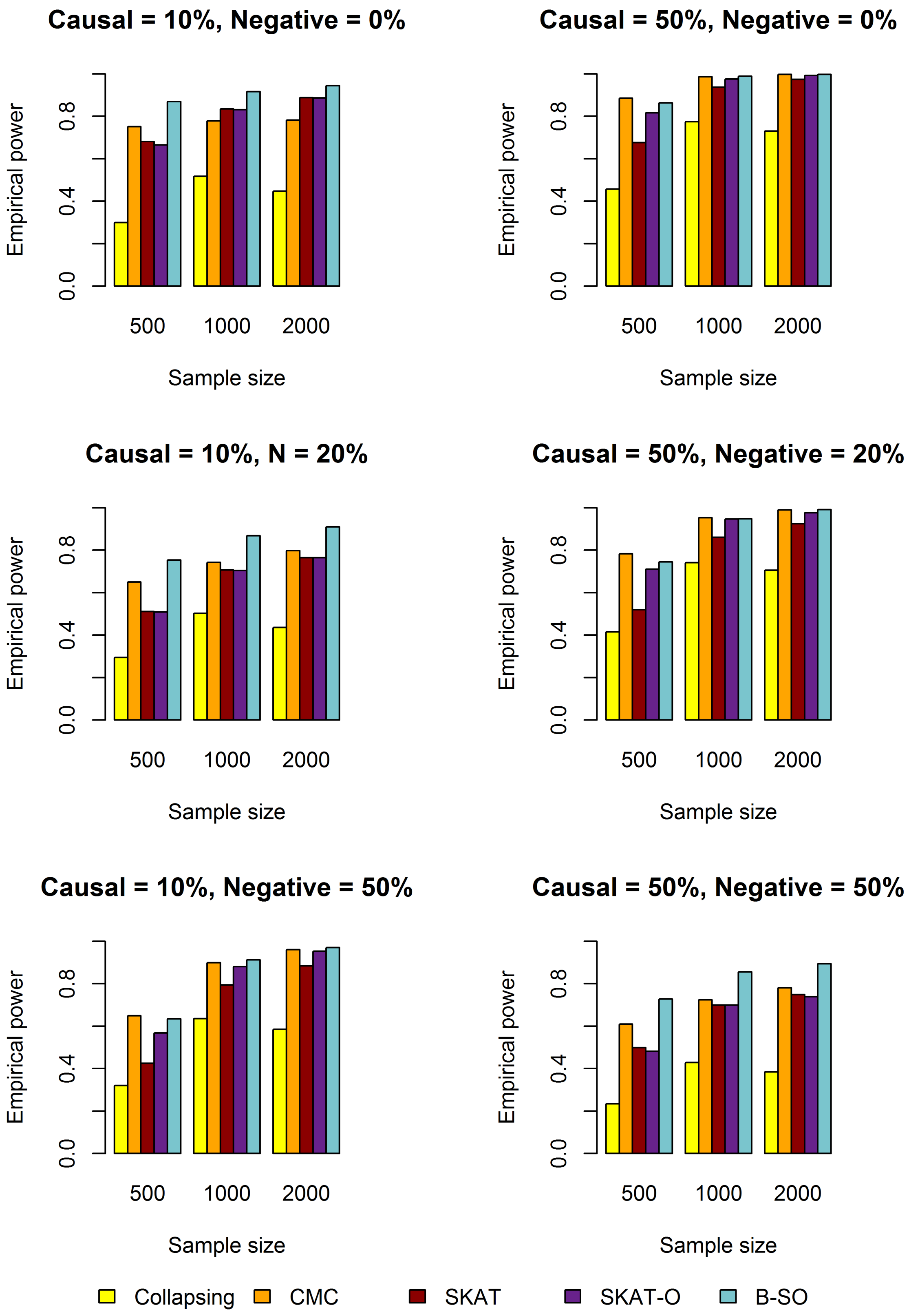


Figure 7.2: Power estimates for an extreme phenotype study with $\alpha = 0.01$.

α	Sample size	% causal	Method	Collapse	CMC	SKAT	SKAT-O	β-SO
0.01	250	10	EPS	0.30	0.75	0.68	0.66	0.87
			RS	0.12	0.64	0.59	0.58	0.78
		20	EPS	0.41	0.91	0.82	0.83	0.95
			RS	0.23	0.82	0.74	0.74	0.90
		50	EPS	0.46	0.88	0.68	0.82	0.86
			RS	0.26	0.74	0.52	0.66	0.71
	500	10	EPS	0.52	0.78	0.83	0.83	0.92
			RS	0.38	0.71	0.74	0.74	0.87
		20	EPS	0.68	0.94	0.95	0.95	0.98
			RS	0.57	0.91	0.90	0.90	0.97
		50	EPS	0.77	0.99	0.94	0.97	0.99
			RS	0.66	0.95	0.83	0.90	0.95
	1000	10	EPS	0.45	0.78	0.89	0.89	0.94
			RS	0.33	0.71	0.78	0.77	0.89
		20	EPS	0.64	0.95	0.97	0.98	0.99
			RS	0.51	0.92	0.93	0.94	0.98
		50	EPS	0.73	1.00	0.97	0.99	1.00
			RS	0.61	0.99	0.93	0.97	0.99

Table 7.8: Estimated power under the EPS design and the cross-sectional based on a small RS (random sample). All causal variants are causal in the positive direction, causal variants are both rare and common.

α	Sample size	% causal	Method	Collapse	CMC	SKAT	SKAT-O	β-SO
0.01	250	10	EPS	0.29	0.65	0.51	0.51	0.75
			RS	0.13	0.64	0.58	0.56	0.78
		20	EPS	0.36	0.84	0.68	0.71	0.90
			RS	0.16	0.76	0.70	0.69	0.90
		50	EPS	0.42	0.78	0.52	0.71	0.74
			RS	0.15	0.60	0.48	0.52	0.65
	500	10	EPS	0.50	0.74	0.71	0.70	0.87
			RS	0.35	0.71	0.74	0.74	0.88
		20	EPS	0.66	0.93	0.88	0.89	0.97
			RS	0.44	0.85	0.88	0.87	0.96
		50	EPS	0.74	0.95	0.86	0.95	0.95
			RS	0.44	0.84	0.79	0.81	0.89
	1000	10	EPS	0.44	0.80	0.76	0.76	0.91
			RS	0.29	0.68	0.75	0.74	0.88
		20	EPS	0.62	0.95	0.93	0.94	0.98
			RS	0.38	0.85	0.92	0.92	0.97
		50	EPS	0.71	0.99	0.92	0.98	0.99
			RS	0.42	0.94	0.90	0.91	0.97

Table 7.9: Estimated power under the EPS design and the cross-sectional based on a small RS (random sample). 20%/80% of causal variants are causal in the positive/negative direction, causal variants are both rare and common.

α	Sample size	% causal	Method	Collapse	CMC	SKAT	SKAT-O	β-SO
0.01	250	10	EPS	0.23	0.61	0.50	0.48	0.73
			RS	0.06	0.57	0.58	0.54	0.75
		20	EPS	0.31	0.78	0.66	0.67	0.86
			RS	0.08	0.69	0.69	0.66	0.86
		50	EPS	0.32	0.65	0.42	0.57	0.63
			RS	0.04	0.44	0.44	0.39	0.56
	500	10	EPS	0.43	0.72	0.70	0.70	0.85
			RS	0.22	0.61	0.70	0.69	0.84
		20	EPS	0.58	0.90	0.87	0.88	0.96
			RS	0.28	0.76	0.85	0.83	0.93
		50	EPS	0.64	0.90	0.79	0.88	0.91
			RS	0.19	0.67	0.73	0.69	0.83
	1000	10	EPS	0.38	0.78	0.75	0.74	0.89
			RS	0.24	0.63	0.75	0.73	0.88
		20	EPS	0.53	0.94	0.91	0.92	0.98
			RS	0.29	0.81	0.90	0.89	0.97
		50	EPS	0.58	0.96	0.88	0.95	0.97
			RS	0.19	0.84	0.90	0.87	0.96

Table 7.10: Estimated power under the EPS design and the cross-sectional based on a small RS (random sample). 50%/50% of causal variants are causal in the positive/negative direction, causal variants are both rare and common.

