NTNU
Norwegian University of
Science and Technology

Statistical methods to detect genotype-phenotype association using genetic similarity matrices

Lene Maria Sundbakk

Master of Science in Physics and Mathematics
Submission date: June 2016
Supervisor: Mette Langaas, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem description

The objective of this Master's thesis is to perform a genotype-phenotype association analysis for maximal oxygen uptake and SNP data. The data set under study consists of a cohort of 1472 men from the HUNT $VO_{2_{max}}$ study. No information about relatedness of these individuals is available, and the challenging part of the data analysis is to correctly account for unknown and cryptic relatedness among the study participants and for possible population substructure. Failure to properly account for this can lead to spurious association or reduced power. Relatedness among the study participants will be estimated from the GWA data. Then, two different statistical models will be pursued. Firstly, a reduced data set is constructed by removing individuals that are estimated to have a high degree of relatedness. This reduced data set is then assumed to be a sample of independent individuals, and a multiple linear regression model is fitted, with maximal oxygen uptake as the response, and age and physical activity index as covariates. Each SNP is then tested for association with the response using a score test, and the family-wise error rate is controlled using the method of Halle et al. (2016). Secondly, the original data set is analyzed using a linear mixed model with the same fixed effects as in the multiple linear regression model, but including correlated random effects for each individual and using twice the estimated kinship matrix from GWA data as the covariance matrix of the random effect. Also for this model each SNP is tested separately, along the same lines as for the first analysis. Finally, the results from the two analyses are compared.

Preface

This Master's thesis constitutes the course TMA4905 - Statistics for the Industrial Mathematics program at NTNU. The topic of this thesis, detecting phenotype–genotype associations using genetic similarity matrices estimated from the HUNT data set, evolved from the mandatory project of the course TMA4500, written in the autumn of 2015. The main focus of the TMA4500 project was estimation of the kinship coefficient, and a comparison of different estimators. I would like to thank my supervisor Mette Langaas at the Department of Mathematical Sciences for excellent guidance and motivation in the process of writing this thesis. I would also like to thank Anja Bye at the Department of Circulation and Medical Imaging for making the HUNT data set available to me and for support in understanding the genetic part of the results.

Lene Maria Sundbakk
Trondheim, Norway
June, 2016

Abstract

The main focus of this thesis is to investigate and compare different statistical methods to perform genotype–phenotype association analyses of HUNT $VO_{2_{max}}$ data from 1472 men, while accounting for genetic confounding. The methods of interest are fitting a linear regression model to a reduced sample of 1274 men, and fitting a linear mixed model to the full sample, with maximal oxygen uptake as response, and age and activity index as covariates. The covariance matrix of the linear mixed model is a scaled version of an estimated genetic similarity matrix, the *kinship matrix*, estimated from 102 477 SNPs. Each SNP is then tested for association with the response using a score test. The analyses are performed using `GenABEL`, which is an R-package for statistical analyses of genome-wide association studies. We analyze only the 9069 SNPs on chromosome 1. The results from both methods show no significant associations between any SNP and the $VO_{2_{max}}$, when controlling the family-wise error rate at level 0.05. Based on the results of the most significant SNPs and estimation of the genomic control inflation factor for both methods, we find that the preferred procedure for performing genotype–phenotype association analyses is using linear mixed models. The incorporation of the estimated kinship matrix accounts for the correlation between the individuals caused by population structure and cryptic relatedness.

Sammendrag

I denne oppgaven ser vi på ulike statistiske metoder for å utføre genetiske assosiasjonsanalyser samtidig som det tas hensyn til korrelasjonen mellom individene i studien. Vi benytter et datasett fra HUNT (Helseundersøkelsen i Nord-Trøndelag) med 1472 menn, for å se etter mulige sammenhenger mellom genetiske markører og maksimalt oksygenopptak ($VO_{2_{max}}$). Metodene vi undersøker er å tilpasse en lineær regresjonsmodell til et redusert datasett med 1274 menn, og å tilpasse en lineær mixed modell til det originale datasettet. Responsen er $VO_{2_{max}}$, og kovariatene er alder og aktivitetsnivå. Kovariansmatrisen i den lineære mixed-modellen er en skalert versjon av en estimert genetisk likhetsmatrise, også kalt kinship-matrisen, estimert fra 102 477 SNPer. Hver SNP er testet for assosiasjon med responsen ved å bruke en scoretest. Analysene er utført ved Ãě benytte `GenABEL`, som er en R-pakke med funksjoner for å utføre statistiske analyser av genetiske assosiasjonsstudier. Vi analyserer kun de 9069 SNPene på kromosom 1. Resultatene fra begge metodene viser at det ikke er signifikant assosiasjon mellom noen SNPer og $VO_{2_{max}}$, når det korrigeres for multippel testing. Basert på resultatene for de mest signifikante SNPene og estimert inflasjonsfaktor for genomisk kontroll for begge metodene, mener vi at den foretrukne metoden for å analysere genetiske assosiasjonsdata er å tilpasse en lineær mixed modell til datasettet. Modellen inkorporerer den estimerte kinshipmatrisen, som fører til at korrelasjonen mellom individene blir tatt i betraktning.

Contents

Chapter 1

Introduction

1.1 Motivation

Genetic association studies are statistical studies of relationships between individuals' genotypes and phenotypes, which seek to examine if genetic variants are associated with the phenotype. It is common to exclude close family members in a *genome-wide association* (GWA) analysis, to be able to analyze data from a sample of independent individuals. However, GWA studies are sensitive to genetic confounding, which can come in the forms of *cryptic relatedness* and *population structure*.

Cryptic relatedness refers to the idea that some individuals in a sample might actually be close relatives, in which their genotypes are not independent draws from the population frequencies. Cryptic relatedness may be noticeable for small and isolated populations. Population structure is used to describe the subdivision of the population. Instead of a single, simple, panmictic population, the population has large-scale systematic differences in ancestry, and groups of individuals share more recent ancestors than expected (Astle and Balding, 2009).

The aim of this Master's thesis is to study different statistical methods to perform a GWA analysis to identify genetic loci associated with maximal oxygen uptake, $VO_{2_{max}}$. The data set from the HUNT $VO_{2_{max}}$ study contains information for 1472 men, but no information about relatedness among the individuals is part of the data set. The demanding part of the analysis is to properly account for correlation between the individuals, arising from cryptic relatedness and population structure. Failure to properly account for genetic confounding can lead to spurious associations or reduced power.

Each pair of individuals' degree of relationship will be estimated from the GWA data, and all the coefficients are arranged in a genetic similarity matrix - the kinship matrix. Based on the coefficients in this matrix it is possible to create a reduced sample, by removing individuals with a high degree of relatedness. This

reduced sample is then assumed to be a sample of independent individuals, which can be analyzed using a linear regression model, with $VO_{2_{max}}$ as response, and age and physical activity index as covariates. Each SNP will be tested separately for association with $VO_{2_{max}}$, and the family-wise error rate is controlled using the method of Halle et al. (2016).

The original data set will be analyzed by fitting a linear mixed model to the data, using the same fixed effects as in the analysis of the reduced sample, but including correlated random effect for each individual. The covariance matrix of the random effect is twice the estimated kinship matrix. Each SNP is tested separately for association with the maximum oxygen uptake, using the same local significance level as for the analysis of the reduced sample.

Lastly, it is of interest to compare the different methods for analyzing GWA data, and their ability to correctly account for genetic confounding.

1.2 The HUNT study

The Nord-Trøndelag Health Study (the HUNT study) is one of the largest health studies ever performed, and Norway's biggest collection of health information from a population. The data are obtained from three studies; HUNT1 (1984-1986), HUNT2 (1995-1997) and HUNT3 (2006-2008), and altogether 120 000 people have been a part of the studies[1]. The data from the HUNT studies are used in several research projects within several subject areas concerning health and diseases.

The data set analyzed in this thesis is from the HUNT3 study, which had 58 000 participants[2]. The participants were asked to answer questionnaires, clinical surveys were performed and different tests were taken. The participants in the study are seen to be a good representation of trends in the Norwegian population. All participants were inhabitants of Nord-Trøndelag county at the time of participation, that means a limited geographical area. The population is relatively homogeneous and stable and for some areas of research the trends could also be valid for other Caucasian populations[3].

Among the tests taken, blood tests were taken of each participant, in order to analyze cholesterol, glucose level and metabolism. The blood samples are also used for extraction of DNA, which is applied in research projects to investigate the relation between inheritance and environment in the progress of a disease. The

[1] NTNU.no (2016). Om HUNT - Helseundersøkelsen i Nord-Trøndelag - NTNU. Available at: http://www.ntnu.no/hunt/om [Accessed 15 Jun. 2016].

[2] NTNU.no(2016). HUNT3 - Helseundersøkelsen i Nord-Trøndelag - NTNU. Available at: http://www.ntnu.no/hunt/hunt3 [Accessed 15 Jun. 2016].

[3] NTNU.edu (2016). HUNT Databank - NTNU. Available at: http://www.ntnu.edu/hunt/databank [Accessed 15 Jun. 2016].

participants were asked to be a part of one or several other tests, like breathing test, urine specimen or physical test.

Physical fitness is important to be in good health. Until recently there has been no robust material that describes the distribution of maximal oxygen uptake across a healthy, adult population[4]. In order to discover this, and to further study the connection between physical fitness and health, some HUNT3 participants were asked to take part in the maximal oxygen uptake test. This test aimed at running or walking on a tread mill until complete exhaustion. 3796 individuals was a part of the physical test where the maximal oxygen uptake was measured. Firstly, 1472 men was genotyped, and then in the second round, 2324 men and women were participating, but only the ones not related to others were genotyped. 295 people were then excluded from the analysis.

In order to analyze both the sample including related individuals and the reduced sample of unrelated individuals, genotype data of all individuals are needed. Consequently, only the 1472 men from the first round of genotyping are part of the analysis in this thesis. When the Cardiac Exercise Research Group [5](CERG) analyzed the HUNT3 $VO_{2_{max}}$ data, they used data from 3470 individuals not related (Bye et al., 2016).

The data set analyzed in this thesis consists of maximal oxygen uptake ($VO_{2_{max}}$), age and physical activity for the 1472 men, and genotype data from 196 725 SNPs. $VO_{2_{max}}$ is measured as millilitres of oxygen per minute (ml/min), and subsequently calculated as $VO_{2_{max}}$ relative to the scaled body mass ($ml/kg^{0.75}/min$).

The physical activity of a person is likely to be the most important behavioural factor influencing $VO_{2_{max}}$, and is therefore an important confounder to adjust for when analysing the genetic contribution to the phenotype. The physical activity of each participant was registered based on the responses from questionnaires. The questionnaires included three questions and each participants' responses to the questions (i.e. numbers in parentheses) were multiplied to calculate a physical activity index score:

- *Question 1: How often do you exercise?*, with the response options: *Never* (0), *Less than once a week* (0), *Once a week* (1), *2-3 times a week* (2.5) and *Almost every day* (5).

- *Question 2: If you exercise as frequently as once or more times a week, how hard do you push yourself?*, with the response options: *I take it easy without breaking a sweat or losing my breath* (1), *I push myself so hard that I lose my breath and break into sweat* (2) and *I push myself to near exhaustion* (3).

[4]NTNU.edu (2016). Fitness numbers - CERG - NTNU. Available at: http://www.ntnu.edu/cerg/fitness-numbers [Accessed 15 Jun. 2016].

[5]NTNU.edu (2016). Cardiac Exercise Research Group - NTNU. Available at: http://www.ntnu.edu/cerg [Accessed 15 Jun. 2016].

- *Question 3: How long does each session last?* with the response options: *Less than 15 minutes* (0.1), *16-30 minutes* (0.38), *30 minutes to 1 hour* (0.75) and *More than 1 hour* (1.0).

As the second and third question only addressed people who exercised at least once a week, both *Never* and *Less than once a week* as a response to question one yielded an index score of zero. Participants with a zero score were categorized as inactive, 0.05-1.5 as low activity, 1.51-3.75 as medium activity, and 3.76-15.0 as high activity. We will use the scores as covariates in the analysis.

Maximal oxygen uptake

Maximal oxygen uptake is the highest oxygen (O_2) uptake that can be achieved by an individual during exercise with dynamic use of a large muscle mass. It is considered as the best indication of cardiorespiratory capacity. The higher $VO_{2_{max}}$, the more O_2 has been transported to and used by exercising muscles, which increases the level of intensity at which the individual can exercise.

$VO_{2_{max}}$ is determined both by genetic and environmental factors. The untrained fitness level is partly based on genetic factors, but the genetic factors also contribute to the potential of training-induced improvements (Bye, 2008). The HERITAGE Family Study examined the genetic contribution to the individual response to endurance training (Bouchard et al., 1999), and reported a significantly association between genetic components and the trainability of $VO_{2_{max}}$. This means that the variation in the human population's ability to improve $VO_{2_{max}}$ by exercise is large. The heritability of the $VO_{2_{max}}$ response to training was estimated to be 47%.

The CERG researchers observed that the mean maximal oxygen uptake in women and men were 35 and 44 $ml/kg/min$, respectively[6]. The material suggested a $\sim 7\%$ decline in maximal oxygen uptake with every 10 year raise in age in both genders.

[6]NTNU.edu (2016). Fitness numbers - CERG - NTNU. Available at: http://www.ntnu.edu/cerg/fitness-numbers [Accessed 15 Jun. 2016].

1.3 Structure of the thesis

In Chapter 2 we give an introduction to the basics of genetics needed in order to understand the estimators of the kinship coefficient and the results of the analyses in Chapter 4. We explain the concept of linkage and define single nucleotide polymorphisms (SNPs). We also introduce the expected kinship coefficient for pedigree data, and the kinship coefficient estimated from SNP data.

In Chapter 3 we present the theory of the statistical models and methods used in the analysis of the GWA data. We define linear models and linear mixed models, hypothesis testing and methods to correct for multiple testing. A comparison of principal components regression and linear mixed models is also presented, as well as the method of genomic control.

In Chapter 4 we perform the analysis of the HUNT $VO_{2_{max}}$ data set, applying the methods of linear model regression and linear mixed model regression presented in Chapter 4. In the end of the chapter we compare the results of the different methods, and give a genetic interpretation of the results.

Finally, the thesis ends with discussion and conclusions in Chapter 5.

Chapter 2

Introduction to genetics

This chapter will give an introduction to basic concepts of genetics. We will introduce the concept of kinship, and present formulas to calculate the kinship coefficient for a pair of individuals from a pedigree. The last section presents an estimator for the kinship coefficient based on SNP data.

2.1 Genes and inheritance

The history of genetics started with the work of Gregor Mendel in the 19th century. He postulated his first law in 1866, based on his analysis of pea plants. The first law is the law of *segregation* (separation), which states that in gamete formation the two copies of the same gene segregate so that each gamete receives only one copy. One is a copy of a corresponding gene in the father of the individual, and the other is a copy of a gene in the mother of the individual. The probabilistic process of the random choice of genes to be copied is known as Mendelian segregation. The biological process forming the chromosomes of the gamete (sperm or egg) cell is known as *meiosis*. Mendel's second law states that alleles at any one gene segregate independently of alleles at any other gene (Thompson, 2000; Fletcher and Hickey, 2013).

DNA contains the biological information needed by and organism to reproduce. The information is encoded in the base sequence of the DNA and is organized as a large number of genes. DNA is located in the nuclei of cells in individuals, and consists of about 3 x 10^9 base pairs, which is packed into *chromosomes*. A human individual has 46 chromosomes in each cell; 22 pairs of *autosomes* and a pair of *allosomes* (sex chromosomes). The *diploid* cells have two homologous copies of each chromosome. Homologous means that the chromosomses have the same size and shape, and carries the same genes in the same order. One of the copies is from the mother and one from the father, thus 46 chromosomes, while the *haploid* cells have only one copy, thus 23 chromosomes. The genes occur at specific positions

along the chromosome, defined as *loci*. The DNA at a locus may come in different variants, or *alleles*. Any human has two chromosomes of a given pair, and thus has two (possibly identical) alleles at each locus. A locus is a *polymorphism* if it has two or more alleles at appreciable frequencies. The two most used definitions of appreciable frequency is that the most frequent allele has a frequency of less than 0.99, or less than 0.95 (Halliburton, 2004). A single homologous chromosome or chromosomal region can be described in terms of the alleles it possesses, also referred to as a *haplotype* (Thompson, 1986, 2000; Fletcher and Hickey, 2013).

A *phenotype* is any character (trait) that is inherited, such as body height, eye color, leaf shape, or a disease. The genes carried by individuals may be labelled according to type, depending on their effects on observed traits. The alleles at each locus constitutes an individual's *genotype* at that locus (Lange, 2003). A *homozygous* locus is a locus with identical alleles, while a *heterozygous* locus has two different alleles. To explain this more in detail, and to present the concepts of recessive and dominant, an illustration of the Punnett square is shown Figure 2.1. The figure presents the crossing of two pea plants with violet flowers, both heterozygous at the locus for flower color. The alleles for these two pea plants are at the top edge and left edge of the square. It is clear that the genotype of the offspring is either VV, Vv or vv and the corresponding phenotype is either violet or white. The V allele is *dominant* to the v allele, which is why the heterozygote individuals exhibit the same phenotype as one of the homozygotes. The v allele is said to be *recessive* to the V allele. The individuals with genotypes VV and Vv at this locus both have violet flowers as their phenotype, while the individuals with genotype vv have white flowers (Fletcher and Hickey, 2013).

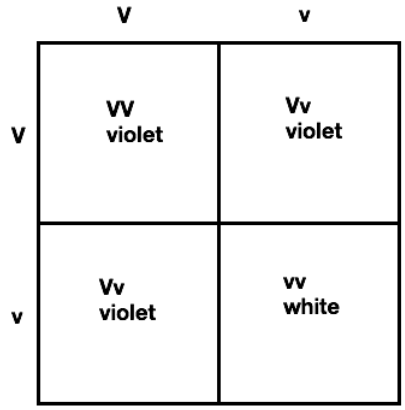

Figure 2.1: Punnett square for the crossing between two pea plants with violet flowers, thus heterozygous for flower color. The phenotype ratio is (3 violet):(1 white), and the genotype ratio is $(1VV)$:$(2Vv)$:$(1vv)$. Created with inspiration from Fletcher and Hickey (2013).

2.1.1 The Hardy-Weinberg principle

Mendel's laws provided a mechanism of heredity that preserves genetic variation. This was not immediately obvious, and many initially thought that the dominant form would take over the population. In 1908 G. H. Hardy and Wilhelm Weinberg derived a principle that is now known as the *Hardy-Weinberg* principle. Halliburton (2004) defines the The Hardy-Weinberg principle as a proposition which says that, under certain conditions, allele frequencies and genotype frequencies will remain constant in a population, and that they are related in specific ways. Consider the two alleles from Section 2.1, V and v, with frequencies p and q, respectively ($p + q = 1$). The genotypes are VV, Vv, and vv, with corresponding frequencies P_{VV}, P_{Vv} and P_{vv}, respectively. The underlying conditions for the principle are:

1. The reproduction within the population is sexual.
2. The population is a *panmictic* population, which means that individuals mate at random.
3. Natural selection is not affecting the locus.
4. Mutation is negligible.
5. Individuals do not move into or out of the population.
6. The population is infinitely large.

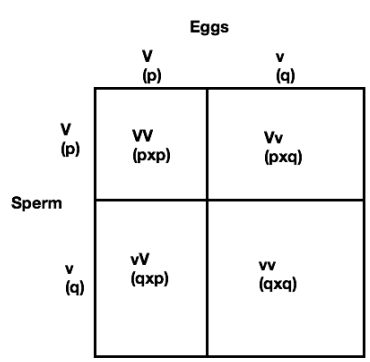

Figure 2.2: Random union of gametes at an autosomal locus with two alleles, V and v. Created with inspiration from Halliburton (2004).

With these assumptions, we define the probability of a V sperm fertilizing a V egg as $p \cdot p = p^2$. Similarly, the probability of a V sperm fertilizing a v egg is $p \cdot q$, and so forth for all possible combinations, shown in Figure 2.2, where the entries in the square are the genotypes of the zygotes, and their frequencies. From

this figure it is clear that the genotype Vv will follow from two different matings. Both the V sperm + v egg and V egg + v sperm will result in a heterozygote. A new generation of zygotes are denoted $(t+1)$, and thus it is seen from the square in Figure 2.2 that the expected genotype frequencies are

$$P_{VV}(t+1) = p^2$$
$$P_{Vv}(t+1) = 2pq$$
$$P_{vv}(t+1) = q^2.$$

From these equations it is clear that the genotype frequencies depend on the allele frequencies. The allele frequencies in the zygotes will remain unchanged, as seen in the following equation,

$$p(t+1) = P_{VV}(t+1) + \frac{1}{2}P_{Vv}(t+1) = p^2 + \frac{1}{2}(2pq) = p(p+q) = p.$$

The allele frequencies and genotype frequencies will remain constant as long as the assumptions above hold (Halliburton, 2004).

2.1.2 Linkage

Along the human genome we find thousands of loci, which may interact with one another to a greater or lesser degree. *Recombination* is defined according to Halliburton (2004) as any process that creates new combinations of alleles in the offspring. He also states that two loci are linked if they are on the same chromosome, close enough together that the frequency of recombination between them is less than 50%. Pairs of genetic loci that are not tightly linked are unassociated at the population level. Astle and Balding (2009) say that such a *linkage equilibrium* arises because recombination events ensure the independent assortment of alleles when they are transmitted across generations, and that tightly linked loci are generally correlated, or in *linkage disequilibrium* in the population. Linkage disequilibrium is the non-random association of alleles at different loci. The association is different from what would be expected if alleles were independently, randomly samples based on their individual allele frequencies.

A convenient measure of the linkage disequilibrium is r^2. Consider two biallelic loci, in which locus 1 has alleles a and A and locus 2 has alleles b and B. We suppose that the frequencies of the alleles a and A are p_a and $1-p_a$, respectively, and the frequencies of the alleles b and B are p_b and $1-p_b$, respectively. The frequency of haplotypes having the a allele at locus 1 and the b allele at locus 2 is p_{ab}. The linkage disequilibrium measure r^2 is according to VanLiere and Rosenberg (2008) defined as

$$r^2 = \frac{(p_{ab} - p_a p_b)^2}{p_a(1-p_a)p_b(1-p_b)}. \tag{2.1}$$

The measure ranges between 0 and 1, and a value of 0 indicates that the loci are in perfect equilibrium, while a value of 1 indicates that the loci provide identical information.

2.1.3 Single nucleotide polymorphisms

The organic molecules that serve as monomers, or subunits, of the DNA, are defined as *nucleotides*. Each nucleotide is composed of a sugar, a nitrogen-containing ring-structure called a base, and a phosphate group. The four possible bases for a nucleotide are adenine (A), guanine (G), thymine (T) and cytosine (C). Adenine and guanine are known as purines, while thymine and cytosine are called pyrimidines. DNA molecules are composed of two strands wrapped around each other to form a double helix. The sugar-phosphate part of the molecule forms a backbone, while the bases face inwards and are stacked on top of each other. There are hydrogen bonds between the bases on the two strands which stabilizes the double helix. The available space between the strands restricts the bases that can interact such that a purine always interacts with a pyrimidine. Thus, A interacts only with T, and G only with C, and this is called complementary base pairing. This means that the sequence of one strand determines and predicts the sequence of the other.

The human *genome* is a term used to describe the different types of sequences that together make up all the DNA in a human cell. The DNA in the human genome is about 3 billion base pairs long, and is estimated to contain 20 000 to 21 000 genes (Fletcher and Hickey, 2013, p.68). Mutations of a single base pair which show variation in the population are called *single nucleotide polymorphisms* (SNPs). The '1000 genome project' have identified 15 million SNPs, which corresponds to one every 200 base pairs on average (Fletcher and Hickey, 2013).

Some SNP alleles are the actual functional variants that contribute to a specific phenotype. Individuals with such a SNP allele have a higher chance of having that phenotype than do individuals without that SNP allele. Most SNPs are not these functional variants, but are useful markers for finding them. To find these regions with genes that contribute to a phenotype, the frequencies of many SNP alleles are compared in individuals with and without the specific phenotype. When a particular region has SNP alleles that are more frequent in individuals with the phenotype than in individuals without the phenotype, those SNPs and their alleles may be associated with the phenotype (Fletcher and Hickey, 2013).

The term *minor allele frequency* (MAF) is used to describe the frequency of the allele with the lowest frequency at a locus, within a population (Speed and Balding, 2015).

2.1.4 Identical by descent and identical by state

The attention is now restricted to a single Mendelian autosomal locus. The use of the word *relatives* corresponds to blood relatives, individuals with a common ancestor. Any given set of individuals may carry the same alleles; there are many copies of an allele in the population. Two copies of the same allele are assumed to have the same function, and such functionally equivalent alleles are designated as *identical by state* (IBS). However, relatives are more likely to share the same alleles, because they may carry copies of a single gene inherited from a common ancestor. Alleles that are *identical by descent* (IBD) are matching DNA-segments that two individuals share, inherited from a recent common ancestor without any recombination occurred. If two alleles are IBD, it entails that they are aslo IBS unless mutation has altered the function of one (Thompson, 1986; Halliburton, 2004). Halliburton (2004) provides the statement that whether two alleles are considered identical by descent depends on how far back in time we follow the population.

2.1.5 Inbreeding

Inbreeding is a term used to describe mating between related individuals. These individuals share a common ancestor, and thus have alleles that are IBD. One of the consequences of inbreeding in a population is increasing frequency of homozygous genotypes. The probability that two copies of a gene are IBD is called the inbreeding coefficient, symbolized by f, with subscript to indicate the individual (Halliburton, 2004).

2.2 Kinship coefficient and kinship matrix from pedigree data

2.2.1 Pedigree

A *pedigree* is a diagram that is used to record ancestry. Males are denoted by squares and females by circles, as seen in the pedigree in Figure 2.3. A pair of parents is connected by a horizontal line, and a vertical line descends to the offspring. If there is more than one child of the same pair of parents, these are shown under a second horizontal sibship line. Between two individuals, the degree of their relationship, deriving from a single ancestor or ancestral couple, relates to the number of generations that separate the ancestor from the individuals (Thompson, 1986).

For the family of individuals in Figure 2.3, the individuals are numbered from 1 to 9, where individuals 1 and 2 are called founders, since they don't have any parents specified. The pedigree embodies several different relations between individ-

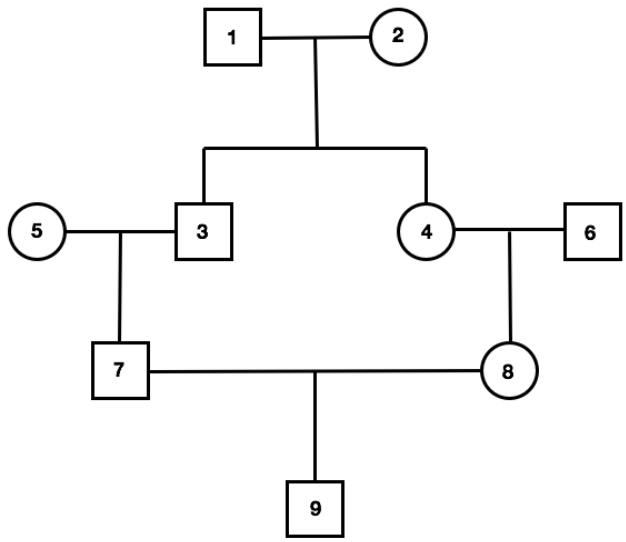

Figure 2.3: Pedigree, cousin mating.

uals; parent-offspring, full siblings, grandparent-grandchild, first cousins, parent-offspring (inbred), grandparents-grandchild (inbred) and great-grandparent-great-grandchild (inbred). The individuals in the pedigree are numbered in such a way that every parent precedes his or her children; every individual are numbered higher than its ancestors. Take individual 7 for example; his parents are numbered 5 and 3, his aunt and uncle are numbered 4 and 6, and his grandparents are numbered 1 and 2.

2.2.2 The kinship coefficient

To measure the degree of relationship between two individuals, the *kinship* coefficient is a useful measure. It is denoted by Φ, usually with subscripts to indicate the individuals involved. The kinship coefficient Φ_{ij} between two individuals i and j is defined as the probability that an allele chosen randomly from i is identical to a homologous allele chosen at random from j (Thompson, 1986). This is the same for each autosomal locus, thus $\Phi_{ij} = \Phi_{ji}$. The coefficient of consanguinity, coefficient of relatedness or the coefficient of coancestry are other terms used for the kinship coefficient (Halliburton, 2004), but we will in the following use kinship coefficient.

The kinship coefficient pertains to a generic autosomal locus and depends only on the relevant pedigree connecting two relatives, and not on any phenotypes

observed in the pedigree. When i and j are the same person, the same gene can be drawn twice because kinship sampling is done with replacement. In the following, necessary rules for computing the kinship coefficient will be specified, and the kinship coefficient for each relation in the pedigree in Figure 2.3 will be calculated.

The kinship coefficient can also be used to describe the inbreeding coefficient of an individual i, f_i, defined in Section 2.1.5. Since the alleles of an inbred individual are chosen at random, one from each of its parents, the inbreeding coefficient of an individual is the same as the kinship coefficient of its parents (Halliburton, 2004).

A systematic way of presenting the relationship between every possible pair of individuals in a pedigree is by a genetic similarity matrix, the *kinship matrix* $\mathbf{\Phi}$, with element Φ_{ij} in row i and column j. Assuming that all pedigree founders in Figure 2.3 are non-inbred and unrelated, then $\Phi_{11} = \frac{1}{2}$, and $\Phi_{22} = \frac{1}{2}$, since the kinship sampling is done with replacement. The founders are assumed to be unrelated, thus the kinship coefficient between them is $\Phi_{12} = \Phi_{21} = 0$ (Lange, 2003). To continue on the diagonal of the kinship matrix $\mathbf{\Phi}$, the formula for the kinship coefficient for individual i with itself (Lange, 2003), when i is not a founder, and i have parents k and l, is

$$\Phi_{ii} = \frac{1}{2} + \frac{1}{2}\Phi_{kl}. \tag{2.2}$$

This formula is easy to understand because in sampling the alleles of i we are equally likely to choose either the same allele twice or both maternally and paternally derived alleles once. Looking at the formula, Φ_{kl} is recognized as f_i, the inbreeding coefficient of individual i. When the parents k and l are not related, $\Phi_{kl} = f_i = 0$, thus $\Phi_{ii} = \frac{1}{2}$.

To compute the kinship coefficient Φ_{13} for the individuals 1 and 3 in the pedigree in Figure 2.3, the following formula is needed (Halliburton, 2004). When i is a founder, and the kinship coefficient between i and a descendant of i, j, depends on n, the number of segregations between the individuals, then

$$\Phi_{ij} = \left(\frac{1}{2}\right)^n \tag{2.3}$$

It follows from Equation (2.3) that $\Phi_{13} = \left(\frac{1}{2}\right)^2 = \frac{1}{4} = \Phi_{31}$, since there are two segregations between the individuals. The kinship coefficient for all parent-offspring relations are computed in the same way, when assuming that the parents are not inbred.

The matrix $\mathbf{\Phi}$ is fully defined with the elements

$$\Phi_{ij} = \Phi_{ji} = \frac{1}{2}\Phi_{jk} + \frac{1}{2}\Phi_{jl}, \tag{2.4}$$

which follows because we are equally likely to compare either the maternal allele of i or the paternal allele of i to a randomly drawn allele from j (Lange, 2003). This formula is also applicable for inbred individuals.

Many of the individuals in the pedigree are unrelated, which corresponds to a kinship coefficient equal to zero. The full-sibling kinship coefficient follows from Equation (2.4), e.g. $\Phi_{34} = \Phi_{43} = \frac{1}{2}\Phi_{41} + \frac{1}{2}\Phi_{42} = \frac{1}{4}$, similarly for the first cousins kinship coefficient, e.g. $\Phi_{78} = \Phi_{87} = \frac{1}{2}\Phi_{83} + \frac{1}{2}\Phi_{85}$, where $\Phi_{83} = \frac{1}{2}\Phi_{34} + \frac{1}{2}\Phi_{36} = \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot 0 = \frac{1}{8}$ and $\Phi_{85} = 0$, so $\Phi_{78} = \Phi_{87} = \frac{1}{2} \cdot \frac{1}{8} = \frac{1}{16}$.

For the pedigree in Figure 2.3, it is clear that there is a cousin-mating between individual 7 and 8, resulting in the offspring 9. Computing the kinship coefficients for these individuals get more interesting. Using Equation (2.4), the kinship coefficient for the relation parent-child, while the child has parents who are cousins, as between individual 9 and individual 7, is $\Phi_{79} = \Phi_{97} = \frac{1}{2}\Phi_{77} + \frac{1}{2}\Phi_{78} = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{16} = \frac{9}{32}$. Using Equation (2.2) for individual 9, $\Phi_{99} = \frac{1}{2} + \frac{1}{2}\Phi_{78} = \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{16} = \frac{17}{32}$. For the relation grandparent-grandchild, while the grandchild is inbred, like between individual 3 and individual 9, Equation (2.4) is applied again; $\Phi_{39} = \Phi_{93} = \frac{1}{2}\Phi_{37} + \frac{1}{2}\Phi_{38} = \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{8} = \frac{3}{16}$. The same procedure is used to compute the kinship coefficient between a great-grandparent and a great-grandchild, while the great-grandchild is inbred, as for individual 1 and individual 9. $\Phi_{19} = \Phi_{91} = \frac{1}{2}\Phi_{17} + \frac{1}{2}\Phi_{18} = \frac{1}{2} \cdot \frac{1}{8} + \frac{1}{2} \cdot \frac{1}{8} = \frac{1}{8}$.

All this elements are used to construct the kinship matrix $\mathbf{\Phi}$ in Equation (2.5). A table of some common values of kinship coefficients Φ are displayed in Table 2.1.

$$\mathbf{\Phi} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & \frac{1}{4} & \frac{1}{8} & \frac{3}{16} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & 0 & \frac{1}{8} & \frac{1}{4} & \frac{3}{16} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{16} & \frac{9}{32} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{16} & \frac{1}{2} & \frac{9}{32} \\ \frac{1}{8} & \frac{1}{8} & \frac{3}{16} & \frac{3}{16} & \frac{1}{8} & \frac{1}{8} & \frac{9}{32} & \frac{9}{32} & \frac{17}{32} \end{pmatrix} \tag{2.5}$$

Table 2.1: Kinship coefficients Φ for some relationships between individuals.

Relationship	Φ
Parent-offspring	$\frac{1}{4}$
Full siblings	$\frac{1}{4}$
Grandparent-grandchild	$\frac{1}{8}$
Half siblings	$\frac{1}{8}$
First cousins	$\frac{1}{16}$

2.3 Estimation of the kinship coefficient

Traditional relatedness coefficients, like the kinship coefficient described in Section 2.2, are based on the probabilities of IBD from common ancestors, and depict the *expected* proportion of shared alleles between the individuals in a pedigree. The actual genome sharing coefficient can deviate substantially from the expected value. Speed and Balding (2015) present a statement saying that human half-siblings are expected to share half of each chromosome that they received from their common parent, but the 95% credibility interval for their actual amount shared genome ranges from 37% to 63%. Nowadays, it is possible to compare the genomes of different individuals to get a measure of the genome sharing coefficient, directly from genome-wide SNP data. There are many different such measures available, and the focus of this section will be on one of the estimators.

2.3.1 SNP-based measures of relatedness

Association studies are designed to identify genetic loci at which the allele is correlated with a phenotype of interest. The associations of interest are causal, arising at loci whose different alleles have different effects on phenotype. Association studies are susceptible to genetic confounding, in the forms of cryptic relatedness and population structure, as presented in Chapter 1. These concepts cause correlation between the individuals in the study.

The advent of GWA data means that the actual genome-sharing coefficient can now be estimated accurately. It can be preferable to use these estimates in association analyses even if pedigree-based estimates are available. There is a slight difference between expectations computed from even a full pedigree, and realized amounts of shared genomic material (Astle and Balding, 2009). For example, if two individuals have a common ancestor many generations in the past, then this ancestor will contribute (slightly) to the pedigree-based kinship coefficient, but may or may not have passed any genetic material to both of them (Astle and Balding, 2009).

The goal now is to estimate the kinship matrix like in Equation (2.5), which was based on pedigree-data, but now based on SNP-data. If for example the minor allele is A and the major allele is a, then the possible genotypes are aa, Aa and AA. In the following, genotype data will be used, assuming that the genotype aa corresponds to the case where the SNP has two major alleles. Conversely, the genotype AA corresponds to the case where the SNP has two minor alleles, while the genotype Aa is the case where the SNP contains one major allele and one minor allele. The MAF is the frequency of the minor allele A in the population.

The genotype of individual i at the kth SNP can be coded as

$$G_{ik} = \begin{cases} 0 & \text{if the genotype of individual } i \text{ at SNP } k \text{ is } aa \\ 1 & \text{if the genotype of individual } i \text{ at SNP } k \text{ is } Aa \\ 2 & \text{if the genotype of individual } i \text{ at SNP } k \text{ is } AA. \end{cases} \tag{2.6}$$

Speed and Balding (2015) present three different estimators for the kinship coefficient Φ_{ij}, for a pair of individuals i and j, using the coding system in Equation (2.6), averaged over m SNPs. The motivation behind the estimators is to indicate whether alleles drawn from each of i and j are some given allelic type (Astle and Balding, 2009), in this case A. In this project we will only focus on one of the estimators, $\hat{\Phi}_{ij}$. In the project we did last semester, the different estimators were analyzed, and their characteristics were examined while varying the number of SNPs included, m, and the minor allele frequency at each SNP. This was done by a simulation study, which showed that $\hat{\Phi}_{ij}$ is the best estimator for the kinship coefficient.

The estimated kinship coefficient $\hat{\Phi}_{ij}$

According to Speed and Balding (2015), the lower the MAF the greater evidence for a recent common ancestor. This proposes giving more weight to the minor shared alleles, which can be done by centering the genotypes around the mean. The population MAF at the kth SNP, p_k, is assumed to be known, and the genotype frequencies are assumed to follow the Hardy-Weinberg principle from Section 2.1.1. The expected value of G_{ik} is calculated as

$$\mathrm{E}[G_{ik}] = \mu_k = 0 \cdot (1 - p_k^2) + 1 \cdot 2p_k(1 - p_k) + 2 \cdot p_k^2 = 2p_k - 2p_k^2 + 2p_k^2 = 2p_k.$$

In order to make each SNP equally informative, Speed and Balding (2015) standardize the genotypes in addition to centering. Thus, the variance of G_{ik} is needed, which can be calculated as

$$\begin{aligned} \mathrm{Var}[G_{ik}] = \sigma_k^2 &= \mathrm{E}[G_{ik}^2] - \mu_k^2 \\ &= 0 \cdot (1 - p_k^2) + 1 \cdot 2p_k(1 - p_k) + 2^2 \cdot p_k^2 - (2p_k)^2 \\ &= 2p_k(1 - p_k) + 4p_k^2 - 4p_k^2 \\ &= 2p_k(1 - p_k). \end{aligned}$$

The estimated kinship coefficient $\hat{\Phi}_{ij}$ is defined by Speed and Balding (2015) and Thornton and McPeek (2010) as

$$\begin{aligned}\hat{\Phi}_{ij} &= \frac{1}{2m}\sum_{k=1}^{m}\left(\frac{G_{ik}-\mu_k}{\sigma_k}\right)\left(\frac{G_{jk}-\mu_k}{\sigma_k}\right)\\ &= \frac{1}{2m}\sum_{k=1}^{m}\left(\frac{G_{ik}-2p_k}{\sqrt{2p_k(1-p_k)}}\right)\left(\frac{G_{jk}-2p_k}{\sqrt{2p_k(1-p_k)}}\right)\\ &= \frac{1}{2m}\boldsymbol{X}_i\boldsymbol{X}_j^{\mathsf{T}} \qquad (2.7)\end{aligned}$$

where $\boldsymbol{X}_i$ is a vector with $\frac{G_{ik}-2p_k}{\sqrt{2p_k(1-p_k)}}$ as the kth element, and $\boldsymbol{X}_j$ is a vector with $\frac{G_{jk}-2p_k}{\sqrt{2p_k(1-p_k)}}$ as the kth element. In the case with SNP data from several individuals, it is also of interest to arrange the estimated kinship coefficients between every pair of individuals in a systematic matrix, similarly as for the pedigree-based kinship coefficients in Section 2.2.2. The estimated kinship matrix can be computed as

$$\hat{\boldsymbol{\Phi}} = \frac{1}{2m}\mathbf{X}\mathbf{X}^{\mathsf{T}} \qquad (2.8)$$

where $\mathbf{X}$ is defined as a matrix with row i equal to $\boldsymbol{X}_i$ (Speed and Balding, 2015).

As stated by Astle and Balding (2009), in practice we do not know the minor allele frequencies p_k, but it can be estimated from the same data as the kinship coefficient, as

$$\hat{p}_k = \frac{\boldsymbol{1}^{\mathsf{T}}\boldsymbol{G}_k}{2n} \qquad (2.9)$$

where $\boldsymbol{G}_k$ is a column vector over the n individuals. To reduce the overfitting effect which can arise when estimating p_k from the same data, Astle and Balding (2009) suggests to iteratively re-estimate p_k after making an initial estimate of $\hat{\boldsymbol{\Phi}}$, with

$$\hat{p}_k^* = \frac{\boldsymbol{1}^{\mathsf{T}}\hat{\boldsymbol{\Phi}}^{-1}\boldsymbol{G}_k}{\boldsymbol{1}^{\mathsf{T}}\hat{\boldsymbol{\Phi}}^{-1}\boldsymbol{1}}$$

where $\boldsymbol{G}_k$ is a column vector over the m individuals.

The estimated kinship coefficient $\hat{\Phi}_{ij}$ in Equation (2.7) can according to Astle and Balding (2009) be interpreted in terms of excess allele sharing beyond what is expected for unrelated individuals, given the MAF. In this context, unrelatedness is a notion used for two alleles that are not IBD, but regarded as random draws from some allele pool.