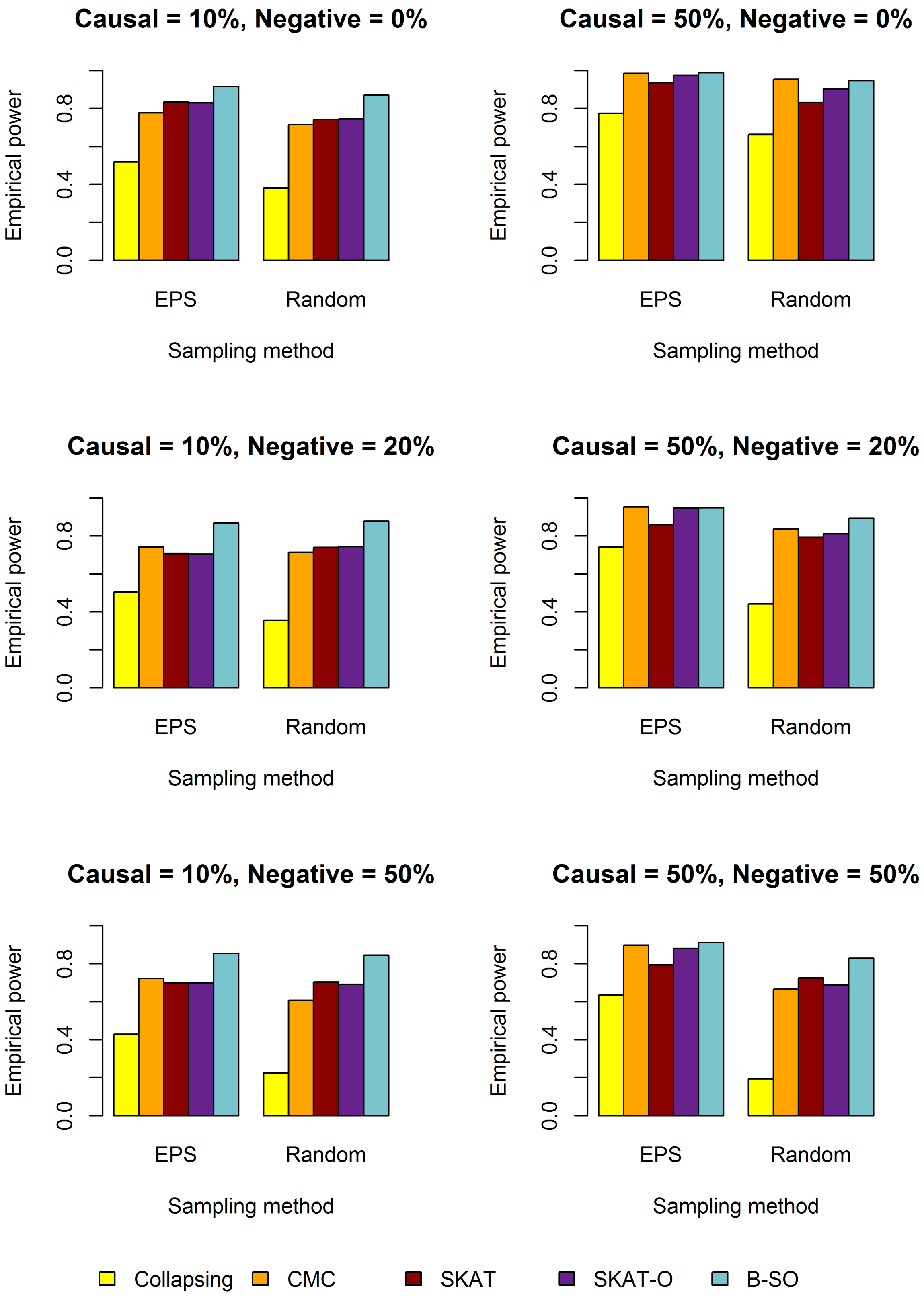


Figure 7.3: Power estimates for an extreme phenotype and cross-sectional study based on a small random sample. Sample size 500, $\alpha = 0.01$.

Chapter 8

Conclusion and discussion

Conclusion

The extreme phenotype design is theoretically considered to be more powerful than a cross-sectional design for discovering associations between loci and phenotypes. Extreme phenotype sampling enriches the presence of causal variants in a dataset of a certain size, and this increases the probability of discovering associations. This enrichment can be argued to be particularly important in rare variant association studies. In this thesis we have investigated the extreme phenotype sampling design and relevant statistical models for both common and rare variant association studies. We have developed statistical methods that can be used to test for a genetic effect, and have applied our methods in a common variant study based on a dataset from HUNT, and a rare variant study based on a simulated dataset from COSI. We have seen that the conditional model for the extreme phenotype sampling design has limitations for real data analysis. It is probable that these limitations occur because the model only incorporates data from the tails of a distribution. The missing genotype model can be more accurate as it uses available information from the entire distribution of a phenotype. We have verified that the CMC and β-SO methods are powerful methods in cross-sectional rare variant association studies, and shown that these methods can be adapted successfully to the extreme phenotype sampling design. Through simulation studies we have seen that the EPS design is often more powerful than the cross-sectional design, but that there are situations in which the cross-sectional design results in similar or higher power to detect associations. Based on common and rare association studies we cannot be certain that the EPS design in general, and the conditional model in particular, are advisable for use in genetic association studies. The methods are theoretically powerful, but may cause spurious results due to sensitivity to model misspecification and violations of model assumptions.

Discussion

At first glance, the HUNT dataset that was used for the study of waist-hip ratio seems to fit the extreme phenotype sampling design perfectly. Waist-hip ratio was calculated for thousands of individuals and the upper and lower quartiles of men and women were sampled for genotyping. The SNPs have known distributions and the cut-off values for sampling are known. However, the distribution of WHR in the population is not perfectly normal distributed and neither is the log transform nor other relevant transforms that we investigated. Additionally it appears as if waist-hip ratio has different variance in different subgroups of the population such as age, smoke and exercise groups. Due to lack of information we cannot check whether sampled individuals are statistically independent, but based on knowledge of the population of Nord-Trøndelag we suggest that population stratification may have occurred and that individuals are therefore not independent. When we fitted the conditional model to the data and saw that, compared to a multiple linear regression model, it over- and underestimated the effects of causal variables, we could not be certain as to which of these model violations - if any - caused the poor fit. The missing genotype model is able to use more information on subjects because non-genetic variables are known for all individuals in HUNT. It was therefore not surprising that the missing genotype model estimated the effects of non-genetic covariates (age, smoke, exercise) more accurately compared to a multiple linear regression model.

In our rare variant simulation studies, the data are by construction independent and normally distributed with constant variance. We used these studies to compare different methods to each other, as well as comparing the power of extreme phenotype sampling with random sampling. We found that the β-SO method is consistently better than other methods. We also showed that the extreme sampling design can be less powerful than the cross-sectional design when sample sizes are low, few variants in the loci are causal, and causal variants are both positively and negatively associated with phenotypes. Based on these results, and what we saw in the HUNT study, we are interested in further investigating the suitability of the extreme phenotype sampling design and evaluate if it is in fact preferable to sample extreme individuals as opposed to individuals from the full range of a phenotype. Perhaps, as suggested by Li et al. (2011), the extreme phenotype sampling design can be used as an initial screening in a two-stage design. In this type of design, the idea is to genotype few extreme individuals in order to detect loci that seem to be associated with a phenotype. In the second stage, these loci are genotyped in a larger cross-sectional sample. Effect estimates of the different loci are found using the cross-sectional sample.

We should also mention that in our simulation studies, we worked with the definition of extreme phenotypes as the upper and lower quartiles of a phenotype based on a dataset of sizes $N = 500, 1000, 2000$. For larger datasets, but similar financial or logistical restrictions, it would be necessary to consider a stricter bound for extreme phenotypes.

What is new in this thesis?

In this thesis we have developed a score test statistic for the conditional model that is flexible with respect to the number of covariates and nuisance parameters. This test can be used in GWA studies where genetic covariates are genotypes of various SNPs. The test can also be used in rare variant association studies where genetic covariates are collapsed discrete variables of functional variables. We have developed a score test statistic for the missing genotype model that was customized to a GWAS study with known distributions and minor allele frequencies of genetic variables.

Huang & Lin (2007) showed that the conditional model and missing genotype model have similarly low bias and high power to estimate and discover genetic association with phenotype in the special case where one only genetic covariate and no non-genetic covariates were present. We have had the opportunity in this thesis to test extended versions of these methods in a large dataset from HUNT and have discovered that the power and bias of the conditional model might be sensitive to violations of assumptions in real data analysis. The missing genotype model, which uses more of the available information, appears to be estimate effects more accurately in the sense that the estimates resemble those of a multiple linear regression model.

For rare variant association studies, we have adapted two important rare variant methods to the extreme phenotype sampling design, namely the CMC (Li & Leal 2008) and β-SO (Fan et al. 2013) method. In an extensive simulation study, we have evaluated five rare variant association methods; (1) the collapsing method, (2) the CMC method, (3) the SKAT method, (4) the SKAT-O method, and (5) the β-SO method, both for cross-sectional and extreme phenotype sampling studies. Through our studies, we found that the β-SO method is consistently more powerful than the other methods, both in cross-sectional and EPS studies. We also compared the extreme sampling design to a random sampling design using all the aforementioned methods. We discovered that the EPS design is not always more powerful than the cross-sectional design.

Future work

We hope to build on this thesis and write two separate articles on extreme phenotype sampling; one concerning common variant association studies and the other concerning rare variant association studies.

For common variants we hope to acquire a full GWAS dataset in which we can control for population stratification and where genotypes are known for a complete sample. In that way, we can fit the data to a linear regression model and in addition synthetically sample extreme phenotype subjects. We can fit the conditional model and the missing genotype model to this extreme sample and compare results to the model fitted to the complete sample. In that way, we can compare the accuracy of the two extreme sampling models. This resembles what we aimed to do with the HUNT dataset, but as we could not estimate genetic effects for a complete sample, it was not clear which model gave the most reliable estimates. It would also be preferable to look further into the effect of model violations, especially in regard to the conditional model as this is used for rare variant studies as well. In a dataset where we have more control over the distribution, independence and variance of the subjects, we could analyse this issue further. It could also be possible to perform a simulation study where one assumption at a time is intentionally violated.

Extreme sampling seems to be used in many studies in disguise of a case-control study, as we saw for the rs9939609 and rs17782313 studies by Frayling et al. (2007) and Loos et al. (2008) respectively. We have shown that the ratio estimated in such studies cannot be interpreted as the standard odds ratio. It would be meaningful to gain insight into the severity of assuming that one has estimated the true odds ratio. We would need a completely genotyped dataset to achieve this.

We also aim to write an article concerning extreme phenotype sampling and rare variant association studies. We have developed two new EPS methods in this thesis (CMC and β-SO) based on methods for rare variant association studies. We have seen evidence that these are both powerful. It is claimed that extreme sampling is more powerful to detect associations between causal variants and phenotypes. However, we have also seen that random sampling in some situations can be as powerful as extreme phenotype sampling. If we combine this observation with observations from the HUNT dataset we have analysed in this thesis, it might seem as if the power gain of extreme phenotype sampling is too small to justify the extreme sampling compared to the more standard statistical method of random sampling. We would like to use a real dataset of rare variants to test all the rare variant methods in both a cross-sectional and extreme phenotype sampling design. We would also like to look further into the definition of extreme samples and investigate whether larger datasets which uses perhaps only 10% of the most extreme phenotypes in each end of the spectrum perhaps gain more power from the extreme sampling.