When estimating the kinship coefficient $\hat{\Phi}_{ij}$, some combinations of genotypes can give negative coefficients. As presented by Astle and Balding (2009) this has caused some authors to avoid such estimators of the kinship coefficient, because

they can't be interpreted as probabilities. Under the interpretation of $\hat{\Phi}_{ij}$ as excess allele sharing, negative values correspond to individuals who share fewer alleles than expected, given the MAF (Astle and Balding, 2009).

Moreover, low MAF gives greater indication of a common ancestor, so individuals who both have two copies of the minor allele at a SNP, have a large estimated kinship coefficient. According to Panagiotou et al. (2010), GWA studies are typically designed to exclude SNPs with MAF $< 5\%$. Associations with minor variants may even have substantial genetic effects, but it requires strong statistical power to make meaningful statements about very minor alleles.

When it comes to the characteristics of the estimator $\hat{\Phi}_{ij}$, one can show that it is an unbiased estimator for the kinship coefficient. Speed and Balding (2015) claims that it is an unbiased and efficient estimator for the kinship coefficient, under the assumptions that alleles drawn from the global population are independent and that p_k are known. According to Thornton and McPeek (2010), the estimator $\hat{\Phi}_{ij}$ provides a consistent estimator for Φ_{ij}, if one assumes that the genotypes at different SNPs are independent with $m \to \infty$ and that p_k is known.

Chapter 3

Statistical models and methods

The main goal of this Master's thesis is to study the statistical models and methods used to perform a genotype–phenotype association analysis for cohort data. A selection of models and methods to be used are presented in this chapter.

Population structure and cryptic relatedness cause correlations between genome-wide SNPs, as presented in Chapter 1. GWA studies look for correlations between the phenotype and a SNP linked to the causal variant. However, relatedness causes correlations between genome-wide loci including causal and non-causal variants, and may therefore lead to spurious associations between the phenotype and unlinked SNPs (Lippert, 2013). Lippert states that as a result of confounding by population structure, some associations reported in the literature could be explained by differences in allele frequencies between populations, and thus do not replicate. Improvements in study design and exclusion of related individuals help to somewhat alleviate the problem of population structure, but the problem of cryptic relationships remains.

If the data is assumed to be a sample of independent individuals, a multiple linear regression model can be fitted. Thus, close relatives must be detected based on genotype data and removed from the study, which is assumed to reduce the statistical power. A complementary way to account for confounding structure is by ways of statistical modelling, an approach that is becoming more important as larger data sets are used to increase power. A range of methods have been proposed to correct for confounding in GWA studies, including genomic control, linear mixed models and principal components analysis (Dadd et al., 2009).

In the following sections, the different methods to analyze GWA data will be presented, and in that manner it is necessary to present the many R-functions available for GWA studies. The functions we will apply to the GWA data are part of `GenABEL`, which is an R package for GWA analysis (R Development Core Team, 2008). `GenABEL` will be presented in detail in Section 4.1.

3.1 Linear models

We will first consider the linear model,

$$\boldsymbol{Y} = \mathbf{X}_e\boldsymbol{\beta}_e + \boldsymbol{X}_{g_{(k)}}\beta_{g_{(k)}} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y}$ is the $n \times 1$ random vector of continuous responses, $\mathbf{X}_e$ is the $n \times p$ fixed effect design matrix, which represents p covariates corresponding to the fixed effects, and $\boldsymbol{\beta}_e$ is the $p \times 1$ fixed effects vector, consisting of the unknown regression coefficients associated with the covariates from the design matrix $\mathbf{X}_e$. Often the first column of $\mathbf{X}_e$ is a column of ones, to allow for an intercept in the model. Further, $\beta_{g_{(k)}}$ is the regression coefficient associated with the n-dimensional vector of genotypes, $\boldsymbol{X}_{g_{(k)}}$ for the k-th SNP. The error term $\boldsymbol{\epsilon}$ is a random vector, which is normally distributed with $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

For simplicity, we will use the following notation in the sequel:

$$\boldsymbol{Y} = \mathbf{X}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon} \tag{3.1}$$

where $\mathbf{X}_k$ is defined as the $n \times (p+1)$ matrix

$$\mathbf{X}_k = \begin{bmatrix}\mathbf{X}_e & \boldsymbol{X}_{g_{(k)}}\end{bmatrix} \tag{3.2}$$

and $\boldsymbol{\beta}_k$ is the $(p+1)$-dimensional vector

$$\boldsymbol{\beta}_k = \begin{bmatrix}\boldsymbol{\beta}_e \\ \beta_{g_{(k)}}\end{bmatrix}. \tag{3.3}$$

The underlying assumptions of the linear model are that (Bingham and Fry, 2010):

- the mean $\mathrm{E}(\boldsymbol{Y})$ is a linear function of the regressors;
- the errors are additive;
- the errors are independent;
- homoscedasticity of the errors;
- the errors are normally distributed.

Any or all of these assumptions may be inadequate, and it is needed to check the adequacy of the assumptions.

From this model we assume that the random vector $\boldsymbol{Y}$ is multivariate normally distributed with mean $\mathbf{X}_k\boldsymbol{\beta}_k$ and covariance matrix $\sigma^2\mathbf{I}$, in symbols:

$$\boldsymbol{Y} \sim \mathcal{N}(\mathbf{X}_k\boldsymbol{\beta}_k, \sigma^2\mathbf{I}).$$

Given an observation $\boldsymbol{Y} = \boldsymbol{y}$, the probability density function at $\boldsymbol{y}$ is

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}}\exp\Big\{-\frac{1}{2\sigma^2}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big)^{\mathsf{T}}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big)\Big\} \tag{3.4}$$

3.1.1 The multivariate normal distribution

The multivariate normal distribution of an n-dimensional random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ can be written as

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with density function

$$f(y_1, \ldots, y_n) = f(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \Big). \tag{3.5}$$

Assume $\boldsymbol{Y}$ is partitioned into $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$:

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix}.$$

The expectation $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are then also partitioned and defined as

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

The conditional distribution of $\boldsymbol{Y}_1$ given $\boldsymbol{Y}_2 = \boldsymbol{y}_2$ is then a multivariate normal distribution with

$$\mathrm{E}(\boldsymbol{Y}_1 | \boldsymbol{Y}_2 = \boldsymbol{y}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2) \tag{3.6}$$

$$\mathrm{Cov}(\boldsymbol{Y}_1 | \boldsymbol{Y}_2 = \boldsymbol{y}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{3.7}$$

3.1.2 Maximum likelihood estimation

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. The parameters needed to be estimated are $\boldsymbol{\beta}_k$ and σ^2. The likelihood function corresponding to Equation (3.4) is

$$L(\boldsymbol{\beta}_k, \sigma^2) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\Big\{ -\frac{1}{2\sigma^2}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big)^{\mathsf{T}}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big) \Big\}$$

and the log-likelihood function is

$$l(\boldsymbol{\beta}_k, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2\sigma^2}\log(\sigma^2) - \frac{1}{2\sigma^2}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big)^{\mathsf{T}}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big). \tag{3.8}$$

Taking the derivative of the log-likelihood function with respect to $\boldsymbol{\beta}_k$ and setting equal to 0,

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}_k} &= \frac{\partial}{\partial \boldsymbol{\beta}_k}\Big(-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - \boldsymbol{y}^\mathsf{T}\mathbf{X}_k\boldsymbol{\beta}_k - (\mathbf{X}_k\boldsymbol{\beta}_k)^\mathsf{T}\boldsymbol{y} \\ &\quad + (\mathbf{X}_k\boldsymbol{\beta}_k)^\mathsf{T}(\mathbf{X}_k\boldsymbol{\beta}_k)\Big) \\ &= 0,\end{aligned}$$

gives the estimator for $\boldsymbol{\beta}_k$:

$$\begin{aligned}\Rightarrow -\mathbf{X}_k^\mathsf{T}\boldsymbol{y} + \mathbf{X}_k^\mathsf{T}\mathbf{X}_k\boldsymbol{\beta}_k &= 0 \\ \hat{\boldsymbol{\beta}}_k &= (\mathbf{X}_k^\mathsf{T}\mathbf{X}_k)^{-1}\mathbf{X}_k^\mathsf{T}\boldsymbol{y}.\end{aligned}$$

The estimates for the environmental coefficients $\boldsymbol{\beta}_e$ can be read off as the p first elements of $\hat{\boldsymbol{\beta}}_k$, and the estimate for the genetic coefficient β_{g_k} is found in element $p+1$ of $\hat{\boldsymbol{\beta}}_k$.

It can be shown that the estimator $\hat{\boldsymbol{\beta}}_k$ is unbiased:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_k) = \boldsymbol{\beta}_k.$$

We observe that if the observations are independent and have constant variance σ^2, then the covariance matrix of the maximum likelihood estimator is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_k) = (\mathbf{X}_k^\mathsf{T}\mathbf{X}_k)^{-1}\sigma^2.$$

Finally it can be shown that the sampling distribution of the estimator $\hat{\boldsymbol{\beta}}_k$ in large samples is multivariate normal with mean and covariance matrix given above.

As a means of finding the estimator for σ^2, we compute the derivative of the log-likelihood function in Equation (3.8) with respect to σ^2 and set the equation equal to 0, which yields the following estimator:

$$\begin{aligned}\frac{\partial l}{\partial \sigma^2} &= 0 \\ 0 &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4}(\boldsymbol{y} - \mathbf{X}_k\hat{\boldsymbol{\beta}}_k)^\mathsf{T}(\boldsymbol{y} - \mathbf{X}_k\hat{\boldsymbol{\beta}}_k) \\ \hat{\sigma}^2 &= \frac{1}{n}\big(\boldsymbol{y} - \mathbf{X}_k\hat{\boldsymbol{\beta}}_k\big)^\mathsf{T}\big(\boldsymbol{y} - \mathbf{X}_k\hat{\boldsymbol{\beta}}_k\big).\end{aligned}$$

Defining the sum of squared differences between observed and predicted values, or residual sum of squares (RSS), as

$$\mathrm{RSS}(\hat{\boldsymbol{\beta}}_k) = \big(\boldsymbol{y} - \mathbf{X}_k\hat{\boldsymbol{\beta}}_k\big)^\mathsf{T}\big(\boldsymbol{y} - \mathbf{X}_k\hat{\boldsymbol{\beta}}_k\big),$$

the expression for $\hat{\sigma}^2$ is abbreviated to

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}(\hat{\boldsymbol{\beta}}_k)}{n}. \tag{3.9}$$

It can be shown that this estimator is biased, and the method of restricted maximum likelihood estimation (REML) gives the unbiased estimator for σ^2:

$$s^2 = \frac{\mathrm{RSS}(\hat{\boldsymbol{\beta}}_k)}{n-p-1}. \tag{3.10}$$

3.1.3 Linear models in `GenABEL` - `mlreg`

The `GenABEL` package includes a function for performing linear regression for genome-wide SNP data; the `mlreg` function. This function uses the standard linear regression approach in Equation (3.1), which results in equivalent estimates from the `lm` function in R and the `mlreg` function in `GenABEL`. The p-values may differ, and the reason for this discrepancy is that while `lm` applies the t-test to test significance, `mlreg` uses the Wald test. The differences between the results from the hypotheses tests are usually small when the number of individuals is large.

3.2 Hypothesis testing

This section describes statistical tests for hypotheses regarding the unknown regression parameters $\boldsymbol{\beta}_k$, as defined in Equation (3.3). A hypothesis test is a method of statistical inference. The goal of a hypothesis test is to decide, based on a sample from the population, which of two complementary hypotheses is true.

Definition: *The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis, H_0 and the alternative hypothesis, H_1 (Casella and Berger, 2002).*

The comparison of the two hypotheses is regarded as statistically significant if the relationship between the data sets is an unlikely realisation of the null hypothesis, given a significance level. Typically, a hypothesis test is specified in terms of a test statistic $W(\boldsymbol{Y})$, a function of the sample (Casella and Berger, 2002).

The most common statistical hypotheses in linear models are (Fahrmeir et al., 2013):

1. Test of significance:

$$H_0 : \beta_{k_j} = 0 \quad \text{against} \quad H_1 : \beta_{k_j} \neq 0.$$

2. Composite test of a subvector $\boldsymbol{\beta}_{k_1} = \left(\beta_{k_1} \dots \beta_{k_r}\right)^\mathsf{T}$:

$$H_0 : \boldsymbol{\beta}_{k_1} = \mathbf{0} \quad \text{against} \quad H_1 : \boldsymbol{\beta}_{k_1} \neq \mathbf{0}.$$

3. Test of equality:

$$H_0 : \beta_{k_j} - \beta_{k_r} = 0 \quad \text{against} \quad H_1 : \beta_{k_j} - \beta_{k_r} \neq 0.$$

Each of these test problems can be treated as special cases of *general linear hypothesis tests*, which are defined as

$$H_0 : \mathbf{C}\boldsymbol{\beta}_k = \boldsymbol{d} \quad \text{against} \quad H_1 : \mathbf{C}\boldsymbol{\beta}_k \neq \boldsymbol{d},$$

where $\mathbf{C}$ is an $r \times (p+1)$ matrix with $\text{rank}(\mathbf{C}) = r \leq (p+1)$. Accordingly, under H_0 there are a total of r linear-independent conditions.

In order to test the significance of one regression parameter (case 1), $\boldsymbol{d} = 0$ and $\mathbf{C}$ is a $1 \times (p+1)$ matrix given by $\mathbf{C} = (0, \dots, 0, 1, 0, \dots, 0)$, in which the one is in the jth position. When testing the first r components of $\boldsymbol{\beta}_k$ (case 2), one obtain the r-dimensional vector $\boldsymbol{d} = \mathbf{0}$ and the $r \times (p+1)$ matrix

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & & & & \ddots & & & \vdots & \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{pmatrix}.$$

The last case, when testing the equality of two regression coefficients, one obtains the scalar $\boldsymbol{d} = 0$ and the $1 \times (p+1)$ matrix given by $\mathbf{C} = (0, \dots, 1, \dots, -1, \dots, 0)$, where the one is in the jth column and the minus one is in the rth column.

It can be shown that the test statistic for the general linear hypothesis is (Fahrmeir et al., 2013):

$$F_{\text{obs}} = \frac{1}{r}(\mathbf{C}\hat{\boldsymbol{\beta}}_k - \boldsymbol{d})^\mathsf{T}(\hat{\sigma}^2\mathbf{C}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{C}^\mathsf{T})^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}_k - \boldsymbol{d}) \sim \mathcal{F}_{r,n-p-1},$$

and we reject the null hypothesis H_0 if the test statistic is larger than the $(1-\alpha)$-quantile of the $\mathcal{F}$-distribution with r and $n-p-1$ degrees of freedom, i.e.,

$$F_{\text{obs}} > \mathcal{F}_{r,n-p-1}(1-\alpha).$$

The test statistic of hypothesis test 1 (t-test) is then

$$t_j = \frac{\hat{\beta}_{k_j}}{\text{se}_j} \sim t_{n-p-1},$$

where se_j denotes the unbiased estimated standard error of $\hat{\beta}_{k_j}$, as given in Equation (3.10). Equivalently, squaring the test-statistic gives

$$t_j^2 = \frac{\hat{\beta}_{k_j}^2}{\text{se}_j^2} \sim \mathcal{F}_{1,n-p-1}.$$

It can be shown that the relationship between the F-test and the Wald test is

$$W = rF \overset{a}{\sim} \mathcal{X}_r^2,$$

where F is a F-test statistic with r and $n-p-1$ degrees of freedom. Thus the $\mathcal{F}_{r,n-p-1}$-distribution converges in distribution with $n \to \infty$ to a $\mathcal{X}_r^2$-distribution.

Score test

We present the score test in a more general context. The score test is a convenient tool for testing whether one or more parameters equal some null value, whenever the probability distribution can be assumed for the data. Suppose we have a sample of independent random variables $\boldsymbol{Y}$ which is described by the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, and the hypothesis to test is $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. Often it is of interest to test whether a subset of $\boldsymbol{\beta}$ equals some null value, so $\boldsymbol{\beta}$ are partitioned into $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, where $\boldsymbol{\beta}_2$ are the parameters to be tested, and $\boldsymbol{\beta}_1$ are regarded as unknown nuisance parameters. The nuisance parameters are replaced by their maximum likelihood estimates $\hat{\boldsymbol{\beta}}_1$. The null hypothesis can be written as $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{2,0}$, and we define $\boldsymbol{\beta}_0 = (\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_{2,0})$.

The score vector for the parameter vector $\boldsymbol{\beta}_0$ is defined as the p-dimensional vector (Bjørnland, 2014)

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

According to Bjørnland (2014) it can be shown that by assuming necessary conditions, $\boldsymbol{S}(\boldsymbol{\beta}_0, \boldsymbol{Y})$ has an approximate $\mathcal{N}(\boldsymbol{0}, I(\boldsymbol{\beta}_0))$-distribution, where $I(\boldsymbol{\beta}_0)$ is the Fisher information matrix corresponding to $\boldsymbol{S}(\boldsymbol{\beta}_0, \boldsymbol{Y})$ (Casella and Berger, 2002).

The score vector is separated according to derivatives of nuisance parameters and parameters to be tested under H_0:

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

where $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are defined as

$$\boldsymbol{S}_1 = \frac{\partial l}{\partial \boldsymbol{\beta}_1}, \; \boldsymbol{S}_2 = \frac{\partial l}{\partial \boldsymbol{\beta}_2}.$$

The elements of $\boldsymbol{S}(\boldsymbol{\beta}_0, \boldsymbol{Y})$ are distributed as

$$\boldsymbol{S}(\boldsymbol{\beta}_0, \boldsymbol{Y}) = \begin{bmatrix} \boldsymbol{S}_1(\boldsymbol{\beta}_0, \boldsymbol{Y}) \\ \boldsymbol{S}_2(\boldsymbol{\beta}_0, \boldsymbol{Y}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 = \boldsymbol{0} \\ \boldsymbol{\mu}_2 = \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} I(\boldsymbol{\beta}_0)_{11} & I(\boldsymbol{\beta}_0)_{12} \\ I(\boldsymbol{\beta}_0)_{21} & I(\boldsymbol{\beta}_0)_{22} \end{bmatrix} \right).$$

By definition, $\boldsymbol{S}_1(\boldsymbol{\beta}_0, \boldsymbol{Y}) = \frac{\partial l}{\partial \boldsymbol{\beta}_1}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \boldsymbol{0}$. The distribution of $\boldsymbol{S}_{2|1} = \boldsymbol{S}_2(\boldsymbol{\beta}_0, \boldsymbol{Y})|$ $(\boldsymbol{S}_1(\boldsymbol{\beta}_0, \boldsymbol{Y}) = \boldsymbol{0})$ is given by the expressions for conditional normal distributed variables in Equations (3.6) and (3.7),

$$\boldsymbol{\mu}_{2|1} = \mathrm{E}(\boldsymbol{S}_{2|1}) = \boldsymbol{\mu}_2 + I(\boldsymbol{\beta}_0)_{21} I(\boldsymbol{\beta}_0)_{11}^{-1} (\boldsymbol{S}_1 - \boldsymbol{\mu}_1) = \boldsymbol{0}$$
$$\Sigma_{2|1} = \mathrm{Var}(\boldsymbol{S}_{2|1}) = I(\boldsymbol{\beta}_0)_{12} - I(\boldsymbol{\beta}_0)_{21} I(\boldsymbol{\beta}_0)_{11}^{-1} I(\boldsymbol{\beta}_0)_{12}.$$

Thus, the score test statistic answering to $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{2,0}$ is defined as

$$T_{2|1} = \boldsymbol{S}_{2|1}^{\mathsf{T}} \Sigma_{2|1}^{-1} \boldsymbol{S}_{2|1},$$

which can be shown to be asymptotically χ^2-distributed with degrees of freedom equal to the difference in the number of parameters between H_0 and H_1 (Lehmann and Romano, 2005). In the case of testing only one parameter β_k, the degrees of freedom are 1.

3.3 Genomic control

Assume that we perform a hypothesis test with the aid of a test statistic that is exactly, or asymptotically, χ_1^2-distributed, and that population stratification and cryptic relatedness have not been taken into account. This test might be the Wald or score test for one regression coefficient β_k. The basic idea of genomic control methods is to detect and correct for stratification based on the genome-wide inflation of the statistics (Price et al., 2010). The empirical distribution of the statistics is inflated from χ_1^2 to $\lambda\chi_1^2$ (Dadd et al., 2009) because of several potential confounders. The possible confounders are population stratification, cryptic relatedness and *differential bias*. Differential bias refers to the spurious differences in allele frequencies between samples due to differences in sample collection, sample preparation and/or genotyping assay procedures.

In order to determine the inflation λ in the test statistics, a set of unlinked SNPs are used, and it is assumed that the test statistics are observed under the complete H_0: all null hypotheses are true. The inflation factor is then applied to the test statistics to correct them as appropriate. The inflation factor λ does not depend on the allele frequency, hence, λ is a constant, under the condition that the amount of genetic variance that can be explained by population structure is constant over the genome (F_{ST} constant) (Aulchenko, 2014). Therefore λ can be

estimated from genomic data, using a set of random SNPs which are believed not to be associated with the trait.

We assume that the inflation factor λ is constant for all SNPs across the genome. Given the test statistics X_k^2 from a set of L unlinked SNPs not associated with the trait, and spread across the genome, the estimator for λ is (Aulchenko, 2014; Dadd et al., 2009)

$$\hat{\lambda} = \frac{\text{median}(X_1^2, X_2^2, \ldots, X_L^2)}{0.4549}. \tag{3.11}$$

The factor 0.4549 is the median of a χ_1^2-distribution, $\mathrm{P}(\chi_1^2 < 0.5)$. For the tested markers, the corrected value of the test statistic is obtained by simple division of the original test statistic value by $\hat{\lambda}$, $X^2_{\text{corrected}} = X^2_{\text{original}}/\hat{\lambda}$.

The estimated value of λ, $\hat{\lambda}$, is also used as a quality control for detection of stratification. According to Price et al. (2010), a value of $\hat{\lambda} \approx 1$ indicates no stratification, whereas $\hat{\lambda} > 1$ indicates stratification or other confounders presented above. However, values of $\hat{\lambda} < 1.05$ are generally acceptable.

The expected proportion of markers that are associated with the trait is generally small. Therefore, practically all loci are used to estimate λ. If very strong associations are present, true associations will increase the average value of the test, and genomic control correction will lead to too few significant findings. Aulchenko (2014) suggests to use 95% of the least significant associations to estimate the inflation factor. However, there are several problems with using the genomic control method to account for population stratification. The method assumes uniform F_{ST} across the genome (Aulchenko, 2014), which may not be the case for some genomic regions, and the method is highly variable and depend strongly on the number of SNPs genotyped (Dadd et al., 2009). Both Aulchenko (2014) and Dadd et al. (2009) recommend use of other methods to correct for genetic confounding, which take the structure of the sample into account in a direct manner.

3.4 Linear mixed models

Among the methods to correct for confounding presented in the opening of this chapter, only the linear mixed models have the ability to correct for population structure and cryptic relatedness, while retaining sufficient power to detect true associations (Lippert, 2013). The key idea behind using linear mixed models to correct for confounding is that while it is hard to reliably give point estimates for the effects of confounding genetic structure, it is often possible to describe these in terms of random effects, for which covariation can be quantified in terms of the degree of genetic relatedness between the samples (Lippert, 2013). In linear mixed

models the phenotype vector $\boldsymbol{Y}$ is written as the mixed sum of linear terms in the fixed effects $\boldsymbol{\beta}_k$ and random effects $\boldsymbol{Q}$ and $\boldsymbol{\epsilon}$, i.e. a mixed model:

$$\boldsymbol{Y} = \mathbf{X}_k\boldsymbol{\beta}_k + \boldsymbol{Q} + \boldsymbol{\epsilon} \tag{3.12}$$

where $\boldsymbol{Y}$ is a vector of n responses. The regression coefficients $\boldsymbol{\beta}_k$, defined in Equation (3.3), are fixed effects, corresponding to the design matrix $\mathbf{X}_k$, defined in Equation (3.2). The vector $\boldsymbol{Q}$ is a vector of random effects, which is assumed to follow the distribution $\boldsymbol{Q} \sim \mathcal{N}(0, 2\sigma_g^2\boldsymbol{\Phi})$, where $\boldsymbol{\Phi}$ is the $n \times n$ matrix of pairwise kinship coefficients. The errors $\boldsymbol{\epsilon}$ are assumed to follow the distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_e^2\mathbf{I})$, where $\mathbf{I}$ is the $n \times n$ identity matrix. The parameters σ_e and σ_g are to be estimated, and are representing the environmental and genetic components of variance, where the genetic variance is due to so called polygenes (Eu-ahsunthornwattana et al., 2014). The matrix $\boldsymbol{\Phi}$ is in general unknown and is replaced by the estimated kinship matrix $\hat{\boldsymbol{\Phi}}$.

When testing a marker for association with the phenotype, the variables of interest are modelled as fixed, whereas the random effects account for nuisance variation and are therefore not of interest (Lippert, 2013).

Lippert (2013) uses the notation

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \bar{\mathbf{G}}\boldsymbol{u} + \boldsymbol{\epsilon}$$

where the random effect is $\boldsymbol{u}$ and the $n \times m$ matrix $\bar{\mathbf{G}}$ is the design matrix holding all the loci. Lippert (2013) continues to define the total random genetic effect as in Equation (3.12), $\boldsymbol{Q} = \bar{\mathbf{G}}\boldsymbol{u}$, and he says that $\boldsymbol{Q} \sim \mathcal{N}(0, 2\sigma_g^2\boldsymbol{\Phi})$, where the covariance matrix is proportional to $\boldsymbol{\Phi} = \frac{1}{2m}\bar{\mathbf{G}}\bar{\mathbf{G}}^\mathsf{T}$. We will not use the notation of Lippert, but continue using the notation established in Equation (3.12).

3.4.1 The implied marginal model

Based on the normality assumption of the random effects and the errors, we can derive the marginal distribution of the responses. The random effects and the errors are assumed to be independent of each other. The marginal linear model is defined as (Østgård, 2011)

$$\boldsymbol{Y} = \mathbf{X}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon}^*,$$

where

$$\boldsymbol{\epsilon}^* = \boldsymbol{Q} + \boldsymbol{\epsilon}.$$

Thus, $\boldsymbol{\epsilon}^*$ is normally distributed with expected value

$$\mathrm{E}(\boldsymbol{\epsilon}^*) = \mathrm{E}(\boldsymbol{Q}) + \mathrm{E}(\boldsymbol{\epsilon}) = 0,$$

and covariance matrix

$$\begin{aligned}\mathbf{V} = \mathrm{Cov}(\boldsymbol{\epsilon}^*) &= \mathrm{Cov}(\boldsymbol{Q}) + \mathrm{Cov}(\boldsymbol{\epsilon}) \\ &= 2\sigma_g^2\boldsymbol{\Phi} + \sigma_e^2\mathbf{I}. \end{aligned} \tag{3.13}$$

It follows that $\boldsymbol{\epsilon}^*$ is multivariate normally distributed,

$$\boldsymbol{\epsilon}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{V}).$$

Thus, the marginal distribution of $\boldsymbol{Y}$ is

$$\boldsymbol{Y} \sim \mathcal{N}(\mathbf{X}_k\boldsymbol{\beta}_k, \mathbf{V}).$$

3.4.2 Maximum likelihood estimation

As in the case of linear regression, the likelihood is maximized by equating the gradient with respect to all parameters to zero, and jointly solving the resulting equations. When defining the variance parameters $\boldsymbol{\theta} = \left[\sigma_e^2, \sigma_g^2\right]$ and the corresponding covariance term $\mathbf{V}_{\boldsymbol{\theta}}$, the marginal distribution of $\boldsymbol{Y}$ is the multivariate probability density function given in Equation (3.5)

$$f(\boldsymbol{y}|\boldsymbol{\beta}_k, \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|\mathbf{V}_{\boldsymbol{\theta}}|^{1/2}} \exp\Big(-\frac{1}{2}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big)^\mathsf{T}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big)\Big).$$

Given the observed data $\boldsymbol{Y} = \boldsymbol{y}$, the likelihood function is defined as

$$L(\boldsymbol{\beta}_k, \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|\mathbf{V}_{\boldsymbol{\theta}}|^{1/2}} \exp\Big(-\frac{1}{2}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big)^\mathsf{T}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big)\Big),$$

and the log-likelihood function is

$$l(\boldsymbol{\beta}_k, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_{\boldsymbol{\theta}}| - \frac{1}{2}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big)^\mathsf{T}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big). \tag{3.14}$$

Taking the derivative of the log-likelihood function with respect to $\boldsymbol{\beta}_k$ and setting the equation equal to 0, yields the corresponding estimator, which is the same procedure as for linear regression,

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}_k} &= \frac{\partial}{\partial \boldsymbol{\beta}_k}\Big(-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_{\boldsymbol{\theta}}| - \frac{1}{2}(\boldsymbol{y}^\mathsf{T}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y} - \boldsymbol{y}^\mathsf{T}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X}_k\boldsymbol{\beta}_k \\ &\quad - (\mathbf{X}_k\boldsymbol{\beta}_k)^\mathsf{T}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y} + (\mathbf{X}_k\boldsymbol{\beta}_k)^\mathsf{T}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X}_k\boldsymbol{\beta}_k\Big) \\ &= 0\end{aligned}$$

$$\Rightarrow -\mathbf{X}_k^\mathsf{T}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y} + \mathbf{X}_k^\mathsf{T}\mathbf{V}\mathbf{X}_k\boldsymbol{\beta}_k = 0$$
$$\hat{\boldsymbol{\beta}}_k = (\mathbf{X}_k^\mathsf{T}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X}_k)^{-1}\mathbf{X}_k^\mathsf{T}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y} \tag{3.15}$$

While for linear regression the maximum likelihood parameters can be found in closed form from the gradient equations, this is not the case for linear mixed models. The log marginal likelihood function is not jointly convex in the variance parameters, rendering it hard to ensure global maximization of the likelihood (Lippert, 2013). Lippert (2013) proposes a simplified version of maximum likelihood estimation, where the log-likelihood in Equation (3.14) is written as a function of the ratio $\gamma = \frac{\sigma_g^2}{\sigma_e^2}$:

$$l(\gamma, \sigma_e^2, \boldsymbol{\beta}_k) = -\frac{n}{2}\log(2\pi\sigma_e^2) - \frac{1}{2}\log|\mathbf{H}_\gamma| - \frac{1}{2\sigma_e^2}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big)^\mathsf{T}\mathbf{H}_\gamma^{-1}\big(\boldsymbol{y} - \mathbf{X}_k\boldsymbol{\beta}_k\big) \tag{3.16}$$

where the matrix $\mathbf{H}_\gamma$ is defined as $\mathbf{H}_\gamma = \mathbf{I} + \gamma\boldsymbol{\Phi}$. The estimator of the fixed effect $\boldsymbol{\beta}_k$ is computed in the similar way as for the estimators in Equation (3.15)

$$\hat{\boldsymbol{\beta}}_{k_\gamma} = (\mathbf{X}_k^\mathsf{T}\mathbf{H}_\gamma^{-1}\mathbf{X}_k)^{-1}\mathbf{X}_k^\mathsf{T}\mathbf{H}_\gamma^{-1}\boldsymbol{y}. \tag{3.17}$$

In order to find the estimator of the environmental variance σ_e^2 as a function of γ, the estimator $\hat{\boldsymbol{\beta}}_{k_\gamma}$ from Equation (3.17) is substituted into the log-likelihood in Equation (3.16). Taking the derivative with respect to σ_e^2 and setting this equal to 0, yields

$$\begin{aligned}\frac{\partial l}{\partial \sigma_{e_\gamma}^2} &= -\frac{n}{2\sigma_{e_\gamma}^2} + \frac{1}{2\sigma_{e_\gamma}^4}\big(\boldsymbol{y} - \mathbf{X}_{k_\gamma}\hat{\boldsymbol{\beta}}_{k_\gamma}\big)^\mathsf{T}\mathbf{H}_\gamma^{-1}\big(\boldsymbol{y} - \mathbf{X}_{k_\gamma}\hat{\boldsymbol{\beta}}_{k_\gamma}\big)\\ &= 0,\end{aligned}$$

which gives that the estimator for the environmental variance given γ is

$$\hat{\sigma}_{e_\gamma}^2 = \frac{1}{n}\big(\boldsymbol{y} - \mathbf{X}_{k_\gamma}\hat{\boldsymbol{\beta}}_{k_\gamma}\big)^\mathsf{T}\mathbf{H}_\gamma^{-1}\big(\boldsymbol{y} - \mathbf{X}_k\hat{\boldsymbol{\beta}}_{k_\gamma}\big).$$

The expression can be further simplified as

$$\hat{\sigma}_{e_\gamma}^2 = \frac{1}{n}\boldsymbol{y}^\mathsf{T}\mathbf{P}_\gamma^\mathsf{T}\mathbf{H}_\gamma^{-1}\mathbf{P}_\gamma\boldsymbol{y} \tag{3.18}$$

where

$$\mathbf{P}_\gamma = \mathbf{I} - \mathbf{X}_e(\mathbf{X}_k^\mathsf{T}\mathbf{H}_\gamma^{-1}\mathbf{X}_k)^{-1}\mathbf{X}_k^\mathsf{T}\mathbf{H}_\gamma^{-1}\boldsymbol{y}.$$

When plugging the estimators of $\sigma_{e_\gamma}^2$ and $\boldsymbol{\beta}_{k_\gamma}$ back into the log-likelihood in Equation (3.16), we obtain the profile log-likelihood as

$$l(\gamma) = -\frac{n}{2}\log(2\pi\hat{\sigma}_{e_\gamma}^2) - \frac{1}{2}\log|\mathbf{H}_\gamma^{-1}| - \frac{1}{2\hat{\sigma}_{e_\gamma}^2}\big(\boldsymbol{y} - \mathbf{X}_k\hat{\boldsymbol{\beta}}_k\big)^\mathsf{T}\mathbf{H}_\gamma^{-1}\big(\boldsymbol{y} - \mathbf{X}_k\hat{\boldsymbol{\beta}}_k\big).$$

The profile log-likelihood can be simplified by using the expressions from Equations (3.17) and (3.18),

$$\begin{aligned} l(\gamma) &= -\frac{n}{2}\log(2\pi\frac{1}{n}\boldsymbol{y}^\mathsf{T}\mathbf{P}_\gamma^\mathsf{T}\mathbf{H}_\gamma^{-1}\mathbf{P}_\gamma\boldsymbol{y}) - \frac{1}{2}\log|\mathbf{H}_\gamma^{-1}| \\ &\quad - \frac{1}{\frac{2}{n}\boldsymbol{y}^\mathsf{T}\mathbf{P}_\gamma^\mathsf{T}\mathbf{H}_\gamma^{-1}\mathbf{P}_\gamma\boldsymbol{y}}(\boldsymbol{y}^\mathsf{T}\mathbf{P}_\gamma^\mathsf{T}\mathbf{H}_\gamma^{-1}\mathbf{P}_\gamma\boldsymbol{y}) \\ &= -\frac{n}{2}(1+\log\frac{2\pi}{n}) - \frac{1}{2}\log|\mathbf{H}_\gamma^{-1}| - \frac{n}{2}\log(\boldsymbol{y}^\mathsf{T}\mathbf{P}_\gamma^\mathsf{T}\mathbf{H}_\gamma^{-1}\mathbf{P}_\gamma\boldsymbol{y}). \end{aligned}$$

In principle, a local optimum with respect to γ of this profile log-likelihood could be obtained by the use of gradient descent methods. Alternatively, a grid search can be used to find an optimum for γ (Lippert, 2013).

According to Lippert (2013), the maximum likelihood estimate underestimate the variances, because it is biased as for the estimator in Equation (3.9). He states that restricted maximum likelihood estimation has been proposed to overcome this problem, but this will not be of focus here.

3.4.3 Best linear unbiased prediction

The best linear unbiased predictor (BLUP) is a minimum variance predicted value of the random effects $\boldsymbol{Q}$ in a linear mixed model (Lippert, 2013), similar to the best linear unbiased estimator (BLUE) of fixed effects. The BLUP $\hat{Q}_i$ of an individual i is obtained by maximizing the joint distribution of the vector of all observed phenotypes $\boldsymbol{Y}$ and the random genetic effect Q_i of that individual of interest. Let $\mathbf{V}$ be the total covariance term of $\boldsymbol{Y}$ as in Equation (3.13), $\boldsymbol{\Phi}_{:i}$ is the $1 \times n$ vector of genetic relatedness between individual i and all observed individuals, and the genetic relatedness of individual i with itself is Φ_{ii}. The joint distribution of $\boldsymbol{Y}$ and Q_i is given as

$$\begin{bmatrix} \boldsymbol{Y} \\ Q_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{X}_k\boldsymbol{\beta}_k \\ 0 \end{bmatrix}; \begin{bmatrix} \mathbf{V} & \sigma_g^2\boldsymbol{\Phi}_{:i}^\mathsf{T} \\ \sigma_g^2\boldsymbol{\Phi}_{:i} & \sigma_g^2\Phi_{ii} \end{bmatrix} \right).$$

In order to find the BLUP for Q_i, the results for conditional distributions of multivariate normal vectors in Equation (3.6) and (3.7) are used, and the BLUP $\hat{Q}_i$ is equal to the expected value of the conditional distribution of Q_i given $\boldsymbol{Y}$,

$$Q_i|\boldsymbol{Y} \sim \mathcal{N}\left(\sigma_g^2\boldsymbol{\Phi}_{:i}\mathbf{V}^{-1}(\boldsymbol{Y} - \mathbf{X}_k\boldsymbol{\beta}_k); \sigma_g^2\Phi_{ii} - \sigma_g^2\boldsymbol{\Phi}_{:i}\mathbf{V}^{-1}\sigma_g^2\boldsymbol{\Phi}_{:i}^\mathsf{T}\right)$$

so that

$$\hat{Q}_i = \hat{\sigma}_g^2\boldsymbol{\Phi}_{:i}\mathbf{V}^{-1}(\boldsymbol{Y} - \mathbf{X}_k\hat{\boldsymbol{\beta}}_k).$$

In practice, $\boldsymbol{\Phi}_{:i}$ is replaced by the estimated matrix $\hat{\boldsymbol{\Phi}}_{:i}$.

3.4.4 Linear mixed models in GenABEL

There are several R-packages available for fitting linear mixed effects models, like `nlme` and `lme4`. `lme4` is a newer version to `nlme`, which is more efficient and uses restricted maximum likelihood estimation. In GWA studies, in which hundreds of thousands of SNPs are to be analyzed, it is quite time consuming to use these packages. The gold standard of genetic association analysis is a likelihood ratio test-based method using variance component analysis applied to a mixed effects model. However, the method requires estimation of all the model parameters for every tested SNP k in Equation (3.12), and is thus computationally demanding (Svishcheva et al., 2012). Instead, the focus is on fast approximate tests which were developed for the purposes of GWA analysis in samples of relatives. The `GenABEL` package include functions for efficient GWA analysis, which will be described in the following sections.

polygenic

`polygenic` is a `GenABEL` function which estimates the linear mixed (polygenic) model based on trait and covariates data, and includes the estimated kinship matrix. The function maximizes the likelihood of the data, given in Equation (3.14), and reports twice the negative maximum likelihood estimates. The main use of this function is to estimate regression coefficients, environmental residuals and the inverse of the covariance matrix for further use in analysis with `mmscore` and `GRAMMAR`. It is difficult to compute the inverse of the covariance matrix given in Equation (3.13), so the eigenvectors of the inverse of $\hat{\mathbf{\Phi}}$ are used instead of taking the inverse. This method is partly based on the paper of E. A. Thompson (1990), and by taking the Moore-Penrose generalized inverse of $\hat{\mathbf{\Phi}}$.

mmscore

The `polygenic` and `mmscore` functions together allow implementation of the family based score test approximation (FASTA) method proposed by Chen and Abecasis (2007), based on the estimated kinship matrix $\hat{\mathbf{\Phi}}$. The FASTA approach divides the model parameters into two categories; segregation parameters related to trait heritability and parameters describing the effects of SNPs on this trait. The segregation parameters are estimated, and the covariance matrix for the phenotypes of the study participants is computed once for a given trait. This step corresponds to the `polygenic`-step. The next step involves evaluating the effect of every SNP, making corrections for the covariance matrix. This approach approximates the likelihood ratio test well if many loci of small effects are involved in trait determination. This approach is much less computationally complex than the likelihood

ratio test-based method, but it becomes rather slow when millions of SNPs are analyzed in large samples (Svishcheva et al., 2012).

GRAMMAR

In order to increase the computation speed, Aulchenko et al. (2007) proposed the GRAMMAR method using environmental residuals estimated from the `polygenic` function. The basic idea is to perform a single linear mixed model analysis using all the individuals, but ignoring genotype data. Subsequently, we use residuals from this analysis, which are now adjusted for polygenic covariation and fixed effects, as a novel quantitative trait for association analyses with each of many SNPs using classical methods for unrelated individuals (Aulchenko et al., 2007).