List of references

Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Rogers, K. & Walter, P. (2010), *Essential Cell Biology*, third edn, Garland Science.

Barnett, I. J., Lee, S. & Lin, X. (2013), 'Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies', *Genetic Epidemiology* **37**.

Casella, G. & Berger, R. L. (2002), *Statistical Inference*, second edn, Brooks/Cole, Cengage Learning.

Fan, R., Wang, Y., Mills, J. L., Wilson, A. F., Bailey-Wilson, J. E. & Xiong, M. (2013), 'Functional Linear Models for Association Analysis of Quantitative Traits', *Genetic Epidemiology* **37**.

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**.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Chang, L.-Y., Huang, W., Liu, B., Shen, Y. et al. (2003), 'The International HapMap Project', *Nature* **426**.

Huang, B. & Lin, D. (2007), 'Efficient Association Mapping of Quantitative Trait Loci with Selective Genotyping', *The American Journal of Human Genetics* **80**.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R. & Herring, A. H. (2005), 'Missing Data Methods for Generalized Linear Models: A Comparative Review', *Journal of American Statistical Association* **100**.

Ibrahim, Q. C. J. G., Cheng, M.-H. & Senchaudhuri, P. (2008), 'Theory and Inference for Regression Models with Missing Responses and Covariates', *Journal of Multivariate Analysis* **99**.

Johnson, R. A. & Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis*, fifth edn, Pearson Education International.

Johnson, R. A. & Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis*, sixth edn, Pearson Prentice Hall.

Karr, A. F. (1993), *Probability*, Springer-Verlag.

Langaas, M. & Bakke, Ø. (2013), 'Robust Methods to Detect Disease-Genotype Association in Genetic Association Studies: Calculate P-values Using Exact Conditional Enumeration Instead of Asymptotic Approximations', *arXiv preprint arXiv:1307.7536* .

Langhammer, A., Krokstad, S., Romundstad, P., Heggland, J. & Holmen, J. (2012), 'The HUNT Study: Participation is Associated with Survival and Depends on Socioeconomic Status, Diseases and Symptoms', *BMC medical research methodology* **12**.

Lee, S., Wu, M. C. & Lin, X. (2012), 'Optimal Tests for Rare Variant Effects in Sequencing Association Studies', *Biostatistics* **13**.

Li, B. & Leal, S. M. (2008), 'Method for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data', *The American Journal of Human Genetics* **83**.

Li, D., Lewinger, J. P., Gauderman, W. J., Murcray, C. E. & Conti, D. (2011), 'Using Extreme Phenotype Sampling to Identify the Rare Causal Variants of Quantitative Traits in Association Studies', *Genetic Epidemiology* **35**.

Loos, R. J., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., Inouye, M., Freathy, R. M., Attwood, A. P., Beckmann, J. S. et al. (2008), 'Common Variants near MC4R are Associated with Fat Mass, Weight and Risk of Obesity', *Nature genetics* **40**.

Luo, L., Zhu, Y. & Xiong, M. (2013), 'Smoothed Functional Principle Component Analysis for Testing Association of the Entire Allelic Spectrum of Genetic Variation', *European Journal of Human Genetics* **21**.

Madsen, B. E. & Browning, S. R. (2009), 'A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic', *PLoS Genetics* **5**.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, second edn, Chapman and Hall/CRC.

Morgenthaler, S. & Thilly, W. G. (2006), 'A Strategy to Discover Genes That Carry Multi-allelic or Mono-allelic Risk for Common Diseases: A Cohort Allelic Sums Test (CAST)', *Mutation Research* **615**.

Mostad, I. L., Langaas, M. & Grill, V. (2014), 'Central Obesity is Associated with Lower Intake of Whole-grain Bread and Less Frequent Breakfast and Lunch: Results from the HUNT Study, an Adult All-population Survey', *Applied Physiology, Nutrition, and Metabolism* **39**.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006), 'Principal Components Analysis Corrects for Stratification in Genome-wide Association Studies', *Nature Genetics* .

Ramsay, J. & Silverman, B. (2005), *Functional Data Analysis*, second edn, Springer.

Robinson, L. D. & Jewell, N. P. (1991), 'Some Surprising Results about Covariate Adjustment in Logistic Regression Models', *International Statistical review* **59**.

Rothman, K. J., Greenland, S. & Lash, T. L. (2008), *Modern Epidemiology*, third edn, Lippincott Williams and Wilkins.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J. & Altshuler, D. (2005), 'Calibrating a Coalescent Simulation of Human Genome Sequence Variation', *Genome Research* **15**.

Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. (2009), 'Common vs. Rare Allele Hypotheses for Common Diseases', *Current Opinion in Genetics and Development* **19**.

Smyth, G. K. (2003), 'Pearson's Goodness of Fit Statistic as a Score Test Statistic', *Science and Statistics: A Festschrift for Terry Speed* **40**.

Tang, Y. (2010), 'Equivalence of Three Score Tests for Association Mapping of Quantitative Trait Loci Under Selective Genotyping', *Genetic Epidemiology* **34**.

The International HapMap Project (2002-2009).
URL: *http://hapmap.ncbi.nlm.nih.gov/*

Walpole, R. E., Myers, R. H., Myers, S. L. & Ye, K. (2007), *Probability and Statistics for Engineers and Scientists*, eight edn, Pearson Prentice Hall.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. (2011), 'Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test', *The American Journal of Human Genetics* **89**.

Ziegler, A. & König, I. K. (2010), *A Statistical Approach to Genetic Epidemiology*, second edn, Wiley-Blackwell.

Appendix A

Extreme phenotype sampling and case-control studies

In Chapter 4 we discussed how to use a logistic regression model in a case-control study, and how this model estimates the true odds ratio of the population. Here, we show that when the sampling of controls is made from below some threshold, leaving middle-valued phenotypes out of the study, the odds ratio from a logistic regression model does not estimate the true odds ratio of the population.

When sampling is based on extreme phenotypes, the population, denoted Ω, is split into three disjoint groups. We define the subset $D \in \Omega$ as those individuals with extreme phenotypes that can be considered "cases". We define the subset $N \in \Omega$ as the individuals with extreme phenotypes in the other end of the spectrum, often considered the healthier of the two extremes. Finally, $\tilde{N} \in \Omega$ contains all remaining individuals. We therefore have the partition $D \cap \tilde{N} \cap N = \Omega$.

The sampling probabilities for the study are given by $\tau_N = P(S|N)$, $\tau_{\tilde{N}} = P(S|\tilde{N})$ and $\tau_D = P(S|D)$, where by design $\tau_{\tilde{N}} = 0$. Further, the continuous disease measure must be dichotomized such that all $y \in D$ are are given the value 1, while all $y \in N$ are set to zero.

When fitting a logistic regression model to such data we find an estimate of $\rho_i^* = P(D|S \cap \tilde{N}^c; \mathbf{x}_i)$. By the same steps as were taken to obtain Equation (4.2), it can be shown that

$$\rho_i^* = \frac{P(S|D)P(D|\tilde{N}^c; \mathbf{z}_i)}{P(S|D)P(D|\tilde{N}^c; \mathbf{z}_i) + P(S|\tilde{N})P(\tilde{N}|\tilde{N}^c; \mathbf{z}_i) + P(S|N)P(N|\tilde{N}^c; \mathbf{z}_i)},$$

which, because $P(S|\tilde{N}) = 0$, simplifies to

$$\rho_i^* = \frac{P(S|D)P(D|\tilde{N}^c; \mathbf{z}_i)}{P(S|D)P(D|\tilde{N}^c; \mathbf{z}_i) + P(S|N)P(N|\tilde{N}^c; \mathbf{z}_i)}.$$

As $P(D|\tilde{N}^c; \mathbf{z}_i) = P(D \cap \tilde{N}^c \cap \mathbf{z}_i)P(\tilde{N}^c; \mathbf{z}_i)$, and $P(D \cap \tilde{N}^c; \mathbf{z}_i) = P(D; \mathbf{z}_i)$ we find that this expression further simplifies into

$$\rho_i^* = \frac{\tau_D P(D; \mathbf{z}_i)}{\tau_D P(D; \mathbf{z}_i) + \tau_N P(N; \mathbf{z}_i)} = \frac{\tau_D \pi_i}{\tau_D \pi_i + \tau_N P(N; \mathbf{z}_i)}.$$

This expression is quite similar to what we found for the case-control study in Equation (4.2), but the difference in this case is that $P(N; \mathbf{z}_i) \neq 1 - \pi_i$. However, when using the logistic regression model it is assumed that there are only two possible outcomes, and that the sum of their probabilities is one. The interpretation of the odds ratio

$$\frac{\rho_i^*/(1-\rho_i^*)}{\rho_j^*/(1-\rho_j^*)},$$

is therefore the odds of being diseased under covariate levels $\mathbf{z}_i$ relative to $\mathbf{z}_j$, assuming that there is *no probability, with these covariate levels, that the corresponding phenotype belongs in $\tilde{N}$*. The interpretation of such an odds ratio is unnatural and not easily understood.

Appendix B

Score test calculations

B.1 Derivatives of log likelihood functions

For simplicity of notation, define the function

$$f_i = Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{G}_i.$$

B.1.1 The cross-sectional design

In a cross-sectional study with normally distributed phenotypes, the log likelihood function is given by

$$l(\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma; \mathbf{Y}) = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{G}_i)^2.$$

The first derivatives of the log likelihood are

$$\frac{\partial l}{\partial \alpha_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} f_i, \qquad \frac{\partial l}{\partial \boldsymbol{\alpha}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} f_i \mathbf{X}_i,$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} f_i \mathbf{G}_i, \qquad \frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} f_i^2.$$

The second derivatives are

$$\frac{\partial^2 l}{\partial \alpha_0^2} = -\frac{n}{\sigma^2}, \qquad \frac{\partial^2 l}{\partial \alpha_0 \partial \boldsymbol{\alpha}} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{X}_i,$$

$$\frac{\partial^2 l}{\partial \alpha_0 \partial \boldsymbol{\beta}} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{G}_i, \qquad \frac{\partial^2 l}{\partial \alpha_0 \partial \sigma} = -\frac{2}{\sigma^3} \sum_{i=1}^{n} f_i,$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha}^2} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T, \qquad \frac{\partial^2 l}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{G}_i^T,$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha} \partial \sigma} = -\frac{2}{\sigma^3} \sum_{i=1}^{n} f_i \mathbf{X}_i, \qquad \frac{\partial^2 l}{\partial \boldsymbol{\beta}^2} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{G}_i \mathbf{G}_i^T,$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \sigma} = -\frac{2}{\sigma^3} \sum_{i=1}^{n} f_i \mathbf{G}_i, \qquad \frac{\partial^2 l}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{n} f_i^2.$$

B.1.2 The EPS design and the conditional model

For the conditional model applied to EPS data, the conditional log likelihood function is given by

$$l_c = -n\log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n}\frac{1}{2\sigma^2}f_i^2 - \sum_{i=1}^{n}\log(1-\Phi_{u,i}+\Phi_{l,i}),$$

where

$$\Phi_{l,i} = \Phi\left(\frac{c_l-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}_i-\boldsymbol{\beta}^T\mathbf{G}_i}{\sigma}\right), \qquad \Phi_{u,i} = \Phi\left(\frac{c_u-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}_i-\boldsymbol{\beta}^T\mathbf{G}_i}{\sigma}\right).$$

We use the following notation;

$$h_{ji} = \frac{-\phi_{u,i}\cdot\left(\frac{c_u-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}_i-\boldsymbol{\beta}^T\mathbf{G}_i}{\sigma}\right)^j + \phi_{l,i}\cdot\left(\frac{c_l-\alpha_0-\boldsymbol{\alpha}^T\mathbf{X}_i-\boldsymbol{\beta}^T\mathbf{G}_i}{\sigma}\right)^j}{1-\Phi_{u,i}+\Phi_{l,i}}, j=0,1,2,3.$$

The first derivatives of the log likelihood are

$$\frac{\partial l_c}{\partial\alpha_0} = \frac{1}{\sigma^2}\sum_{i=1}^{n}f_i + \frac{1}{\sigma}\sum_{i=1}^{n}h_{0i}, \qquad \frac{\partial l_c}{\partial\boldsymbol{\alpha}} = \frac{1}{\sigma^2}\sum_{i=1}^{n}f_i\mathbf{X}_i + \frac{1}{\sigma}\sum_{i=1}^{n}h_{0i}\mathbf{X}_i,$$

$$\frac{\partial l_c}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}\sum_{i=1}^{n}f_i\mathbf{G}_i + \frac{1}{\sigma}\sum_{i=1}^{n}h_{0i}\mathbf{G}_i, \qquad \frac{\partial l_c}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}f_i^2 + \frac{1}{\sigma}\sum_{i=1}^{n}h_{1i}.$$

The second derivatives are

$$\frac{\partial^2 l_c}{\partial\alpha_0^2} = -\frac{n}{\sigma^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n}(h_{1i}-h_{0i}^2), \qquad \frac{\partial^2 l_c}{\partial\alpha_0\partial\boldsymbol{\alpha}} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(-1+h_{1i}-h_{0i}^2)\mathbf{X}_i,$$

$$\frac{\partial^2 l_c}{\partial\alpha_0\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(-1+h_{1i}-h_{0i}^2)\mathbf{G}_i, \qquad \frac{\partial^2 l_c}{\partial\boldsymbol{\alpha}^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(-1+h_{1i}-h_{0i}^2)\mathbf{X}_i\mathbf{X}_i^T,$$

$$\frac{\partial^2 l_c}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(-1+h_{1i}-h_{0i}^2)\mathbf{X}_i\mathbf{G}_i^T, \qquad \frac{\partial^2 l_c}{\partial\boldsymbol{\beta}^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(-1+h_{1i}-h_{0i}^2)\mathbf{G}_i\mathbf{G}_i^T,$$

$$\begin{aligned}
\frac{\partial^2 l_c}{\partial\alpha_0\partial\sigma} &= -\frac{2}{\sigma^3}\sum_{i=1}^{n}f_i + \frac{1}{\sigma^2}\sum_{i=1}^{n}(-2h_{0i}+h_{2i}+h_{0i}h_{1i}),\\
\frac{\partial^2 l_c}{\partial\boldsymbol{\alpha}\partial\sigma} &= -\frac{2}{\sigma^3}\sum_{i=1}^{n}f_i\mathbf{X}_i + \frac{1}{\sigma^2}\sum_{i=1}^{n}(-2h_{0i}+h_{2i}+h_{0i}h_{1i})\mathbf{X}_i,\\
\frac{\partial^2 l_c}{\partial\boldsymbol{\beta}\partial\sigma} &= -\frac{2}{\sigma^3}\sum_{i=1}^{n}f_i\mathbf{G}_i + \frac{1}{\sigma^2}\sum_{i=1}^{n}(-2h_{0i}+h_{2i}+h_{0i}h_{1i})\mathbf{G}_i,\\
\frac{\partial^2 l_c}{\partial\sigma^2} &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}f_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(-3h_{1i}+h_{3i}+h_{1i}^2).
\end{aligned}$$

B.1.3 The EPS design and the missing genotype model

For simplicity of notation, we define

$$f_i = Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{G}_i,$$

as well as

$$f_i(\mathbf{g}) = Y_i - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{X}_i - \boldsymbol{\beta}^T \mathbf{g}.$$

The log likelihood function for the missing genotype model is given by

$$l_m = -n \log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{1}{2\sigma^2} f_i^2 - \sum_{i=n+1}^{N} \log \left(\sum_{\mathbf{g}} \phi_i(\mathbf{g}) P(\mathbf{G} = \mathbf{g}) \right),$$

where

$$\phi_i(\mathbf{g}) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left(\frac{1}{2\sigma^2} f_i(\mathbf{g})^2 \right).$$

Define the functions

$$k_{ji} = \frac{\sum_{\mathbf{g}} f_i(\mathbf{g})^j \phi_i(\mathbf{g}) P(\mathbf{G} = \mathbf{g})}{\sum_{\mathbf{g}} \phi_i(\mathbf{g}) P(\mathbf{G} = \mathbf{g})}, \qquad k_{ji}(\mathbf{g}) = \frac{\sum_{\mathbf{g}} f_i(\mathbf{g})^j \phi_i(\mathbf{g}) P(\mathbf{G} = \mathbf{g}) \mathbf{g}}{\sum_{\mathbf{g}} \phi_i(\mathbf{g}) P(\mathbf{G} = \mathbf{g})},$$

for $j = 0, 1, 2, 3, 4$. Additionally, define

$$k_{2i}(\mathbf{g}\mathbf{g}^T) = \frac{\sum_{\mathbf{g}} f_i(\mathbf{g})^2 \phi_i(\mathbf{g}) P(\mathbf{G} = \mathbf{g}) \mathbf{g}\mathbf{g}^T}{\sum_{\mathbf{g}} \phi_i(\mathbf{g}) P(\mathbf{G} = \mathbf{g})}.$$

The first derivatives of the missing genotype log likelihood are

$$\frac{\partial l_m}{\partial \alpha_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} f_i + \frac{1}{\sigma^2} \sum_{i=n+1}^{N} k_{1i}, \qquad \frac{\partial l_m}{\partial \boldsymbol{\alpha}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} f_i \mathbf{X}_i + \frac{1}{\sigma^2} \sum_{i=n+1}^{N} k_{1i} \mathbf{X}_i,$$

$$\frac{\partial l_m}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} f_i \mathbf{G}_i + \frac{1}{\sigma^2} \sum_{i=n+1}^{N} k_{1i}(\mathbf{g}), \qquad \frac{\partial l_m}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} f_i^2 + \frac{1}{\sigma^3} \sum_{i=n+1}^{N} k_{2i}.$$