In the initial step, a reduced version of linear mixed model is fitted, similar to Equation (3.12)

$$\boldsymbol{Y} = \mathbf{X}_e\boldsymbol{\beta}_e + \boldsymbol{Q} + \boldsymbol{\epsilon},$$

including only the environmental regression covariates. The residuals from this analysis are given by

$$\boldsymbol{Y}^* = \boldsymbol{Y} - \left(\mathbf{X}_e\hat{\boldsymbol{\beta}}_e + \hat{\boldsymbol{Q}}\right) \tag{3.19}$$

where $\hat{\boldsymbol{\beta}}_e$ is the estimate of the fixed effect and $\hat{\boldsymbol{Q}}$ is the estimated contribution from the polygenes (Aulchenko et al., 2007). In the second step, these residuals are used as the predictor in a simple linear regression model for each SNP k,

$$\boldsymbol{Y}^* = \boldsymbol{X}_{g_{(k)}}\tilde{\beta}_{g_{(k)}} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}^*$ is a vector of residuals from Equation (3.19), $\boldsymbol{X}_{g_{(k)}}$ is the vector of genotypes at the SNP under study, $\tilde{\beta}_{g_{(k)}}$ is the effect of SNP k, and $\boldsymbol{\epsilon}$ is the vector of random errors. This analytical approach is called GRAMMAR (Aulchenko et al., 2007). Estimation of $\tilde{\beta}_{g_{(k)}}$ can be accomplished through maximum likelihood or least squares approaches (Eu-ahsunthornwattana et al., 2014), and testing of the null hypothesis that $\tilde{\beta}_{g_{(k)}} = 0$ is done by applying the score test. Subsequent to the GRAMMAR analysis, SNPs showing test statistics greater than some predefined threshold are selected for a final analysis using the linear mixed model

$$\boldsymbol{Y} = \mathbf{X}_e\boldsymbol{\beta}_e + \boldsymbol{X}_{g_{(k)}}\beta_{g_{(k)}} + \boldsymbol{Q} + \boldsymbol{\epsilon}.$$

This version of GRAMMAR is shown to produce a conservative test and biased estimates of the regression coefficients. In order to fix these problems, Amin et al. (2007) proposed a genomic control corrected version of GRAMMAR which is denoted GRAMMAR-GC. The method involves the same steps as the original GRAMMAR, but the final score test statistic is re-inflated by multiplying by an

appropriate estimated correction factor. This is analogous to the deflation of χ_1^2-statistic in the method of genomic control in Section 3.3, and result in a final test statistic with the appropriate null distribution (Eu-ahsunthornwattana et al., 2014). Thus, the conservativity of the test is solved, but the method generates biased estimates of the regression coefficients and the genomic control λ is by definition 1, and can not serve as an indicator of goodness of the model.

Svishcheva et al. (2012) present an extremely fast variance components-based two-step method, which solves all the problems of the original GRAMMAR and GRAMMAR-GC. According to Svishcheva et al. (2012), this new method called GRAMMAR-gamma produces a correct distribution of the test statistics, interpretable values of the genomic control λ, and unbiased estimates of the regression coefficients. We will not go into details of the method, but it involves calculating a correction factor which is used to adjust a test statistic. The new GRAMMAR-gamma statistic can be shown to be approximately equivalent to the FASTA statistic (Eu-ahsunthornwattana et al., 2014) of the `mmscore` function, but more efficiently calculated. Details for the GRAMMAR-gamma method can be seen in Svishcheva et al. (2012).

All the different methods are applied by using the `GenABEL`-function `GRAMMAR`, and choosing method *raw*, *gc* or *gamma*.

3.5 Comparison of principal components regression and linear mixed models

It has been standard practice to include principal components of the genotypes constructed from a genotype matrix in a regression model in order to account for population structure. However, more recently the linear mixed model has been shown to be a powerful method to account for both population structure and cryptic relatedness, which the principal components method fails to account for (Price et al., 2010). Hoffman (2013) examines the relationship between the principal components method and the linear mixed model, and the statistical theory underlying the differences in empirical performance between modelling principal components as fixed versus random effects.

When handling a large set of correlated variables, the principal components epitomize the set using a smaller number of variables that taken together explain most of the variability in the original set. To perform principal regression, in general we simply use principal components as predictors in a regression model in place of the original larger set of variables. *Principal component analysis* (PCA) refers to the statistical procedure by which principal components are computed, and the subsequent use of these components in understanding the data (James et al., 2013). It involves an orthogonal transformation to convert a set of obser-

vations of possible correlated variables into a set of linearly uncorrelated variables called principal components.

PCA provides a tool to find a low-dimensional representation of a data set that contains as much as possible of the total variance. Suppose we have a $n \times m$ data matrix $\mathbf{X}$, with columns $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m$, which is centered to have column-wise zero empirical mean, resulting in $\mathbf{X}^*$. The idea is that each of the n observations are in a m-dimensional space, but not all of these dimensions are equally interesting. The first principal component has the largest possible variance, and thus accounts for as much of the variability in the data as possible.

3.5.1 The data matrix X

In our case, the data matrix $\mathbf{X}$ $(n \times m)$ consists of genotype data for n individuals at m SNPs, where $\mathbf{X}_{i,k} \in \{0, 1, 2\}$, equivalently to G_{ik} in Equation (2.6). In order to apply the tool of principal components, the data matrix needs to have column-wise zero mean. According to Patterson et al. (2006), this is done by subtracting the column mean for each entry. The column mean for column k is:

$$\overline{X}_k = \frac{\sum_{i=1}^{m} \mathbf{X}_{i,k}}{n}.$$

In addition to subtracting the mean, Patterson et al. (2006) include a normalizing step, which normalizes each data column to have the same variance. By setting $p_k = \overline{X}_k/2$ as an estimate of the minor allele frequency, equivalently as done in Equation (2.9), each entry in the corrected matrix $\mathbf{X}^*$ is

$$\mathbf{X}^*_{i,k} = \frac{\mathbf{X}_{i,k} - \overline{X}_k}{\sqrt{2p_k(1-p_k)}}.$$

The matrix $\mathbf{X}^*$ is identical to the matrix $\mathbf{X}$ in Equation (2.8), which means that an estimator for the kinship matrix is

$$\hat{\mathbf{\Phi}} = \frac{1}{2m}\mathbf{X}^*\mathbf{X}^{*\mathsf{T}} \propto \mathbf{X}^*\mathbf{X}^{*\mathsf{T}}$$

and an estimator for the matrix of correlations between the SNP data is proportional to $\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*$.

3.5.2 Singular value decomposition

We define the principal components of $\mathbf{X}^*$ by taking the singular value decomposition of the data matrix $\mathbf{X}^*$. The singular value decomposition separates the $n \times m$ matrix $\mathbf{X}^*$, into the following matrices (Ripley, 1996)

$$\mathbf{X}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}} \tag{3.20}$$

where $\mathbf{U}$ is an $n \times m$ matrix with orthonormal columns, $\mathbf{V}$ is an $m \times m$ orthogonal matrix of principal directions and $\mathbf{\Lambda}$ is an $m \times m$ diagonal matrix of decreasing non-negative singular values s_i of $\mathbf{X}^*$. Normally, the number of SNPs is larger than the number of individuals in a study, $m > n$, so it can be shown that the rank of $\mathbf{X}^*$ is at most $n - 1$. Thus, the diagonal elements s_i for $i > (n - 1)$ are 0. The principal components of $\mathbf{X}^*$ are defined by Ripley (1996) as the columns of the PCA score matrix $\mathbf{T}$, which is defined as

$$\mathbf{T} = \mathbf{X}^*\mathbf{V} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}.$$

There is a connection between the principal components and both the matrices $\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*$ and $\mathbf{X}^*\mathbf{X}^{*\mathsf{T}}$. The matrix $\mathbf{X}^*\mathbf{X}^{*\mathsf{T}}$ is an $n \times n$ symmetric matrix, and can thus be diagonalized in the following way:

$$\mathbf{X}^*\mathbf{X}^{*\mathsf{T}} = \mathbf{W}\mathbf{L}\mathbf{W}^\mathsf{T} \tag{3.21}$$

where $\mathbf{W}$ is a matrix of eigenvectors (each column is an eigenvector), and $\mathbf{L}$ is a diagonal matrix with eigenvalues λ_i in decreasing order on the diagonal. The same can be done for the matrix $\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*$ since it is an $m \times m$ symmetric matrix,

$$\mathbf{X}^{*\mathsf{T}}\mathbf{X}^* = \mathbf{M}\mathbf{C}\mathbf{M}^\mathsf{T} \tag{3.22}$$

where $\mathbf{M}$ is a matrix of eigenvectors (each column is an eigenvector), and $\mathbf{C}$ is a diagonal matrix with eigenvalues c_i in decreasing order on the diagonal.

Using the singular value decomposition of $\mathbf{X}^*$ from Equation (3.20) on the matrix $\mathbf{X}^*\mathbf{X}^{*\mathsf{T}}$, it is clear that

$$\mathbf{X}^*\mathbf{X}^{*\mathsf{T}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{\Lambda}\mathbf{U}^\mathsf{T} = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^\mathsf{T}. \tag{3.23}$$

It follows from Equation (3.21) that the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{X}^*\mathbf{X}^{*\mathsf{T}}$, and the corresponding eigenvalues are the diagonal elements of $\mathbf{\Lambda}^2$, $\lambda_i = s_i^2$.

Applying the same procedure to the matrix $\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*$, we get

$$\mathbf{X}^{*\mathsf{T}}\mathbf{X}^* = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\mathsf{T} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^\mathsf{T}. \tag{3.24}$$

It follows from Equation (3.22) that the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*$, and that the eigenvalues are the diagonal elements of $\mathbf{\Lambda}^2$, $c_i = s_i^2$, correspondingly as for the matrix $\mathbf{X}^*\mathbf{X}^{*\mathsf{T}}$.

It can be shown that

$$\text{rank}(\mathbf{X}^*) = \text{rank}(\mathbf{X}^*\mathbf{X}^{*\mathsf{T}}) = \text{rank}(\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*). \tag{3.25}$$

From Equation (3.23), (3.24) and (3.25) it is clear that $\mathbf{X}^*\mathbf{X}^{*\mathsf{T}}$ and $\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*$ have the same non-zero eigenvalues but different eigenvectors, and that the PCA scores of $\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*$ are the eigenvectors of the matrix $\mathbf{X}^*\mathbf{X}^{*\mathsf{T}}$.

3.5.3 Modelling principal components as fixed versus random effects

The kinship matrix can be directly used as a part of the model for the correlations between outcomes in the random effects method. The principal components approach involves extraction of the leading eigenvectors of the kinship matrix, and usage of these components as additional fixed effects in the model for the outcomes.

Hoffman (2013) includes the first i principal components as fixed effects in a linear model, which takes the form

$$\boldsymbol{Y} = \mathbf{X}_k \boldsymbol{\beta}_k + \mathbf{T}_{1:i} \boldsymbol{\tau} + \boldsymbol{\epsilon} \tag{3.26}$$

where $\boldsymbol{Y}$ is a vector of phenotype values, $\boldsymbol{\beta}_k$ is an unkown vector of fixed effects, $\mathbf{X}_k$ is the design matrix relating to $\boldsymbol{\beta}_k$, $\mathbf{T}_{1:i}$ are the first i principal components with coefficient vector $\boldsymbol{\tau}$ and $\boldsymbol{\epsilon}$ is the normally distributed error term with variance σ_e^2. The principal components are treated as fixed effects, such that maximizing the likelihood involves directly estimating all parameters.

If we now consider the linear mixed model,

$$\boldsymbol{Y} = \mathbf{X}_k \boldsymbol{\beta}_k + \mathbf{Q} + \boldsymbol{\epsilon}$$

as defined in Equation (3.12), we have that $\mathbf{Q} \sim \mathcal{N}(0, 2\hat{\boldsymbol{\Phi}}\sigma_g^2)$. Using Equation (3.23) we get

$$\hat{\boldsymbol{\Phi}} = \frac{1}{2m}\mathbf{X}^*\mathbf{X}^{*\mathsf{T}} = \mathbf{U}\frac{\boldsymbol{\Lambda}^2}{2m}\mathbf{U}^\mathsf{T} = \mathbf{U}\frac{\boldsymbol{\Lambda}}{\sqrt{2m}}\Big(\mathbf{U}\frac{\boldsymbol{\Lambda}}{\sqrt{2m}}\Big)^\mathsf{T} = \mathbf{R}\mathbf{R}^\mathsf{T} \tag{3.27}$$

so that the columns of $\mathbf{R}$ are the principal components of $\mathbf{X}^*$. Using the property of a multivariate Gaussian, $z \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \mathbf{B}z \sim \mathcal{N}(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\mathsf{T})$, and the decomposition in Equation (3.27), it is clear that $\boldsymbol{\alpha} \sim \mathcal{N}(0, 2\sigma_g^2\mathbf{I}) \Rightarrow \mathbf{R}\boldsymbol{\alpha} \sim \mathcal{N}(0, 2\hat{\boldsymbol{\Phi}}\sigma_g^2)$, so that the linear mixed model can be written as

$$\boldsymbol{Y} = \mathbf{X}_e \boldsymbol{\beta}_e + \mathbf{X}_g \boldsymbol{\beta}_g + \mathbf{R}\boldsymbol{\alpha} + \boldsymbol{\epsilon}. \tag{3.28}$$

Hoffman (2013) claims that based on the relationship between Equation (3.26) and (3.28), it is apparent that modelling principal components as fixed or random effects share the same underlying regression model.

3.6 Multiple testing

The number of tests performed when doing GWA studies are ranging from ten to hundreds of thousands, which makes it important to control the type I error rate.

Table 3.1: Possible situations for testing a statistical hypothesis (Walpole et al., 2012).

	H_0 true	H_0 false
Do not reject H_0	Correct decision 1-α	Type II error β = P(type II error)
Reject H_0	Type I error α = P(Type I error)	Correct decision 1-β

As presented in Table 3.1, rejection of the null hypothesis when it is true is called a type I error, and nonrejection of the null hypothesis when it is false is called a type II error (Walpole et al., 2012). Traditionally, type I errors are considered more problematic than type II errors. If a rejected hypothesis allows publication of a scientific finding, a type I error brings a false discovery and the risk of publication of a potentially misleading scientific result (Goeman and Solari, 2014). Type II errors, on the other hand, mean missing out on a scientific result. As the type I errors are likely to be the most surprising and novel findings, they have a high risk of finding their way into publications (Goeman and Solari, 2014).

In hypothesis tests, researchers have bounded the probability of making a type I error by α, an acceptable risk of type I errors, conventionally set at 0.05. However, problems arise when researchers perform many tests, because each test has a probability of producing a type I error. Performing a large number of hypothesis tests in practice guarantees the presence of type I errors among the findings, as seen in the following equations, if we assume independence between tests.

$$\text{P(not making an error in } m \text{ tests)} = (1-\alpha)^m$$
$$\text{P(making at least 1 error in } m \text{ tests)} = 1-(1-\alpha)^m$$

For example, if $m = 10000$ and $\alpha = 0.05$ this gives P(not making an error in m tests) = 0 and P(making at least 1 error in m tests) = 1.

There are many methods of dealing with type I errors when performing multiple tests: those that estimate the false discovery proportion, those that control the *false discovery rate* (FDR) and those that control the *family-wise error rate* (FWER). In this study, the focus will be on the methods that control the FWER.

3.6.1 Family-wise error rate

We have a collection $\mathcal{H} = (H_1, \ldots, H_m)$ of null hypotheses which we would like to investigate. Assume that m_0 of these hypotheses are true, and $m_1 = m - m_0$ are false. The collection of true hypotheses is called $\mathcal{T}$ and the remaining collection of false hypotheses $\mathcal{F} = \mathcal{H} \backslash \mathcal{T}$. The goal of multiple testing is to choose a collection of hypotheses to reject, $\mathcal{R}$. If the p-values for each of the hypotheses $H_1, \ldots, H_m$ are $p_1, \ldots, p_m$, a straightforward choice of $\mathcal{R}$ is all the hypotheses with a p-value less than a threshold T, $\mathcal{R} = \{H_i : p_i \leq T\}$. The numbers of errors occurring in a multiple hypothesis testing procedure can be summarized in a contingency table, as in Table 3.2. The total number of hypotheses m and the number of rejected hypotheses $R = \#\mathcal{R}$ are observable, but all the other elements of the table are unobservable.

Table 3.2: Contingency table for multiple hypothesis testing (Benjamini and Hochberg, 1995).

	True	False	Total
Rejected	V	U	R
Not rejected	m_0	$m_1 - U$	$m - R$
Total	m_0	m_1	m

The multiple testing methods attempt to reject as many hypotheses as possible while controlling the type I errors. The standard approaches to measure the type I errors is the number V of type I errors or the false discovery proportion Q, defined as

$$Q = \begin{cases} V/R, & \text{if } R > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The FWER method controls

$$\text{FWER} = \text{P}(V > 0) = \text{P}(Q > 0),$$

and the FDR method controls

$$\text{FDR} = \text{E}(Q).$$

The FDR method has its focus on the expected proportion of errors among the rejections, whereas the FWER method looks at the probability that the rejected set contains any error (Goeman and Solari, 2014).

The Bonferroni procedure is the most known method for control of the FWER, and it is popular because of its simplicity (Goeman and Solari, 2014). The Bonferroni method controls the FWER at level α for all types of dependence structures

between the tests, by rejecting hypotheses only if they have a p-value smaller than $T = \alpha/m$.

Another method that controls the FWER is the Šidák method, in which one assumes that all the individual tests are independent. The threshold T is then $T = 1 - (1 - \alpha)^{1/m}$.

The method of Halle et al. (2016) is a competitor to the Bonferroni and Šidák methods. The method is constructed for use in a generalized linear model where hypothesis testing is performed using the score test, and is intended for use in the GWA setting. Further, the vector of score tests for the multiple hypotheses (one for each SNP) is asymptotically multivariate normal, and Halle et al. (2016) calculate T using approximations to a m-dimensional multivariate normal distribution. The method takes into account the correlation between test statistics from neighbouring SNPs.

Chapter 4

Analysis

In this chapter we perform the analysis of the HUNT $VO_{2_{max}}$ data.The analysis is executed using the `GenABEL` package, which provides several important methods to analyze genome-wide association studies. By estimating the kinship matrix for all the individuals, using the estimator in Section 2.3.1, it is possible to create a new data set; a reduced sample including only individuals with an estimated kinship coefficient below a defined value. The individuals in the reduced sample are then assumed to be independent, and a multiple linear regression can be fitted, as presented in Section 3.1. The original sample containing all the individuals can be analyzed using a linear mixed model, which was introduced in Section 3.4.

The outline of Chapter 4 is as follows: In Section 4.1 the `GenABEL` package is presented, Section 4.2 consider summaries of quality controls of the data, Section 4.3 presents the $VO_{2_{max}}$ trait and the covariates age and activity level, while Section 4.4 comprise estimation of the kinship matrix and construction of the two data sets. Moreover, Section 4.5 presents results from the methods to control FWER, and Sections 4.6 and 4.7 embody the analyses using linear models and linear mixed models, respectively. Finally, in Section 4.8 the results of the different analyses are compared, as well as the performance of the different statistical methods.

4.1 `GenABEL`

`GenABEL` is an R package for performing statistical analyses of genome-wide association (GWA) studies. Important challenges of the modern computational genetics are to store, handle and analyze GWA data as effectively as possible. In general, the amount of data generated in GWA studies are enormous, as hundreds of thousands of SNPs are genotyped in hundreds or thousands of individuals. The `GenABEL`-package makes it possible to do GWA analysis on standard desktop computers. The library addresses minimization of the amount of rapid access memory

(RAM) used and the time required for data transactions, and maximization of the throughput of GWA analysis[1]. The use of the `GenABEL` package gives access to a wide range of statistical analysis functions. `GenABEL` will be used in the subsequent data analysis.

4.2 Quality control of genetic data

Performing cleaning and quality control of GWA data is important. To do this we use the `GenABEL` function `check.marker`, which does genotypic quality control. This function helps selecting the SNPs of sufficient quality to enter into GWA analysis based on call rate, minor allele frequency (MAF) and p-value of the $\mathcal{X}^2$-test for Hardy Weinberg equilibrium (HWE). Call rate is defined as the proportion of genotypes per SNP with non-missing data. We choose to filter out the SNPs which have call rate below 90%, and use 90% as a cut-off for individual call rate as well (maximum proportion of missing genotypes in a person). The cut-off p-value in assessing HWE is selected so that the FDR is controlled at level 0.2. For the MAF, the cut-off is set to 0.05, so that SNPs with an estimated MAF less than 0.05 are filtered out. This means that SNPs with less than 5% copies of the minor allele are removed. The MAF is estimated using the estimator in Equation (2.9). The quality check is repeated until no further errors are found. The quality check needs to be done in this iterative way because as soon as you exclude a SNP or an individual, other statistics change and need to be computed again and re-checked.

The output from the quality control is listed in Appendix B.1. In total there are 196 725 SNPs and 1472 individuals in the original sample. The number of individuals excluded due to low call rate and SNPs excluded due to low call rate and low MAF, can be found in the summary below. The output shows that in total 102 477 SNPs and 1459 people passed all criteria.

The summary of SNPs and individuals which did not pass the quality control is as follows:

```
$`Per-SNP fails statistics`
          NoCall NoMAF NoHWE Redundant Xsnpfail
NoCall      1891  2410   688         0        0
NoMAF         NA 83918  1798         0        0
NoHWE         NA    NA  3382         0        0
```

[1]Genabel.org (2016). GenABEL: an R package for Genome Wide Association Analysis. Available at: http://genabel.org/genabel/genabel-package.html [Accessed 13 Jun. 2016]

```
$`Per-person fails statistics`
            IDnoCall HetFail IBSFail isfemale ismale
IDnoCall          13       0       0        0      0
HetFail           NA       0       0        0      0
IBSFail           NA      NA      16        0      0
```

From this output we can see that some SNPs fail both call rate and HWE. This result is natural because poor call rate of a SNP often is a sign of problems during genotyping, which also can cause SNPs to be out of HWE.

Descriptive summary tables for the markers that passed the quality control can be seen in Appendix B.2. The tables show that more than 50% of the markers have a MAF>0.2, and the remaining markers have a MAF between 0.05 and 0.2. Moreover, 100 367 SNPs, or 97.9% of all SNPs, had a proportion of 99% successful genotypes, and all individuals had a proportion of 99% successful genotypes. The mean heterozygosity for a SNP is 0.3374 with a corresponding standard deviation of 0.1288, while the mean heterozygosity for a person is 0.3368 with a standard deviation of 0.0075.