The second derivatives are

$$
\begin{aligned}
\frac{\partial^2 l_m}{\partial \alpha_0^2} &= -\frac{N}{\sigma^2} + \frac{1}{\sigma^4} \sum_{i=n+1}^{N} k_{2i} - k_{1i}^2, \\
\frac{\partial^2 l_m}{\partial \alpha_0 \partial \boldsymbol{\alpha}} &= -\frac{1}{\sigma^2} \sum_{i=1}^{N} \mathbf{X}_i + \frac{1}{\sigma^4} \sum_{i=n+1}^{N} (k_{2i} - k_{1i}^2)\mathbf{X}_i, \\
\frac{\partial^2 l_m}{\partial \alpha_0 \partial \boldsymbol{\beta}} &= -\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{G}_i - \frac{1}{\sigma^2} \sum_{i=n+1}^{N} k_{0i}(\mathbf{g}) + \frac{1}{\sigma^4} \sum_{i=n+1}^{N} k_{2i}(\mathbf{g}) - k_{1i}k_{1i}(\mathbf{g}), \\
\frac{\partial^2 l_m}{\partial \alpha_0 \partial \sigma} &= -\frac{2}{\sigma^3} \sum_{i=1}^{n} f_i - \frac{2}{\sigma^3} \sum_{i=n+1}^{N} k_{1i} + \frac{1}{\sigma^5} \sum_{i=n+1}^{N} k_{3i} - k_{2i}k_{1i}, \\
\frac{\partial^2 l_m}{\partial \boldsymbol{\alpha}^2} &= -\frac{1}{\sigma^2} \sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i^T + \frac{1}{\sigma^4} \sum_{i=n+1}^{N} (k_{2i} - k_{1i}^2)\mathbf{X}_i \mathbf{X}_i^T, \\
\frac{\partial^2 l_m}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}} &= -\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{G}_i^T - \frac{1}{\sigma^2} \sum_{i=n+1}^{N} \mathbf{X}_i k_{0i}(\mathbf{g})^T + \frac{1}{\sigma^4} \sum_{i=n+1}^{N} \mathbf{X}_i (k_{2i}(\mathbf{g}) - k_{1i}k_{1i}(\mathbf{g}))^T, \\
\frac{\partial^2 l_m}{\partial \boldsymbol{\alpha} \partial \sigma} &= -\frac{2}{\sigma^3} \sum_{i=1}^{n} f_i \mathbf{X}_i - \frac{2}{\sigma^3} \sum_{i=n+1}^{N} k_{1i} \mathbf{X}_i + \frac{1}{\sigma^5} \sum_{i=n+1}^{N} (k_{3i} - k_{2i}k_{1i})\mathbf{X}_i, \\
\frac{\partial^2 l_m}{\partial \boldsymbol{\beta}^2} &= -\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{G}_i \mathbf{G}_i^T + \frac{1}{\sigma^4} \sum_{i=n+1}^{N} k_{2i}(\mathbf{g}\mathbf{g}^T) - k_{1i}(\mathbf{g}) k_{1i}(\mathbf{g})^T, \\
\frac{\partial^2 l_m}{\partial \boldsymbol{\beta} \partial \sigma} &= -\frac{2}{\sigma^3} \sum_{i=1}^{n} f_i \mathbf{G}_i - \frac{2}{\sigma^3} \sum_{i=n+1}^{N} k_{1i}(\mathbf{g}) + \frac{1}{\sigma^5} \sum_{i=n+1}^{N} k_{3i}(\mathbf{g}) - k_{2i}k_{1i}(\mathbf{g}), \\
\frac{\partial^2 l_m}{\partial \sigma^2} &= \frac{N}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{n} f_i^2 - \frac{3}{\sigma^4} \sum_{i=n+1}^{N} k_{2i} + \frac{1}{\sigma^6} \sum_{i=1+n}^{N} k_{4i} - k_{2i}^2.
\end{aligned}
$$

B.2 Proof

Tang (2010) proved that for the simple model $y = \alpha_0 + \beta G + \epsilon$, the score test for testing $H_0 : \beta = 0$ is equivalent for a conditional distribution of y (as under the EPS design) and a normal distribution of y. In the following we present the proof that this result also holds for the case where $y = \alpha_0 + \beta_1 G_1 + \cdots + \beta_p G_p + \epsilon$ under the null hypothesis $H_0 : \beta_1 = \cdots = \beta_p = 0$. Note that if the null hypothesis does not include all covariate coefficients, then the result is no longer valid. In this setting we define

$$f_i = Y_i - \alpha - \boldsymbol{\beta}^T \mathbf{G}_i.$$

The score test statistic for the normal distributed sample

We will now find the expression for the score test statistic T^* (defined in Equation (3.19)), for the regression model that assumes random sampling from a normal distribution. The log likelihood function is

$$l(\alpha, \sigma, \boldsymbol{\beta}; \mathbf{Y}) = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} f_i^2. \tag{B.1}$$

The test is performed under the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$, with α and σ as nuisance parameters. We thus have $\boldsymbol{\theta}_1 = (\alpha, \sigma)$ and $\boldsymbol{\theta}_2 = \boldsymbol{\beta}$ and $H_0 : \boldsymbol{\theta} = (\hat{\boldsymbol{\theta}}_1, \mathbf{0})$.

Under H_0, the maximum likelihood estimates for α and σ will be the same for any dimension of $\boldsymbol{\beta}$, and given by

$$\hat{\alpha} = \bar{Y}, \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = S_{YY}.$$

The scores $\mathbf{s}_1, \ldots, \mathbf{s}_n$, as presented in Equation (3.15), are under H_0 given by

$$\mathbf{s}_i = \begin{bmatrix} \frac{\partial}{\partial \alpha} \log f_{Y_i}|_{(\hat{\alpha}, \hat{\sigma}, \mathbf{0})} \\ \frac{\partial}{\partial \sigma} \log f_{Y_i}|_{(\hat{\alpha}, \hat{\sigma}, \mathbf{0})} \\ \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{Y_i}|_{(\hat{\alpha}, \hat{\sigma}, \mathbf{0})} \end{bmatrix} = \begin{bmatrix} \frac{1}{S_{YY}}(Y_i - \bar{Y}) \\ -\frac{1}{\sqrt{S_{YY}}} + \frac{1}{(S_{YY})^{3/2}}(Y_i - \bar{Y})^2 \\ \frac{1}{S_{YY}}(Y_i - \bar{Y})\mathbf{G}_i \end{bmatrix},$$

for all $i \in \{1, \ldots, n\}$. It is reasonable to assume that the scores constitute a bounded sequence when Y_i represents a ratio such as waist-hip ratio, and $\mathbf{G}_i$ are physical measures or genotypes coded as 0, 1 and 2. By this assumption, the Lindeberg condition is satisfied, and the central limit theorem can be applied.

We will need the Fisher information matrix, I. This is found by taking the negative expectation of the second derivatives of the log likelihood function in (B.1). The second derivatives are

$$\begin{aligned} \frac{\partial^2 l}{\partial \alpha^2} &= -\frac{n}{\sigma^2} & \frac{\partial^2 l}{\partial \sigma^2} &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{n} f_i^2, \\ \frac{\partial^2 l}{\partial^2 \boldsymbol{\beta}} &= -\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{G}_i \mathbf{G}_i^T & \frac{\partial^2 l}{\partial \alpha \partial \sigma} &= -\frac{2}{\sigma^3} \sum_{i=1}^{n} f_i, \\ \frac{\partial^2 l}{\partial \alpha \partial \boldsymbol{\beta}} &= -\frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{G}_i = -\frac{n}{\sigma^2} \bar{\mathbf{G}}, & \frac{\partial^2 l}{\partial \sigma \partial \boldsymbol{\beta}} &= -\frac{2}{\sigma^3} \sum_{i=1}^{n} f_i \mathbf{G}_i, \end{aligned}$$

where $\bar{\mathbf{G}}$ is a vector of the mean values $\bar{G}_1, \ldots, \bar{G}$. When taking the expected value of the second derivatives, we use the fact that $E(f_i) = 0$, while $E(f_i^2) = \sigma^2$. Consequently, the Fisher information matrix is given by

$$I = \frac{1}{\sigma^2} \begin{bmatrix} n & 0 & n\bar{\mathbf{G}}^T \\ 0 & 2n & \mathbf{0}_{1 \times p_g} \\ n\bar{\mathbf{G}} & \mathbf{0}_{p_g \times 1} & \sum_{i=1}^{n} \mathbf{G}_i \mathbf{G}_i^T \end{bmatrix}, \tag{B.2}$$

where $\mathbf{0}_{N \times M}$ is a $N \times M$ matrix with all entries equal to zero. $\mathbf{G}_i$ is a p_g-dimensional vector. The information matrix is estimated by inserting $\hat{\sigma}^2 = S_{YY}$. The estimated matrix is denoted $\hat{I}$. Note that

$$(\hat{I})_{11} = \frac{n}{S_{YY}} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix},$$

is a 2×2 matrix for all p_g as long as there are two nuisance parameters, while $(\hat{I})_{12}$ is a $2 \times p_g$ matrix given by

$$(\hat{I})_{12} = \frac{n}{S_{YY}} \begin{bmatrix} \bar{\mathbf{G}}^T \\ \mathbf{0}_{1 \times p_g} \end{bmatrix},$$

and $(\hat{I})_{21}$ is the transpose of this. Finally, $(\hat{I})_{22}$ is a $p_g \times p_g$ matrix given by $\frac{1}{S_{YY}} \sum_{i=1}^n \mathbf{G}_i \mathbf{G}_i^T$.

The score statistic conditioned on $\partial l / \partial \alpha = 0$ and $\partial l / \partial \sigma = 0$, and evaluated in $(\hat{\alpha}, \hat{\sigma}, \mathbf{0})$, is given by

$$\mathbf{S}^* = \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{(\hat{\alpha}, \hat{\sigma}, \mathbf{0})} = \frac{1}{S_{YY}} \sum_{i=1}^n (Y_i - \bar{Y})(\mathbf{G}_i - \bar{\mathbf{G}}),$$

which follows from

$$\begin{aligned} \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{(\hat{\alpha}, \hat{\sigma}, \mathbf{0})} &= \frac{1}{S_{YY}} \sum_{i=1}^n (Y_i - \bar{Y}) \mathbf{G}_i \\ &= \frac{1}{S_{YY}} \left(\sum_{i=1}^n Y_i \mathbf{G}_i - n \bar{Y} \bar{\mathbf{G}} \right) \\ &= \frac{1}{S_{YY}} \sum_{i=1}^n (Y_i - \bar{Y})(\mathbf{G}_i - \bar{\mathbf{G}}). \end{aligned}$$

The expected value of $\mathbf{S}^*$ is zero, as shown in Equation (3.16), and the variance Σ^* is easily determined by Equation (3.18). Using the information matrix in Equation (B.2), Σ^* is found to be

$$\begin{aligned} \Sigma^* &= \frac{1}{S_{YY}} \sum_{i=1}^n \mathbf{G}_i \mathbf{G}_i^T - \frac{n}{S_{YY}} \begin{bmatrix} \bar{\mathbf{G}} & \mathbf{0}_{p_g \times 1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} \bar{\mathbf{G}}^T \\ \mathbf{0}_{1 \times p_g} \end{bmatrix} \\ &= \frac{1}{S_{YY}} \sum_{i=1}^n \mathbf{G}_i \mathbf{G}_i^T - \frac{n}{S_{YY}} \begin{bmatrix} \bar{\mathbf{G}} & \mathbf{0}_{p_g \times 1} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{G}}^T \\ \mathbf{0}_{1 \times p_g} \end{bmatrix} \\ &= \frac{1}{S_{YY}} \sum_{i=1}^n \mathbf{G}_i \mathbf{G}_i^T - \frac{n}{S_{YY}} \bar{\mathbf{G}} \bar{\mathbf{G}}^T \\ &= \frac{1}{S_{YY}} \sum_{i=1}^n (\mathbf{G}_i - \bar{\mathbf{G}})(\mathbf{G}_i - \bar{\mathbf{G}})^T. \end{aligned}$$

We thus have that the score statistic is given by

$$\mathbf{S}^* = \frac{n}{S_{YY}} S_{Y\mathbf{G}},$$

while the covariance matrix is given by

$$\Sigma^* = \frac{n}{S_{YY}} S_{\mathbf{G}\mathbf{G}}.$$

By the definition of T^* in Equation (3.19), we obtain the score test statistic

$$T^* = \frac{n}{S_{YY}} S_{Y\mathbf{G}}^T S_{\mathbf{G}\mathbf{G}}^{-1} S_{Y\mathbf{G}}.$$

This score test statistic is approximately χ^2-distributed with p_g degrees of freedom.

The score test statistic for the conditional model

As described, the log likelihood function for the conditional model is given by

$$\begin{aligned} l_c &= \sum_{i=1}^{n} \log(\frac{1}{\sigma}\phi_i) - \sum_{i=1}^{n} \log(1 - \Phi_{u,i} + \Phi_{l,i}) \\ &= -n\log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{1}{2\sigma^2} f_i^2 - \sum_{i=1}^{n} \log(1 - \Phi_{u,i} + \Phi_{l,i}). \end{aligned}$$

Following the same procedure as in the previous section, we find the score test statistic by finding the score statistic and its variance conditioned on the MLEs of the nuisance parameters and $\boldsymbol{\beta} = 0$. The derivatives of the conditional log-likelihood are not so easily written out as for the prospective model. The following functions, inspired by Tang (2010), are therefore introduced for ease of notation;

$$h_{ji} = \frac{-\phi_{u,i} \cdot \left(\frac{c_u - \alpha - \boldsymbol{\beta}^T \mathbf{G}_i}{\sigma}\right)^j + \phi_{l,i} \cdot \left(\frac{c_l - \alpha - \boldsymbol{\beta}^T \mathbf{G}_i}{\sigma}\right)^j}{1 - \Phi_{u,i} + \Phi_{l,i}}, j = 0, 1, 2, 3.$$

We note that under $H_0 : \boldsymbol{\beta} = \mathbf{0}$, the vectors $\mathbf{G}_i$ do not appear in the above expressions, and the functions h_{ji} are equal for all $i \in \{1, \dots, n\}$. We can therefore denote them by h_0, h_1, h_2 and h_3 under the null hypothesis.

Let $\hat{\alpha}_c$ and $\hat{\sigma}_c$ be the maximum likelihood estimators under H_0 in the conditional model. By the maximum likelihood procedure, and by inserting $\boldsymbol{\beta} = \mathbf{0}$, we have that

$$\hat{\alpha}_c = \bar{Y} + \hat{\sigma}_c h_0.$$

For the variance estimate we see that

$$\hat{\sigma}_c^2(1 - h_1) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha}_c)^2,$$

where $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha}_c)^2$ can be rewritten as $S_{YY} + \hat{\sigma}_c^2 h_0^2$ by the definition of S_{YY}. Thus

$$\hat{\sigma}_c^2 = \frac{S_{YY}}{1 - h_1 - h_0^2}.$$

The score statistic, conditioned on $\partial l_c/\partial\alpha = 0$ and $\partial l_c/\partial\sigma = 0$, and evaluated in $(\hat{\alpha}_c, \hat{\sigma}_c, \mathbf{0})$, is given by

$$\begin{aligned} \mathbf{S}_c^* = \frac{\partial l_c}{\partial \boldsymbol{\beta}}|_{(\hat{\alpha}_c, \hat{\sigma}_c, \mathbf{0})} &= \frac{a}{S_{YY}}\left(\sum_{i=1}^{n}(Y_i - \hat{\alpha}_c)\mathbf{G}_i + n\hat{\sigma}_c h_0 \bar{\mathbf{G}}\right) \\ &= \frac{a}{S_{YY}}\left(\sum_{i=1}^{n}(Y_i - \bar{Y} - \hat{\sigma}_c h_0)\mathbf{G}_i + \hat{\sigma}_c h_0 \mathbf{G}_i\right) \\ &= \frac{a}{S_{YY}}\sum_{i=1}^{n}(Y_i - \bar{Y})\mathbf{G}_i \\ &= \frac{an}{S_{YY}} S_{Y\mathbf{G}}. \end{aligned}$$

The second derivatives of the conditional log likelihood function are

$$
\begin{aligned}
\frac{\partial^2 l_c}{\partial\alpha^2} &= -\frac{n}{\sigma^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{0i}^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{1i},\\
\frac{\partial^2 l_c}{\partial\sigma^2} &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n} f_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{0i}^2 - \frac{2}{\sigma^2}\sum_{i=1}^{n} h_{1i} + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{3i},\\
\frac{\partial^2 l_c}{\partial\boldsymbol{\beta}^2} &= -\frac{1}{\sigma^2}\sum_{i=1}^{n} \mathbf{G}_i\mathbf{G}_i^T + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{0i}^2\mathbf{G}_i\mathbf{G}_i^T + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{1i}\mathbf{G}_i\mathbf{G}_i^T,\\
\frac{\partial^2 l_c}{\partial\alpha\partial\sigma} &= -\frac{2}{\sigma^3}\sum_{i=1}^{n} f_i - \frac{2}{\sigma^2}\sum_{i=1}^{n} h_{0i} + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{0i}h_{1i} + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{2i},\\
\frac{\partial^2 l_c}{\partial\alpha\partial\boldsymbol{\beta}} &= -\frac{1}{\sigma^2}\sum_{i=1}^{n} \mathbf{G}_i + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{0i}^2\mathbf{G}_i + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{1i}\mathbf{G}_i,\\
\frac{\partial^2 l_c}{\partial\sigma\partial\boldsymbol{\beta}} &= -\frac{2}{\sigma^3}\sum_{i=1}^{n} f_i\mathbf{G}_i - \frac{2}{\sigma^2}\sum_{i=1}^{n} h_{0i}\mathbf{G}_i + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{0i}h_{1i}\mathbf{G}_i + \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{2i}\mathbf{G}_i.
\end{aligned}
$$

By taking the negative expectation of the second derivatives, and inserting the null hypothesis and the corresponding maximum likelihood estimators, we find the Fisher information matrix. For further ease of notation, again inspired by Tang (2010), we introduce the functions

$$
\begin{aligned}
a &= 1 - h_1 - h_0^2,\\
b &= 2h_0 - h_2 - h_0h_1,\\
c &= 2 + 3h_1 - h_3 - h_0^2.
\end{aligned}
$$

Note that we can now write $\hat{\sigma}_c^2 = S_{YY}/a$. The estimated Fisher information matrix is given by

$$
\hat{I}_c = \frac{a}{S_{YY}}\begin{bmatrix} an & bn & an\bar{\mathbf{G}}^T \\ bn & cn & bn\bar{\mathbf{G}}^T \\ an\bar{\mathbf{G}} & bn\bar{\mathbf{G}} & a\sum_{i=1}^{n}\mathbf{G}_i\mathbf{G}_i^T \end{bmatrix}.
$$

Note that

$$
(\hat{I}_c)_{11} = \frac{an}{S_{YY}}\begin{bmatrix} a & b \\ b & c \end{bmatrix},
$$

while $(\hat{I}_c)_{12}$ is a $2 \times p$ matrix given by

$$
(\hat{I}_c)_{12} = \frac{an}{S_{YY}}\begin{bmatrix} a\bar{\mathbf{G}}^T \\ b\bar{\mathbf{G}}^T \end{bmatrix},
$$

and $(\hat{I}_c)_{21}$ is the transpose of this. $(\hat{I}_c)_{22}$ is a $p_g \times p_g$ matrix given by $\frac{a^2}{S_{YY}}\sum_{i=1}^{n}\mathbf{G}_i\mathbf{G}_i^T$.