4.3 $VO_{2_{max}}$, age and activity level

The data set contains information of $VO_{2_{max}}$, age and activity level for each of the 1459 individuals. Figure 4.1 displays a histogram of the maximal oxygen uptake ($VO_{2_{max}}$) of each of the 1459 individuals. The $VO_{2_{max}}$ data has a mean of 136.434 $ml/kg^{0.75}/min$ with associated standard deviation of 25.812, as shown in Table 4.1. The minimum value of $VO_{2_{max}}$ is 65.15 $ml/kg^{0.75}/min$ and the maximum is 222.06 $ml/kg^{0.75}/min$. The blue curve in the figure represents a normal distribution with mean and standard deviation corresponding to the $VO_{2_{max}}$ data. The data follows the normal distribution well, but deviations from the normal distribution is easier to see from a Q-Q plot, which will be presented in Section 4.6.

Table 4.1 also presents summaries for age and activity level for each of the individuals. The age of the individuals ranges from 19.6 years to 84.4 years, with a mean of 49.483 years. The activity level ranges from 0 to 15, and the mean is 3.409.

Table 4.1: Descriptive summary statistics of the data for the 1459 individuals including $VO_{2_{max}}$, age and activity level.

	n	Mean	SD	Min	Max
Age	1459	49.483	12.749	19.6	84.4
Activity level	1447	3.409	2.950	0	15
$VO_{2_{max}}$	1458	136.434	25.812	65.15	222.06

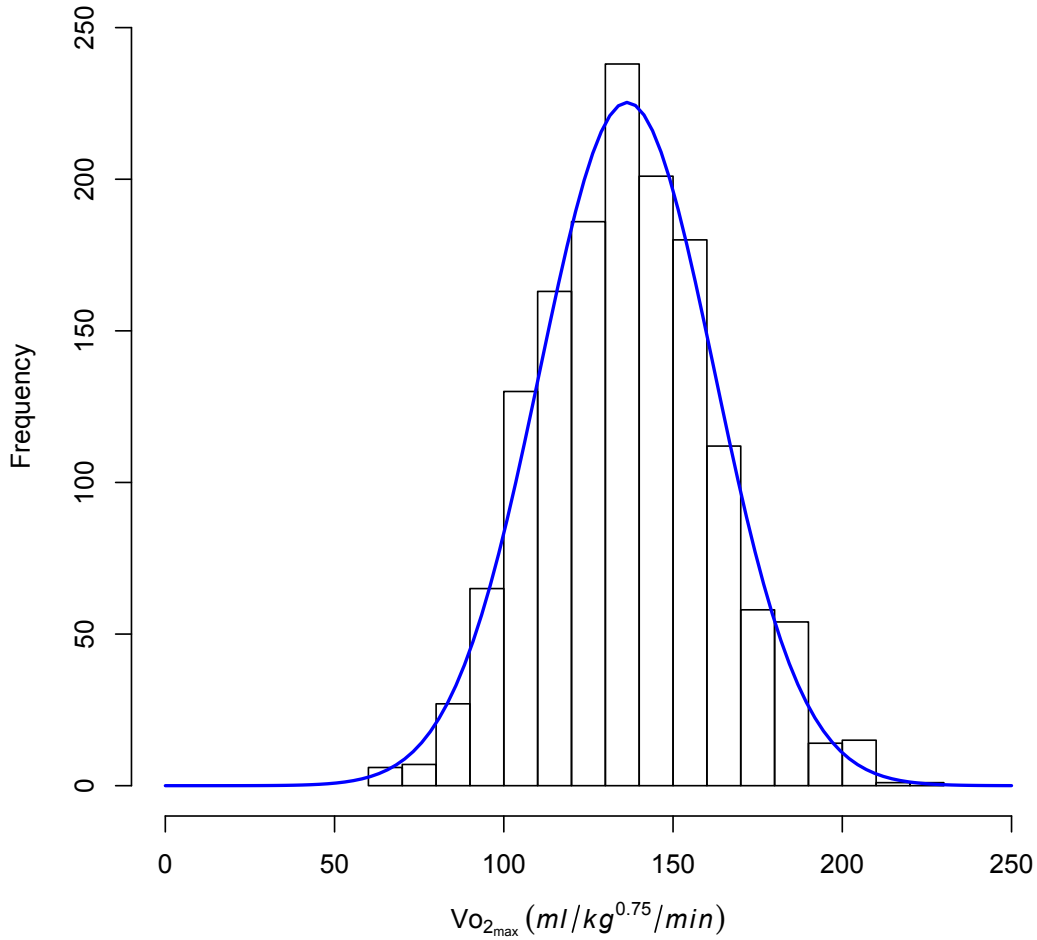

Figure 4.1: Histogram of the $VO_{2_{max}}$ for all the 1459 individuals. The blue line represents a normal distribution with mean of 136.434 and standard deviation of 25.812.

4.4 Estimation of the kinship matrix for the $VO_{2_{max}}$ data

To investigate the possible relatedness between the individuals we estimate the kinship matrix for the individuals. To estimate the kinship matrix, the `GenABEL` function `ibs` is applied. The function computes a matrix of kinship coefficients for a group of individuals, based on a set of SNPs. The function has some arguments, including the argument *weight*, which can either be assigned to *no*, *eVar* or *freq*. In this thesis, *freq* is used, which means what the allelic frequency is applied, and the kinship coefficient is estimated in a similar way as in Equation (2.7). The only difference is that the genotype of an individual at a SNP is coded as 0, 1/2 and 1, while we used the coding 0, 1 and 2 in Equation (2.6). This is made up for in the `ibs` function by using the following expression for estimating the kinship coefficient:

$$\hat{\Phi}_{ij,\text{ibs}} = \frac{1}{m}\sum_{k=1}^{m}\frac{(x_{ik}-p_k)(x_{jk}-p_k)}{p_k(1-p_k)} \tag{4.1}$$

where x_{ik} and x_{jk} are the genotypes of individual i and j, respectively, at the kth SNP, coded as 0, 1/2 and 1. Using Equation (4.1) is equivalent to using Equation (2.7).

The estimated kinship matrix is a 1459×1459 matrix, and the the total time to compute it was 9.1 minutes on a computer of the type Intel Core i7-4770 (quad core, 3.4 GHz). Figure 4.2 shows a histogram of all the estimated pairwise kinship coefficients (not including the diagonal elements). In total the histogram includes $\frac{1459 \cdot 1458}{2}$ =1 063 611 estimated kinship coefficients. From the histogram it is clear that many pairs of individuals have an estimated kinship coefficient that is negative, which correspond to individuals sharing fewer alleles than expected. Very negative coefficients may indicate that there is population structure between the individuals, and is therefore not set to 0.

Plots of the largest, second largest and third largest estimated kinship coefficients for each individual are displayed in Figure 4.3. The largest estimated kinship coefficient for each individual $i \in 1, \ldots n$, is the maximum of $\hat{\Phi}_{ij}$ for one $j \in 1, \ldots n$, $j \neq i$. The plots show that some individuals have an estimated kinship coefficient of 0.3. The histogram of the third largest estimated kinship coefficients indicates that there are not that many individuals that are closely related to several others, as the estimated kinship coefficients are much smaller than for the largest estimated kinship coefficients.

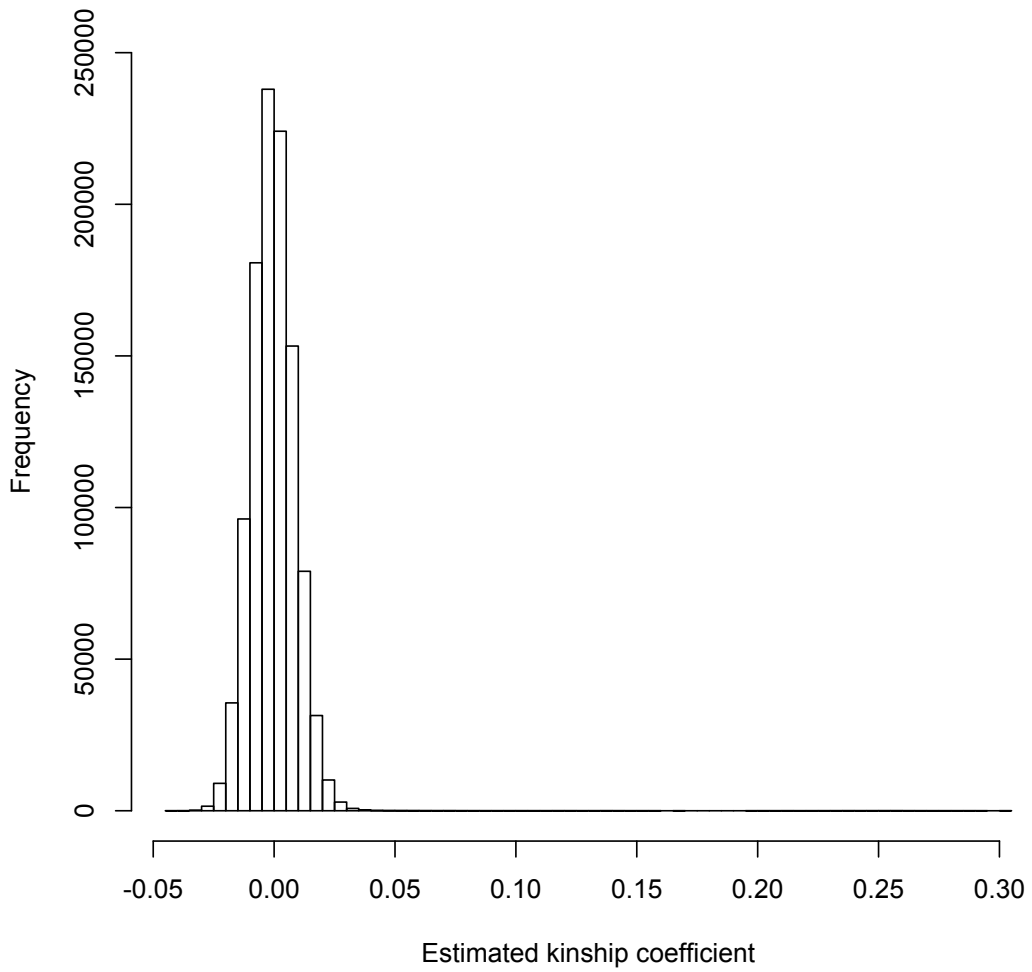

Figure 4.2: Histogram of the estimated kinship coefficients for the 1459 individuals based on 102 477 SNPs.

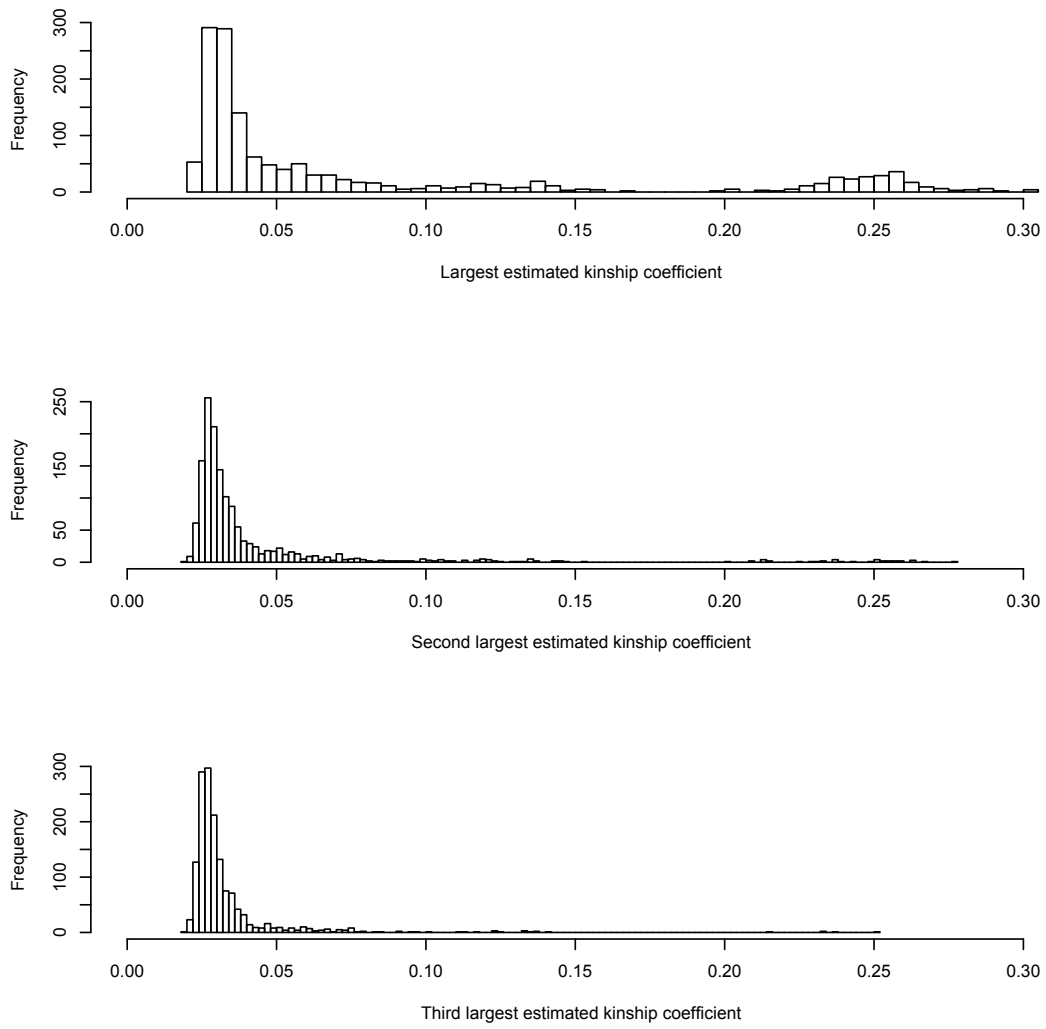

Figure 4.3: Histogram of the largest, second largest and third largest estimated kinship coefficient for each of the 1459 individuals based on 102 477 SNPs. In total 1 063 611 kinship coefficients are estimated.

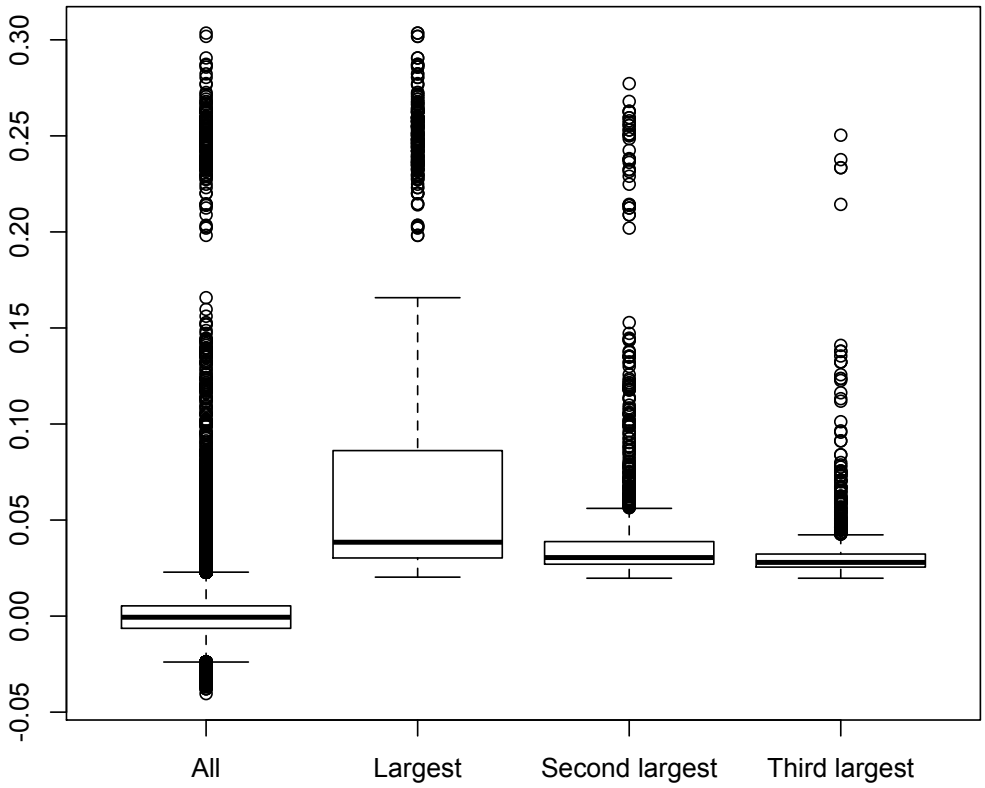

Figure 4.4: Box plots of all the estimated kinship coefficients, the largest, second largest and third largest estimated kinship coefficients for each individual.

Figure 4.4 displays box plots of all the estimated kinship coefficients and the largest, second largest and third largest estimated kinship coefficients for each individual. The box plot of all coefficients shows that the median is 0, and that many coefficients are negative, and that some pairs of individuals seem to be closely related. The box plots of the largest, second largest and third largest coefficients indicate that the median decreases when going from largest to second largest and third largest, as expected. Moreover, there is a reduction in the number of pairs of individuals that are closely related going from the box plot of the largest coefficients to the box plot of the third largest coefficients.

The mean of the estimated kinship coefficients, not including the diagonal elements and setting the negative coefficients to 0, are 0.003345. The minimum esti-

mated kinship coefficient is 0 and the maximum is 0.3035. The standard deviation of the estimated kinship coefficients is 0.00605933. Out of 1 063 611 estimated kinship coefficients, 561 162 was estimated to be negative, and thus set to 0. The fact that many of the estimated kinship coefficients are low and several are negative, means that many individuals in the population are not related.

Table 4.2 is a table of estimated kinship coefficients for 5 individuals of the sample. The individuals are chosen to illustrate the largest values of the estimated kinship matrix, because each of them have a large estimated kinship coefficient with at least one of the other. The diagonal elements are close to the expected value of 0.5. It is clear that person 1 and person 2 are closely related, with an estimated kinship coefficient $\hat{\Phi}_{12} = 0.303518$, which is the maximum estimated kinship coefficient. The same is the case for person 3, 4 and 5, in which $\hat{\Phi}_{34} = 0.301800$, $\hat{\Phi}_{35} = 0.257851$ and $\hat{\Phi}_{45} = 0.250882$.

Table 4.2: Table of the estimated kinship coefficients for 5 selected individuals of the sample.

	Person 1	Person 2	Person 3	Person 4	Person 5
Person 1	0.498258	0.303518	-0.000452	-0.013020	0.013236
Person 2	0.303518	0.493042	-0.002097	-0.012986	0.006839
Person 3	-0.000452	-0.002097	0.502639	0.301800	0.257851
Person 4	-0.013020	-0.012986	0.301800	0.493582	0.250882
Person 5	0.013236	0.006839	0.257851	0.250882	0.511947

The two data sets

The main focus of this subsection is to create two data sets, to be referred to as the *full* sample and the *reduced* sample, where the full sample includes all the 1459 individuals in this study, and the reduced sample only contains the individuals that are not closely related. In this context, not closely related is interpreted as being further apart than cousins, which means having a kinship coefficient less than 0.125. Since the estimated kinship coefficient for cousins can vary around 0.125 (Speed and Balding, 2015), the threshold for the kinship coefficient is set to be 0.1. This ensures that every pair of individuals are not closely related. Only one of the individuals in a pair of individuals with estimated kinship coefficient >0.1 is included in the reduced sample. So for every pair of individuals i and j, the estimated kinship coefficient is computed, and if $\hat{\Phi}_{ij} > 0.1$, one of the individuals is removed from the reduced sample, and the other one is included in the reduced sample. Which of the individuals i and j is removed, is chosen randomly, and if the individual chosen to be removed is already removed then the other individual are kept in the reduced sample. This procedure removed 185 individuals, and thus

included 1274 individuals in the reduced sample. Since the removed individuals are chosen randomly, this procedure does not ensure uniqueness of the reduced sample.

Descriptive summary statistics for the reduced sample are presented in Table 4.3, and descriptive summary statistics for the removed individuals are presented in Table 4.4. The summaries for the full sample are found in Table 4.1.

Table 4.3: Summaries of the data for the reduced sample including $VO_{2_{max}}$, age and activity level.

	n	Mean	SD	Min	Max
Age	1274	49.488	12.621	19.6	84.4
Activity level	1264	3.427	2.956	0	15
$VO_{2_{max}}$	1273	136.397	25.701	65.15	222.06

Table 4.4: Summaries of the data for the removed individuals including $VO_{2_{max}}$, age and activity level.

	n	Mean	SD	Min	Max
Age	185	49.454	13.636	20.7	77.9
Activity level	183	3.283	2.911	0	15
$VO_{2_{max}}$	185	136.690	26.632	65.81	205.17

The results show that the means of the age, activity and $VO_{2_{max}}$ measurement are approximately equal for the different samples. It is of interest to check this statistically, by applying t-tests with the null-hypothesis that the population means of the reduced sample and the removed individuals are equal. The outputs from the t-tests are listed in Appendix B.3. The results show that there is no evidence suggesting rejection of the null hypotheses that the means of activity level, age and $VO_{2_{max}}$ are equal for the reduced sample and the sample of the removed individuals, using significance level $\alpha = 0.05$.