The variance is found using Equation (3.18), and is given by

$$\begin{aligned}
\Sigma_c^* &= \frac{a^2}{S_{YY}} \sum_{i=1}^{n} \mathbf{G}_i \mathbf{G}_i^T - \frac{an}{S_{YY}} \begin{bmatrix} a\bar{\mathbf{G}} & b\bar{\mathbf{G}} \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} \begin{bmatrix} a\bar{\mathbf{G}}^T \\ b\bar{\mathbf{G}}^T \end{bmatrix} \\
&= \frac{a^2}{S_{YY}} \sum_{i=1}^{n} \mathbf{G}_i \mathbf{G}_i^T - \frac{an}{S_{YY}} \begin{bmatrix} a\bar{\mathbf{G}} & b\bar{\mathbf{G}} \end{bmatrix} \frac{1}{ac - b^2} \begin{bmatrix} \bar{\mathbf{G}}^T(ac - b^2) \\ \mathbf{0}_{1\times p_g} \end{bmatrix} \\
&= \frac{a^2}{S_{YY}} \sum_{i=1}^{n} \mathbf{G}_i \mathbf{G}_i^T - \frac{a^2 n}{S_{YY}} \bar{\mathbf{G}} \bar{\mathbf{G}}^T \\
&= \frac{a^2}{S_{YY}} \sum_{i=1}^{n} (\mathbf{G}_i - \bar{\mathbf{G}})(\mathbf{G}_i - \bar{\mathbf{G}})^T \\
&= \frac{a^2 n}{S_{YY}} S_{\mathbf{GG}}.
\end{aligned}$$

We see that $\mathbf{S}_c^* = a\mathbf{S}^*$, and $\Sigma_c^* = a^2\Sigma^*$. Consequently, the score test statistic T_c^* must be equivalent to T^* for the normally distributed cross-sectional sample. We have proved that the score tests for the cross-sectional and conditional models are equivalent when all covariate coefficients are zero under the null hypothesis.

Appendix C

Details of the HUNT study

C.1 Definition of covariates

Smoke level

The smoke variable is summarized based on questions 20 and 21 in questionnaire number 1, of the HUNT3 study;

- Q20: Do you smoke?
- Q21: How many cigarettes do/did you smoke daily?

The variable is defined by the following answers to Q20 and Q21;

- 0: Q20: No, I have never smoked
- 1: Q20: Yes, cigarettes once in a while, or Yes, cigars/cigarillos/pipe once in a while
- 2: Q20: No, I have stopped smoking, and Q2:1 More than 10 each day
- 3: Q20: Yes, cigarettes daily, or Yes, cigars/cigarillos/pipe daily
- 4: Q20: Yes, cigarettes daily, and Q21: More than 10 each day
- 5: Q20: Yes, cigars/cigarillos/pipe daily and Q21: More than 10 each day

Exercise frequency

The exercise frequency was categorized based question 32 in questionnaire number 1, of the HUNT3 study;

- 0: Never
- 1: Less than one time a week
- 2: One time a week
- 3: 2-3 times a week
- 4: Approximately every day

C.2 Analysis of waist-hip ratio

Waist-hip ratio is a of measure central obesity that is not as widely used as BMI. Statistical transformations of WHR are therefore not agreed upon in the scientific community as opposed to BMI for which it is widely acknowledge that a log transform of BMI is approximately normally distributed in a population. We will use our male and female datasets to assess appropriate transformations of WHR in order to create a normally distributed response variable.

We use the function `boxcox` in R. This function returns a plot of a parameter λ which is 1 if the data are already normally distributed, and 0 if a log transform is appropriate. If $0 < \lambda < 1$, the appropriate transformation is found by

$$y = \frac{\mathrm{WHR}^{\lambda} - 1}{\lambda}.$$

The input to the function is a linear model, for which we use

$$\mathrm{WHR} = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X},$$

where $\mathbf{X}$ contains age, smoke and exercise covariates.

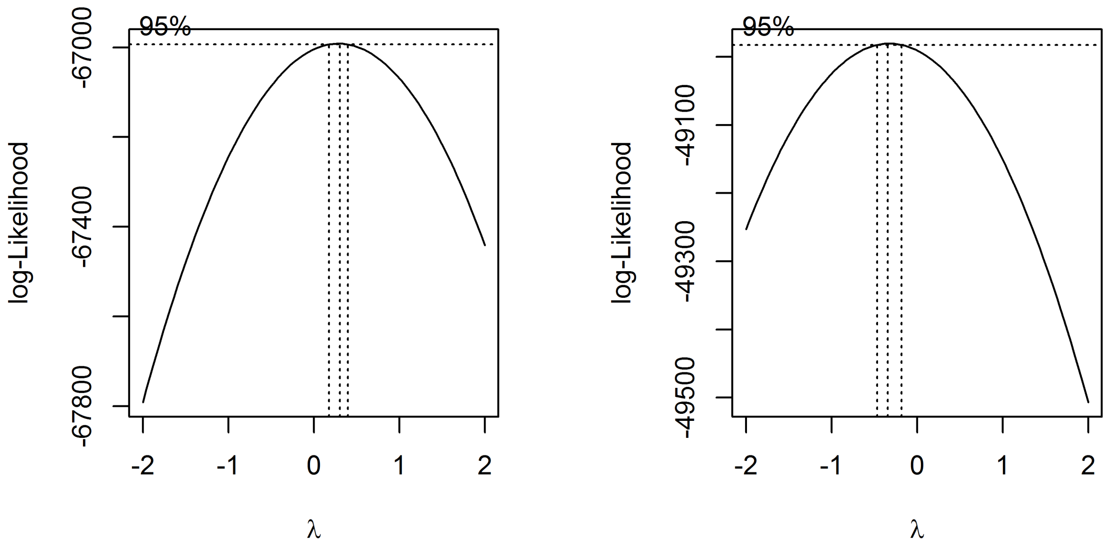


Figure C.1: Boxcox results, women and men

The resulting plots for women and men are shown in Figure C.1. We see that neither of the λs, nor their respective confidence intervals, are zero. For women, the maximal value of λ is 0.3, while for men, the maximal value is -0.3. For medicinal and biological research we deem it inappropriate to propose different transformations for men and women. We use an Anderson-Darling test (function `ad.test()`in R) to test whether the residuals of the transformed linear model, divided by the estimated standard deviations, are $N(0, 1)$ distributed. We perform the same test for log transformed WHR, as well as for a model where no transformation of WHR is applied. The p-values are presented in Table C.1. For women, the boxcox transformed WHR values yield the highest p-value, while for men, the boxcox and log transformations yield similar p-values. Because the log transform is accepted as an appropriate transform for other ratios, and as it seems an acceptable transformation for our data, we decide to work with the log transformed WHR values. For reference, we performed similar analysis to BMI values and did not find stronger evidence for a log transform than with WHR values.

Gender	Transformation	p-value
Women	Boxcox	$3.4030e-2$
	Log	$6.9770e-4$
	None	$7.1590e-4$
Men	Boxcox	$2.3300e-3$
	Log	$3.2250e-3$
	None	$2.7330e-8$

Table C.1: Anderson-Darling p-values

Variance analysis

We can test whether variances are homogeneous across subgroups of the population using the Bartlett and Levene tests. Both of these tests are easily applied in R. For example, if we wish to test age subgroups we can write

```
bartlett.test(log(WHR) ~ as.factor(agegroup))
leveneTest(log(WHR) ~ as.factor(agegroup))
```

These tests return p-values corresponding to the null hypothesis that the data are homogeneous. The Bartlett test is sensitive to data that are not normally distributed while the Levelene test is more robust in this case. We therefore perform both tests. As it is widely agreed upon that log(BMI) is homogeneous with respect to variance of subgroups in adult populations, we perform the same tests for log(BMI) in order to compare results for log(WHR). Resulting p-values are presented in Table C.2. We see that the null hypothesis should be rejected for all tests, both for log(WHR) and log(BMI). This implies that variance is not constant in the dataset. In our separate work where aim to estimate genetic effects as realistically as possible, we must be cautious of this fact.

Gender	Response variable	Bartlett p-value	Levene p-value
Women	log(WHR)	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$
	log(BMI)	$< 2.2 \cdot 10^{-16}$	$3.154 \cdot 10^{-11}$
Men	log(WHR)	$3.411 \cdot 10^{-5}$	$1.026 \cdot 10^{-5}$
	log(BMI)	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$

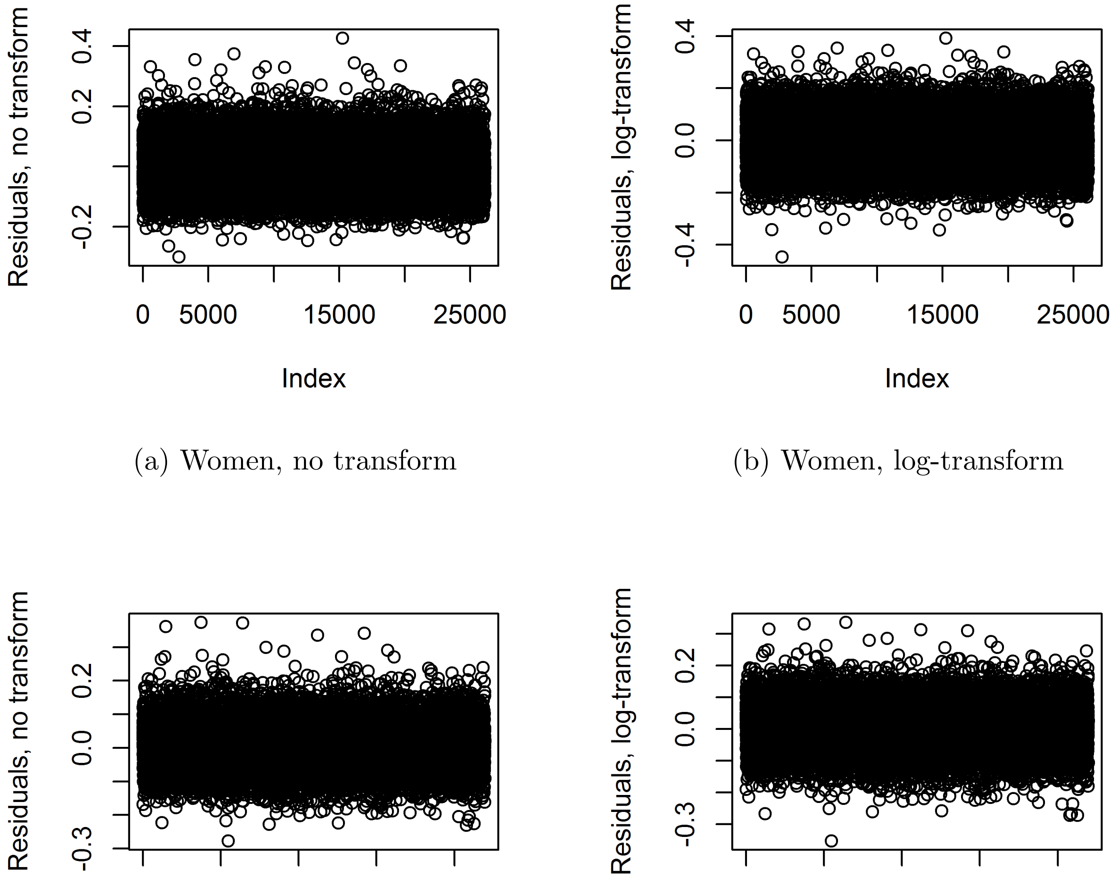
Table C.2: Tests for homogeneity of variance of age subgroups

C.3 Residuals

The plotted residuals for the linear models

$$\text{WHR} = \alpha_0 + \boldsymbol{\alpha}_1^T \mathbf{X}_{age} + \boldsymbol{\alpha}_2^T \mathbf{X}_{smoke} + \boldsymbol{\alpha}_3^T \mathbf{X}_{exercise} + \epsilon,$$
$$\log(\text{WHR}) = \alpha_0 + \boldsymbol{\alpha}_1^T \mathbf{X}_{age} + \boldsymbol{\alpha}_2^T \mathbf{X}_{smoke} + \boldsymbol{\alpha}_3^T \mathbf{X}_{exercise} + \epsilon,$$

are presented in Figure C.2. Due to the large number of observations, we are not able to use the residual plots to discover trends in the distribution of residuals.



(a) Women, no transform

(b) Women, log-transform

(c) Men, no transform

(d) Men, log-transform

Figure C.2: Residuals

Appendix D

R code

D.1 Score tests

Cross-sectional

```
scorenormal=function(phenotypes,covariates,genotypes){
  y=phenotypes; x=covariates; g=genotypes
  n=length(y); px=dim(x)[2]; pg=dim(g)[2]
  fit=mlesnormal(y,x)
  alpha0=fit[1]; alpha=fit[2:(length(fit)-1)];
  sigma=fit[length(fit)]; sigma2=sigma*sigma

  I11_11=n; I11_22=matrix(0,nrow=px,ncol=px)
  for (i in 1:n){I11_22=I11_22 + x[i,]%*%t(x[i,])}; I11_33=2*n
  I11_21=n*matrix(colMeans(x)); I11_12=t(I11_21)
  I11_31=0; I11_13=0
  I11_32=matrix(0,nrow=1,ncol=px); I11_23=t(I11_32)
  I11=cbind(rbind(I11_11,I11_21,I11_31),
            rbind(I11_12,I11_22,I11_32),
            rbind(I11_13,I11_23,I11_33))

  I22=matrix(0,nrow=pg,ncol=pg);
  for (i in 1:n){I22=I22 + g[i,]%*%t(g[i,])}

  I21_1=matrix(n*colMeans(g)); I21_3=matrix(0,nrow=pg,ncol=1)
  I21_2=matrix(0,nrow=pg,ncol=px)
  for (i in 1:n){I21_2=I21_2 + g[i,]%*%t(x[i,])}
  I21=cbind(I21_1,I21_2,I21_3)

  sigmastar=(1/sigma2)*(I22 - I21%*%ginv(I11)%*%t(I21))
  s=matrix(0,nrow=pg,ncol=1)
  for (i in 1:n){s=s + (y[i]-alpha0-alpha%*%x[i,])*g[i,]}
  s=s/sigma2
  t=t(s)%*%ginv(sigmastar)%*%s
  pval=pchisq(t,pg,lower.tail=F)
  return(pval)
}
```

Conditional

```
scoreconditional=function(phenotypes,covariates,genotypes,l,u){
  y=phenotypes; x=covariates; g=genotypes
  n=length(y); px=dim(x)[2]; pg=dim(g)[2]
  fit=mlesconditional(y,x,l,u)
  alpha0=fit[1]; alpha=fit[2:(length(fit)-1)]
  sigma=fit[length(fit)]; sigma2=sigma*sigma

  ax=x%*%alpha
  zl=(l-alpha0-ax)/sigma; zu=(u-alpha0-ax)/sigma

  h0=(-dnorm(zu)+dnorm(zl))/(1-pnorm(zu)+pnorm(zl))
  h1=(-dnorm(zu)*zu+dnorm(zl)*zl)/(1-pnorm(zu)+pnorm(zl))
  h2=(-dnorm(zu)*zu*zu+dnorm(zl)*zl*zl)/(1-pnorm(zu)+pnorm(zl))
  h3=(-dnorm(zu)*zu*zu*zu+dnorm(zl)*zl*zl*zl)/(1-pnorm(zu)+pnorm(zl))

  a=1-h1-h0*h0
  b=2*h0 - h2 - h0*h1
  c=2 + 3*h1 - h3 - h1*h1

  I11_11=sum(a); I11_22=matrix(0,nrow=px,ncol=px)
  for (i in 1:n){I11_22=I11_22 + a[i]*x[i,]%*%t(x[i,])}
  I11_21=matrix(0,nrow=px,ncol=1)
  for (i in 1:n){I11_21=I11_21 + a[i]*x[i,]}; I11_12=t(I11_21)
  I11_33=sum(c); I11_31=sum(b); I11_13=sum(b)
  I11_23=matrix(0,nrow=px,ncol=1)
  for (i in 1:n){I11_23=I11_23 + b[i]*x[i,]}; I11_32=t(I11_23)
  I11=cbind(rbind(I11_11,I11_21,I11_31),
            rbind(I11_12,I11_22,I11_32),
            rbind(I11_13,I11_23,I11_33))

  I22=matrix(0,nrow=pg,ncol=pg)
  for (i in 1:n){I22=I22 + a[i]*g[i,]%*%t(g[i,])}

  I21_1=matrix(0,nrow=pg,ncol=1)
  for (i in 1:n){I21_1=I21_1 + a[i]*g[i,]}
  I21_2=matrix(0,nrow=pg,ncol=px)
  for (i in 1:n){I21_2=I21_2 + a[i]*g[i,]%*%t(x[i,])}
  I21_3=matrix(0,nrow=pg,ncol=1)
  for (i in 1:n){I21_3=I21_3 + b[i]*g[i,]}
  I21=cbind(I21_1,I21_2,I21_3); I12=t(I21)
  sigmastar=(1/sigma2)*(I22 - I21%*%ginv(I11)%*%t(I21))
  s=matrix(0,nrow=pg,ncol=1)
  for (i in 1:n){s=s + (y[i]-alpha0-ax[i] + sigma*h0[i])*g[i,]}
  s=s/sigma2
  t=t(s)%*%ginv(sigmastar)%*%s
  pval=pchisq(t,pg,lower.tail=F)
  return(pval)
}
```

D.2 Rare variant methods

The collapsing method

```
collapsing=function(phenotypes,covariates,genotypes){
  collapsedgens=rowSums(genotypes)
  collapsedgens[collapsedgens>0]=1
  pval=scorenormal(phenotypes,covariates,as.matrix(collapsedgens))
  return(pval)
}
collapsingconditional=function(phenotypes,covariates,genotypes,l,u){
  collapsedgens=rowSums(genotypes)
  collapsedgens[collapsedgens>0]=1
  pval=scoreconditional(phenotypes,covariates,
                        as.matrix(collapsedgens),l,u)
  return(pval)
}
```

The CMC method

```
collapse=function(genotypes,mafs){
  uncollapsed=genotypes[,mafs>=0.05]
  collapsedgens=rowSums(genotypes[,mafs<0.05])
  collapsedgens[collapsedgens>0]=1
  return(cbind(uncollapsed,collapsedgens))
}
cmc=function(phenotypes,covariates,genotypes,mafs){
  newgen=collapse(genotypes,mafs)
  pval=scorenormal(phenotypes,covariates,as.matrix(newgen))
  return(pval)
}
cmcconditional=function(phenotypes,covariates,genotypes,mafs,l,u){
  newgen=collapse(genotypes,mafs)
  pval=scoreconditional(phenotypes,covariates,as.matrix(newgen),l,u)
  return(pval)
}
```

The β-SO method

```
betasmooth=function(phenotypes,covariates,genotypes,positions){
  bbasis=15; order=4
  betabasis =create.bspline.basis(norder=order,nbasis=bbasis)
  B=eval.basis(positions,betabasis)
  UJ=genotypes %*% B
  pval=scorenormal(phenotypes,covariates,UJ)
  return(pval)
}
betasmoothconditional= function(phenotypes,covariates,
                                  genotypes,positions,l,u){
  bbasis= 15; order =  4
  betabasis =create.bspline.basis(norder=order,nbasis=bbasis)
  B=eval.basis(positions,betabasis)
  UJ=genotypes %*% B
  pval=scoreconditional(phenotypes,covariates,UJ,l,u)
  return(pval)
}
```

Using the SKAT methods in a cross-sectional sample

```
  obj=SKAT_Null_Model(pheno ~ cov,out_type="C")
  pval_skat=SKAT(geno,obj)$p.value
  pval4_skato=SKAT(geno,obj,method="optimal.adj")$p.value
```

Using the SKAT methods in an EPS sample

```
  obj=SKAT_Null_Model_CEP(pheno ~ cov,l,u)
  pval_skat=SKAT(geno,obj)$p.value
  pval_skato=SKAT(geno,obj,method="optimal.adj")$p.value
```