4.5 Choice of local significance level

In this section the results from multiple testing are presented, which sets local significance levels. The theory behind and the different methods to control the FWER are discussed in Section 3.6.1. The different methods are Bonferroni, Šidák and the two methods of Halle et al. (2016). All the approaches produce a threshold T, for which hypotheses with a p-value lower than T are rejected. The threshold is also named α_{loc}.

Table 4.5: The local significance level α_{loc} produced by the different methods to contol FWER.

	α_{loc}
Bonferroni	5.513287e-06
Šidák	5.655877e-06
α_{loc2}	6.853087e-06
α_{loc3}	7.694294e-06

The methods of Halle et al. (2016) utilize SNP data, and takes into consideration the linkage or correlation between neighbouring SNPs. In order to make the analysis less complex and time consuming when performing hypotheses tests on each SNP, the focus will only be on SNPs on chromosome 1. In total there are $m = 9069$ SNPs on chromosome 1.

Table 4.5 shows the values of the local significance level α_{loc} for the different methods to control FWER. It is clear from the table that the Bonferroni method is more strict than the other methods, and that the methods of Halle et al. (2016) (α_{loc2} and α_{loc3}) are the least stringent. The significance levels α_{loc2} and α_{loc3} are based on SNP data of chromosome 1 from the individuals in the reduced sample, since the method of Halle et al. (2016) is constructed for generalized linear models and score tests. We choose to control the FWER with the use of the threshold $\alpha_{\text{loc}} = 7.694294 \cdot 10^{-6}$ for the analyses of the linear models and the linear mixed models.

4.6 Analysis using linear models

In this section the reduced sample will be analyzed, first without including any genetic covariates. This is done to check if the data set with only age and activity level as covariates fits a linear model. Subsequently, the genetic covariates will be included, one at a time. Finally, a linear model will be fitted to the full sample in order to compare the results of the reduced and the full sample using linear regression models.

4.6.1 No genetic covariates

It is of interest to check if the variables for the reduced sample fits a linear model. The response y_i is the maximal oxygen uptake for individual i, $i = 1, \ldots, n$, where $n = 1274$, in which y_i is treated as a realization of a random variable Y_i. The random variable Y_i is assumed to follow the distribution given in Equation (3.4), in which the covariates are age and activity level for each individual. The vector of

regression coefficients, $\boldsymbol{\beta}_k$, is a 3-dimensional vector consisting of the intercept β_0, β_{age} and β_{act}. Figure 4.5 shows scatterplots for all pairs of variables. Note that, not surprisingly, $\text{VO}_{2_{\text{max}}}$ is positively associated with activity level and negatively associated with age.

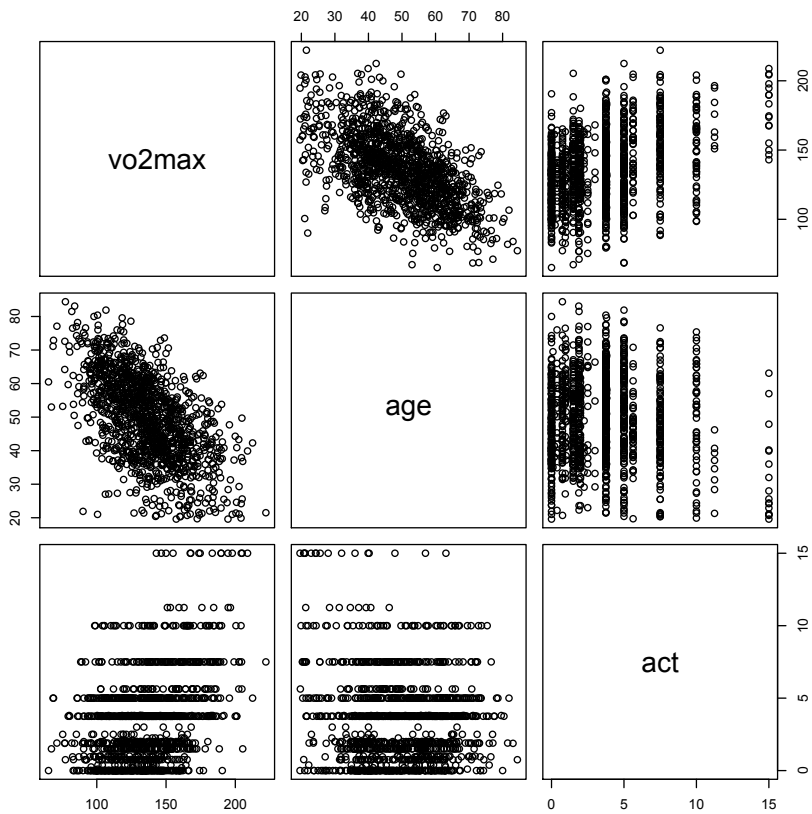

Figure 4.5: Scattergrams of age, activity level and $\text{VO}_{2_{\text{max}}}$ for the reduced sample.

Fitting the two-covariate model, with $\text{VO}_{2_{\text{max}}}$ as response and age and activity level as linear coefficients, produces the output from Listing 4.1.

From the output in Listing 4.1 the estimates of the coefficients, standard errors, test of hypotheses, residual standard error and coefficient of determination R^2 can be read off. The residual standard error is $\hat{\sigma} = 19.14$, which provides a measure of the extent to which individuals with the same age and level of activity can experience different declines in the $\text{VO}_{2_{\text{max}}}$. The degrees of freedom is the number of observations minus the number of parameters estimated. Starting with 1274 observations, 11 was deleted due to missing values, which results in 1263 observations. The estimated parameters are the intercept, age and activity coefficients, which yield a total of 1260 degrees of freedom.

Listing 4.1: Results from `lm` of reduced sample without genetic covariates.

```
Call:
lm(formula = vo2max ~ age + act)

Residuals:
   Min      1Q Median      3Q     Max
-67.73 -12.82   0.51  12.97  63.32

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 182.57263    2.32070   78.67   <2e-16 ***
age          -1.13098    0.04287  -26.38   <2e-16 ***
act           2.88328    0.18273   15.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
   1

Residual standard error: 19.14 on 1260 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.4457,  Adjusted R-squared:  0.4448
F-statistic: 506.5 on 2 and 1260 DF,  p-value: < 2.2e-16
```

The intercept is estimated to be 182.57263, which is interpreted to be the value of the $VO_{2_{max}}$ in the case when both age and activity level is 0. This is not a meaningful interpretation, since the age of the individuals ranges from 19 to 85 years. We find that for every additional year of age we expect the $VO_{2_{max}}$ to decrease by an average of 1.13098, while holding the activity level constant. The estimated standard error of the regression coefficient is 0.04287, and the corresponding t-test (t-statistics is -26.38 on 1260 degrees of freedom) is highly significant. Similarly, we expect the $VO_{2_{max}}$ to increase by an average of 2.88328 for each one-unit difference in the activity level, if the age remains constant. The estimated standard error of the regression coefficient is 0.18273, and the t-test (t-statistic is 15.78 on 1260 degrees of freedom) yields a significant linear association between activity level and maximal oxygen uptake.

The coefficient of determination is $R^2 = 0.4457$ and the adjusted $R^2_{\text{adj}} = 0.4448$ are useful summaries of the proportion of the variation in the data set explained by the regression. The value of R^2 illustrate that 44.57% of the variance is explained as linear effects of age and activity level.

It is necessary to evaluate if the data meets the assumptions of the linear model listed in Section 3.1. Figure 4.6 shows results from the analysis of the reduced sample using `lm`.

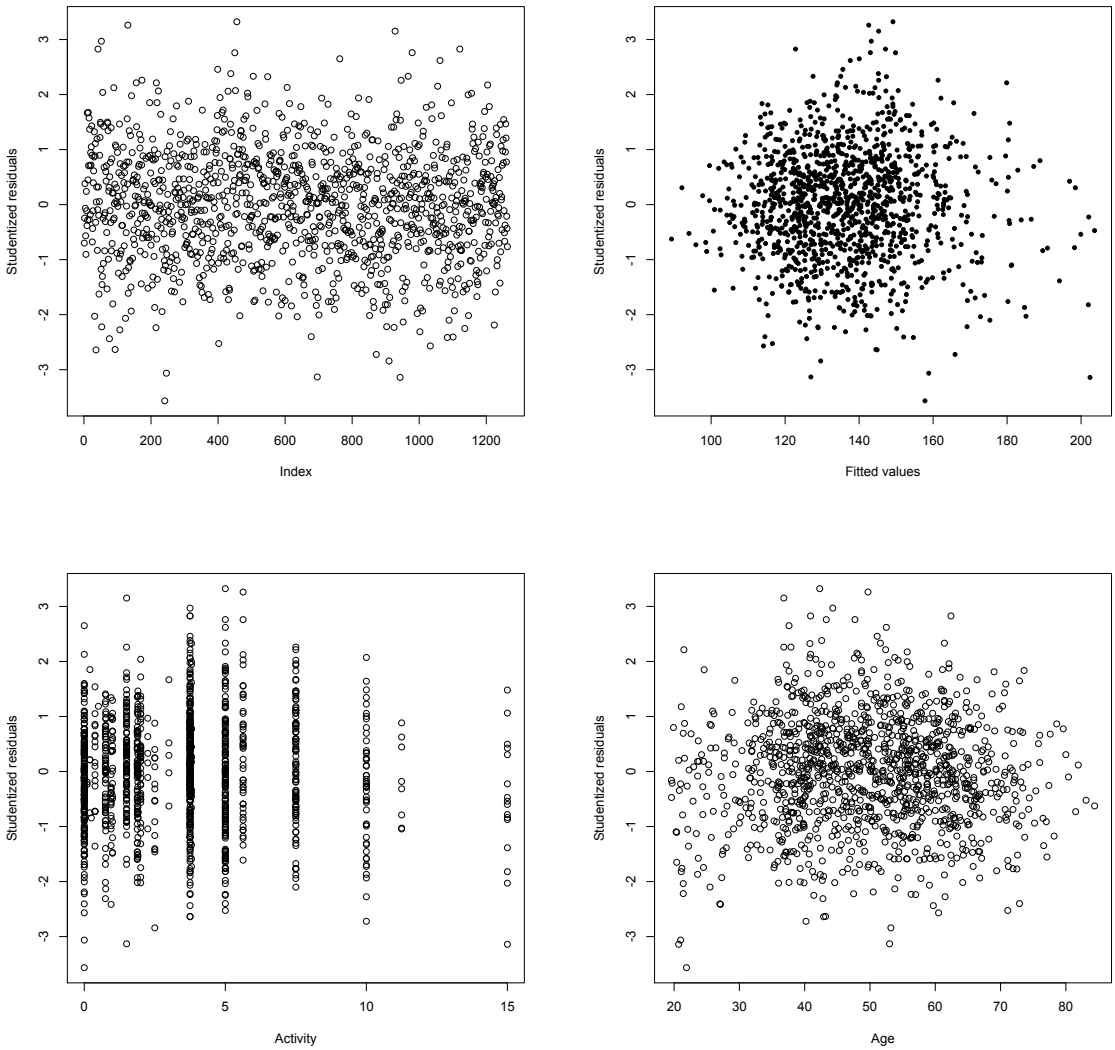

Figure 4.6: Plots from the analysis of the reduced sample using `lm`, including studentized residuals (upper left panel), studentized residuals against fitted values (upper right panel), studentized residuals versus activity level (lower left panel) and studentized residuals versus age (lower right panel).

The upper left panel displays the studentized residuals, which illustrates that the residuals appear independent and that they fluctuate around 0. Since the residuals are estimates of the errors, we conclude that the errors are independent.

In the upper right panel the studentized residuals are plotted against the fitted values, in which the points are centered around a horizontal line. This gives an indication of linearity.

The model is assumed to be additive, in the sense that the effect of each covariate on the response is assumed to be the same for all values of the other covariate. In terms of our data, the model assumes that the effect of age is exactly the same at every activity level. Non-linearity and non-additivity may also be revealed by systematic patterns in plots of the studentized residuals versus individual independent variables. The plots of studentized residuals versus activity level and studentized residuals versus age in Figure 4.6, lower left panel and lower right panel, respectively, show no patterns, which again are signs of linearity and additivity.

Moreover, to check the assumption of equal variance (homoscedasticity) of the errors, the plots of studentized residuals as a function of the fitted values (upper right panel) and the studentized residuals as a function of the independent variables (lower left and right panel) are applied again. The fact that the residuals don't increase or decrease as a function of the fitted values or independent variables, is a verification of homoscedasticity of the errors.

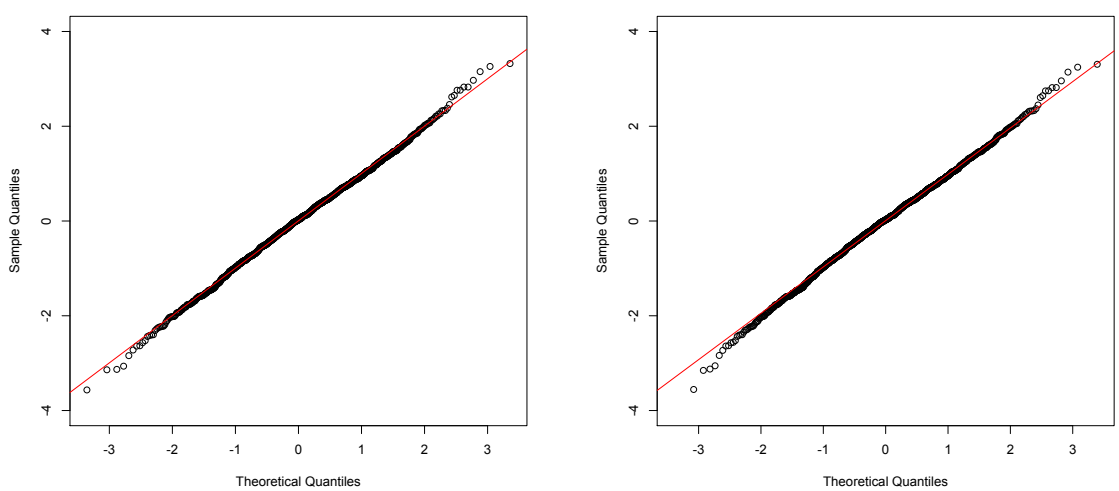

Figure 4.7: Q-Q plot of the residuals from fitting a linear model to the reduced sample in the left panel, and full sample in the right panel.

Finally, the assumption of normality of the error distribution is assessed by a quantile-quantile plot (Q-Q plot). The Q-Q plot of the residuals is displayed in the left panel of Figure 4.7, and it is clear that the points are close to the red line which is a sign that the distribution fits the data well. Comparing to the right panel of Figure 4.7, which shows the Q-Q plot of the residuals from fitting the same linear model to the full sample, the tails do not follow the line as good as for the reduced sample, which is a sign of deviation from the normal distribution.

The Anderson-Darling normality test is also applied to discover the normality of the error distribution. The test gives a statistic $A = 0.51681$ and p-value p= 0.1894 for the reduced sample, which indicates that the normal distribution fits the data well. When analyzing the full sample using linear regression, the result gave an Anderson-Darling test statistic $A = 0.75227$ with a p-value $p = 0.05015$. This resulted in accepting the null hypothesis that the full sample follows the normal distribution. Comparing the two results from the Anderson-Darling normality test show that the reduced sample follows the normal distribution slightly better than the full sample. The results from the Anderson-Darling normality test and the Q-Q plots are part of the motivation to analyze the full sample using a linear mixed regression model.

Figure 4.8 shows the estimated regression equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_{\text{age}} X_{\text{age}} + \hat{\beta}_{\text{act}} X_{\text{act}}$ evaluated for the reduced sample. The regression plane may be considered as an infinite set of regression lines. For any fixed value of age, the expected $\text{VO}_{2_{\text{max}}}$ incline is a linear function of activity level with slope 2.88. Correspondingly, for any fixed value of activity level, the expected $\text{VO}_{2_{\text{max}}}$ decline is a linear function of age with slope -1.13.

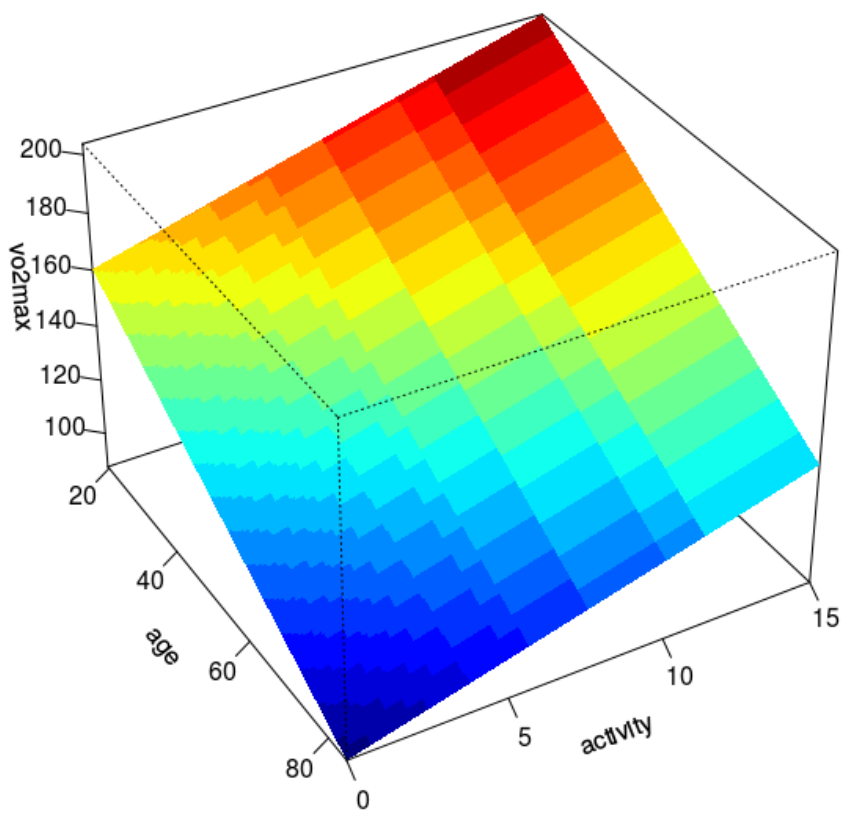

Figure 4.8: Fitted model from multiple linear regression of $VO_{2_{max}}$ on activity level and age, for the reduced sample.

4.6.2 Including genetic covariates

To analyze the reduced sample including one genetic covariate at the time, the `GenABEL` function `mlreg` is applied. The aim is to perform a hypothesis test for each SNP. The total number of genetic covariates to check for association with the response is $m = 9069$. As for the analysis without genetic covariates, the response y_i is the maximal oxygen uptake for individual i, $i = 1, \ldots, n$, where $n = 1274$, in which y_i is treated as a realization of a random variable Y_i. The random variable Y_i is assumed to follow the distribution given in Equation (3.4), in which the covariates are age, activity level and genotype data for the kth SNP for each individual. The vector of regression coefficients, $\boldsymbol{\beta}_k$, is a 4-dimensional vector consisting of the intercept β_0, β_{age}, β_{act} and β_{g_k}, where $k \in 1, \ldots, m$, $m = 9069$.

The summary for the top 10 results of the `mlreg` regression, sorted by p-values, "P1df", is found in Table 4.6. The results show that the top 3 most significant SNPs to be associated with maximal oxygen uptake are rs10921875, rs1218592 and rs155902.

Table 4.6: Results from fitting a linear model to the reduced sample using `mlreg`, sorted by p-values "P1df".

	Position	n	effB	se_effB	chi2.1df	P1df	Pc1df
rs10921875	187604620	1263	-3.001139	0.7477329	16.10940	0.0000597867	0.0002331088
rs1218592	153156109	1260	3.158232	0.8911609	12.55960	0.0003941767	0.0011562126
rs155902	187843884	1263	-2.567510	0.7517374	11.66519	0.0006368039	0.0017383949
chr1:74819417	74819417	1262	-4.015202	1.2031892	11.13645	0.0008464761	0.0022146012
chr1:74823928	74823928	1260	-4.015720	1.2043753	11.11740	0.0008552151	0.0022340389
chr1:74812647	74812647	1263	-3.994218	1.2013972	11.05326	0.0008853159	0.0023007686
chr1:74813852	74813852	1263	-3.994218	1.2013972	11.05326	0.0008853159	0.0023007686
chr1:74832901	74832901	1263	-3.994218	1.2013972	11.05326	0.0008853159	0.0023007686
chr1:74840649	74840649	1263	-3.994218	1.2013972	11.05326	0.0008853159	0.0023007686
chr1:74841681	74841681	1263	-3.994218	1.2013972	11.05326	0.0008853159	0.0023007686

As for the analysis without genetic covariates in Section 4.6.1, some individuals are deleted from the analysis due to missingness. The column "n" gives the number of individuals included for the analysis of each SNP, and it is clear that for most SNPs 1263 individuals are part of the analysis, as for the case without genetic covariates. The column "effB" gives the estimates of the coefficients for each SNP, $\hat{\beta}_{g_{(k)}}$, with standard errors in column "se_effB ". The χ^2_1-statistics, T_k, in column "chi2.1df" are computed from dividing each regression coefficient by its standard deviation, and then squaring the result:

$$T_k = \left(\frac{\beta_{g_{(k)}}}{se_k}\right)^2 \sim \mathcal{F}_{1,n-p-1} \xrightarrow{n\to\infty} \mathcal{X}^2_1. \tag{4.2}$$

Equation (4.2) shows that T_k is asymptotically $\mathcal{X}^2_1$-distributed. The p-values

corresponding to the test statistics are found by calculating

$$P_k = \mathrm{P}(F_{1,n-p-1} > T_k) \approx \mathrm{P}(\chi_1^2 > T_k),$$

and the p-values from the 1-degree of freedom test for association between each SNP and trait are found in column "P1df". In order to find the p-value of the top SNP rs10921875, we first compute the test statistic $T = (-3.001139/0.7477329)^2 = 16.10940$, which is used to calculate the p-value $p = \mathrm{P}(\chi_1 > 16.10940) = 5.978655 \cdot 10^{-5}$.

The genomic control inflation factor from fitting a linear model using `mlreg` to the reduced sample, is estimated to be $\hat{\lambda} = 1.16864$. The function `estlambda` of the `GenABEL`-package is used to estimate λ, applying the method *median* equivalently as in Equation (3.11). The column "Pc1df" gives p-values of the same test for association, as in column "P1df", only that the test statistics are corrected for possible inflation. Dividing the test statistics in column "chi2.1df" by $\hat{\lambda}$ estimated using the function `estlambda` with the method *regression*, and then computing the corresponding p-values, gives the elements in column "Pc1df". The estimate using the *regression* method is $\hat{\lambda} = 1.189464$, so the differences in the estimates using *median* and *regression* are not big. We will continue using the *median* method. Sorting the results from fitting a linear model to the reduced sample using `mlreg` by "Pc1df" gives the same SNPs as in Table 4.6.

The `mlreg` function doesn't provide estimates of the environmental regression coefficients, as a matter of problems with storage of the enormous amounts of data produced by the GWA analyses. The estimates of the coefficients of the SNPs and testing of each SNP are the main priorities.

Manhattan plots are used to illustrate which SNPs are the most associated to the response. The horizontal axis displays the genomic coordinates of chromosome 1, while the vertical axis represents the negative (base 10) logarithm of the association p-value for each SNP. Thus, each point on the plot signifies one SNP. The most significant associations correspond to the largest negative logarithm values.

The upper panel of Figure 4.9 displays a Manhattan plot for the single SNP analyses, while a Manhattan plot of SNPs with -$\log_{10}(p) > 1.5$ is shown in the lower panel, which makes it easier to see the SNPs that have the smallest p-values. Moreover, the roughly blue lines of SNPs make it clear that many of the SNPs with the smallest p-values are positioned in the same region on the chromosome. The red line represents the local significance level $\alpha_{\mathrm{loc}} = 7.694294 \cdot 10^{-6}$, while the green line intersects the 10th most significant SNP. Accordingly, it is clear from the plot that many of the top SNPs are at the same position of the chromosome. Moreover, both the plots and Table 4.6 demonstrate that none of the SNPs have a p-value below the local significance level, and thus can't be considered to have a significant association to the trait.

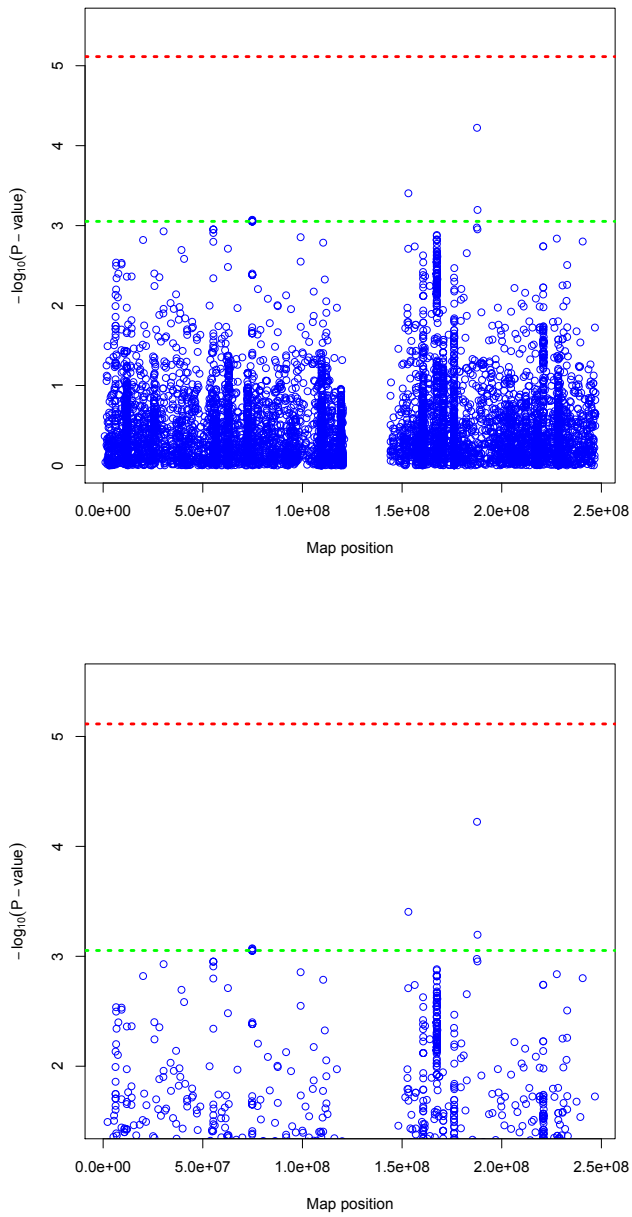

Figure 4.9: Manhattan plots for the single SNP analyses of the reduced sample using `mlreg`. The red lines represent the local significance level α_{loc}, while the green lines intersect the 10th most significant SNP. The upper panel displays all SNPs, while the lower panel for simplicity only displays SNPs with -$\log_{10}(p) > 1.5$.

The five last SNPs in the top 10 list all share the same estimated coefficient, standard error, test statistic and p-value. The reason for this is that the genotype data for the individuals are the same at these SNPs. That is, all the 1263 individuals have the exact same genotype data at these 5 SNPs. Consequently, fitting a linear model including each of these SNPs gives identical results.

Furthermore, the fact that only three SNPs are above the green line in Figure 4.9, and that the rest are clumped and intersected by the green line, are evidences of that these SNPs are correlated or in linkage disequilibrium. In order to demonstrate this, the R package `cgmisc` is utilized. The package is useful for visualizing results of GWA analyses. The function `plot.manhattan.LD` plots local linkage disequilibrium pattern on a Manhattan plot resulting from a GWA analysis, relative to a pre-selected SNP. Each of the SNPs in the fixed interval of the chromosome, is coloured according to its linkage disequilibrium r^2 (see Equation (2.1)) with the reference SNP. The colour codes are not continuous, but the linkage disequilibrium is discretized into intervals, as can be seen in the top left corner. In addition, the MAFs are plotted in the lower panel.

First, the neighbouring SNPs of the top SNP of Table 4.6, rs10921875, will be studied. Figure 4.10 shows the local linkage disequilibrium pattern of rs10921875, which is plotted as a black circle. It is clear from the plot that there are 2 SNPs relatively close and highly correlated to the SNP, the red and orange SNPs. The SNP that is coloured red have a linkage disequilibrium coefficient $0.8 < r^2 < 1$, which corresponds to high level of linkage. Using the function `choose.top.snps` it is possible to find the SNPs with highest r^2 to a given reference SNP. Table 4.7 gives the names of the 2 SNPs highest correlated with SNP rs10921875, the value of r^2 and the coordinates of their position on the chromosome.

Table 4.7: Top SNPs linked with the index SNP rs10921875.

SNP	r^2	coord
rs10921875	INDEX SNP	187604620
rs1935660	0.809637719224218	187509749
rs516084	0.633792841245244	187822397

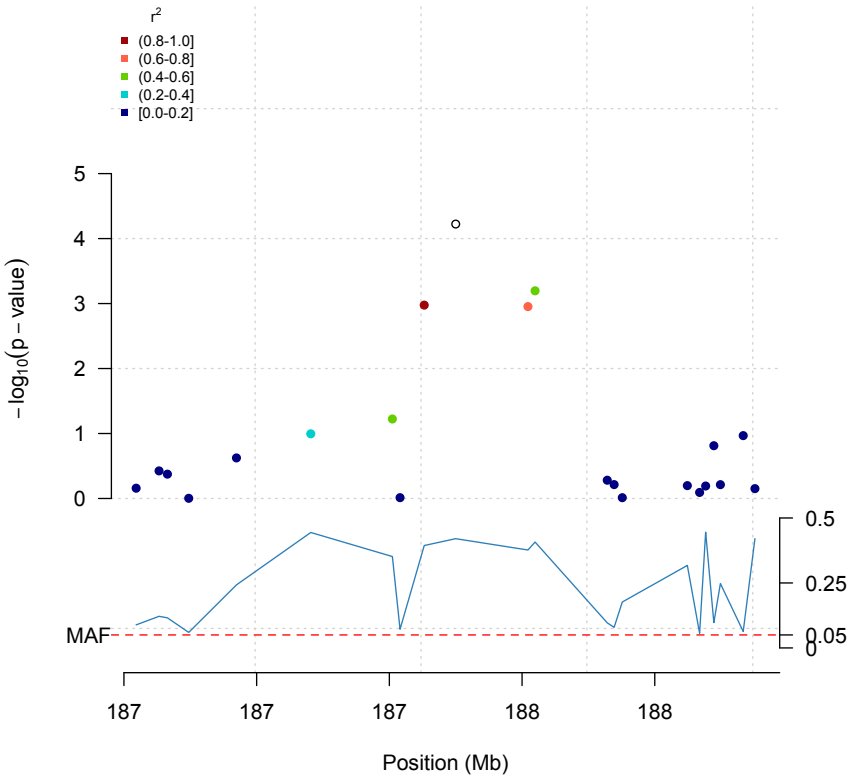

Figure 4.10: Plot of local linkage disequilibrium pattern relative to the pre-selected SNP rs10921875, on a Manhattan plot. Each SNP is coloured according to its linkage disequilibrium with the reference SNP, and the legend in the top left corner represents the colour codes of the linkage r^2. The lower panel displays the MAFs, which are relatively constant for the most correlated SNPs.

Moreover, it is of interest to check the linkage disequilibrium of the SNPs from Table 4.6 which produced the exact same regression coefficients and consists of identical genotype data. Using the SNP chr1:74812647 as reference SNP, the plot of linkage disequilibrium pattern on a Manhattan plot is displayed in Figure 4.11. The reference SNP is barely visible, because a cluster of SNPs highly correlated to the reference SNP is positioned on top of it. The cluster consists of SNPs coloured red, which means that they are in high linkage disequilibrium with the reference SNP.

There are several clusters of SNPs coloured both red, orange and blue, which means that many of the SNPs in this region are in linkage disequilibrium with the reference SNP. The lower panel shows that the MAFs fluctuate for the SNPs around the index SNP. Table 4.8 presents the top 9 SNPs with highest correlation coefficient r^2 with the index SNP chr1:74812647. We recognize that the third to tenth top SNPs of Table 4.6 are representative in Table 4.8, with linkage disequilibrium coefficient $r^2 = 1$. This indicates that these SNPs are in high linkage disequilibrium and provide nearly identical information.

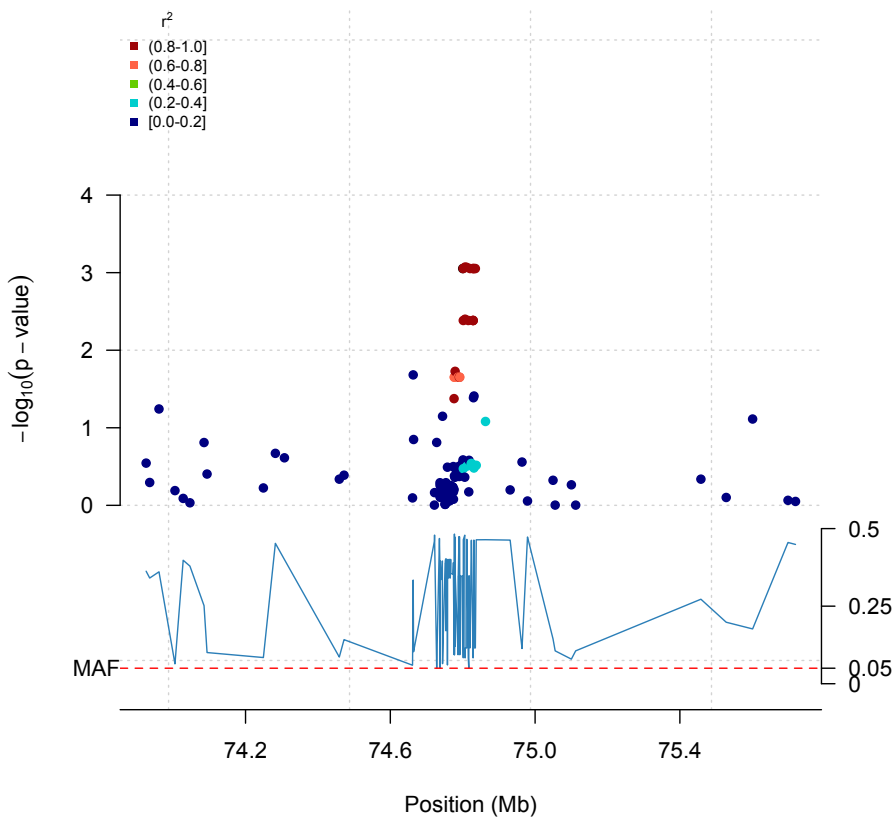

Figure 4.11: Plot of local linkage disequilibrium pattern relative to the pre-selected SNP chr1:74812647, on a Manhattan plot. Each SNP is coloured according to its linkage disequilibrium with the reference SNP, and the legend in the top left corner represents the colour codes of the linkage r^2. The lower panel displays the MAFs, which fluctuate for the SNPs in the region around the reference SNP.

Table 4.8: Top SNPs linked with the index SNP chr1:74812647.

SNP	r^2	coord
chr1:74812647	INDEX SNP	74812647
chr1:74819417	1	74819417
chr1:74847576	1	74847576
chr1:74844626	1	74844626
chr1:74841681	1	74841681
chr1:74840649	1	74840649
chr1:74832901	1	74832901
chr1:74823928	1	74823928
chr1:74813852	1	74813852
chr1:74791630	0.880729101281149	74791630

Finally, we want to fit a linear regression model to the reduced sample including the covariates age, activity level and the top SNP rs10921875. The summary for the linear regression is shown in Listing 4.2. The top SNP has an estimated regression coefficient of -3.00114, which means that when going from 0 to 1 copy of the minor allele, the $VO_{2_{max}}$ decreases on average by 3.00114.

Listing 4.2: Results from `lm` of reduced sample including SNP rs10921875.

```
Call:
lm(formula = vo2max ~ age + act + top_SNP)

Residuals:
    Min      1Q  Median      3Q     Max
-67.296 -12.416   0.531  12.628  63.804

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 185.15348    2.39485  77.313  < 2e-16 ***
age          -1.13179    0.04261 -26.560  < 2e-16 ***
act           2.87759    0.18165  15.842  < 2e-16 ***
top_SNP      -3.00114    0.74773  -4.014 6.33e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
   1

Residual standard error: 19.03 on 1259 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.4527,  Adjusted R-squared:  0.4514
F-statistic: 347.1 on 3 and 1259 DF,  p-value: < 2.2e-16
```

The `mlreg` function doesn't provide the estimates for the environmental regression coefficients, but the above procedure including each SNP gives the estimates. The estimated coefficient for the intercept is $\hat{\beta}_0 = 185.15348$, the estimated coefficient for age is $\hat{\beta}_{\text{age}} = -1.13179$ and the estimated coefficient for activity level is $\hat{\beta}_{\text{act}} = 2.87759$. Comparing the results with the summary of the reduced sample without including the top SNP, in Listing 4.1, we see that the estimated coefficient for the intercept is larger for the analysis including the genetic covariate, while the other coefficients are almost the same for the two analyses. From Listing 4.2 it is clear that the coefficient of determination is $R^2 = 0.4527$, which is slightly greater than for the analysis in Listing 4.1, which was $R^2 = 0.4457$. Thus, the model including the top SNP is preferable to the model with only age and activity level as covariates.

It is also of interest to compare the summary in Listing 4.2 with the results of analyzing the reduced sample using `mlreg`, as these results should be equivalent. From Table 4.6 we see that the estimated coefficient for the SNP rs10921875 is -3.0011391, with standard deviation of 0.7477329. These values are approximately equal to the estimates in Listing 4.2. Moreover, the χ^2_1-statistic of the top SNP in Table 4.6 is the squared t-statistic of the top SNP in Listing 4.2.

The left panel of Figure 4.12 shows a plot of the studentized residuals from the analysis, which looks ok. The right panel presents a box plot of the SNP rs10921875, and the effects of having 0, 1 or 2 copies of the minor allele, on $\text{VO}_{2_{\text{max}}}$. It is clear that this SNP has a negative effect, in the meaning that it is not favourable for the $\text{VO}_{2_{\text{max}}}$ to have 2 copies of the minor allele.

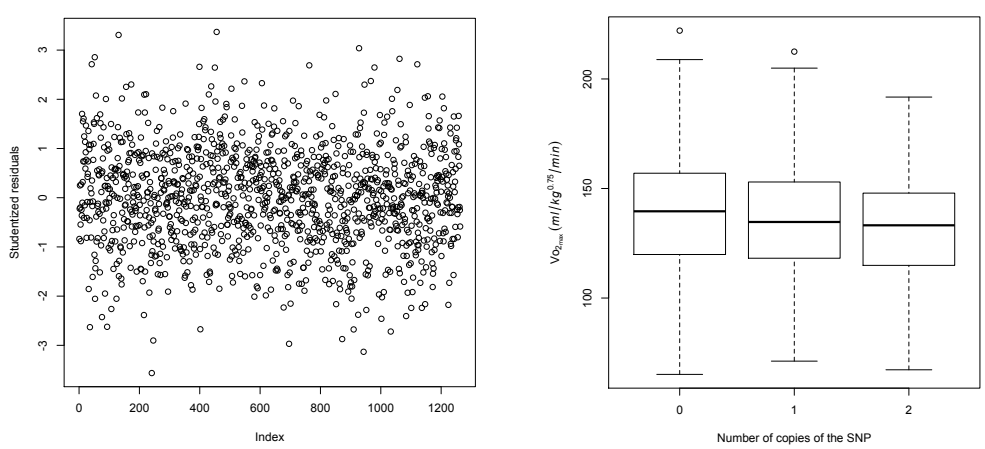

Figure 4.12: Plot of studentized residuals in left panel, and box plot of the SNP rs10921875 in the right panel, showing the effects of having 0, 1 or 2 copies of the minor allele on $\text{VO}_{2_{\text{max}}}$.

Comparison with fitting a linear model to the full sample

Additionally to the analysis of fitting a linear model to the reduced sample, it is of interest to compare the analysis of the reduced sample and the full sample. Thus, we fit a linear model to the full sample without taking into consideration the confounding components. Confounding is the fact that a subset of the sample contain individuals that seem to be related, i.e. not independent. Fitting a linear model to the full sample including genetic covariates using `mlreg`, gives the results in Table 4.9, similarly as for the reduced sample in Table 4.6, sorted by p-values "P1df". The results are the same when sorting by "Pc1df".

The top 10 SNPs of Table 4.9 are not identical to the top 10 SNPs of Table 4.6. The top SNP of the results for the reduced model is rs10921875, while this SNP is number 3 of the results for the full model. Correspondingly, the top SNP of the results for the full model is rs1218592, and this SNP is number 2 of the results for the reduced model. The resulting regression coefficients, χ^2_1-statistics and corresponding p-values are slightly different for the two models.

Table 4.9: Summary for fitting a linear model to the full sample using `mlreg`, sorted by p-values P1df.

	Position	n	effB	se_effB	chi2.1df	P1df	Pc1df
rs1218592	153156109	1443	3.302019	0.8359170	15.60389	0.0000780939	0.0003402991
rs499689	227632355	1446	-2.753119	0.7077048	15.13372	0.0001001582	0.0004185045
rs10921875	187604620	1446	-2.697045	0.7014743	14.78267	0.0001206390	0.0004885243
rs12036597	39302307	1446	-4.541835	1.1885722	14.60197	0.0001327754	0.0005290619
rs155902	187843884	1446	-2.579056	0.7077675	13.27824	0.0002685050	0.0009505222
chr1:160611918	160611918	1446	2.781605	0.7788025	12.75664	0.0003547482	0.0011985854
rs516084	187822397	1446	-2.531598	0.7198733	12.36737	0.0004369021	0.0014256262
chr1:55278035	55278035	1446	-3.368577	1.0247369	10.80608	0.0010116729	0.0028702560
chr1:55278514	55278514	1446	-3.368577	1.0247369	10.80608	0.0010116729	0.0028702560
chr1:55278691	55278691	1446	-3.368577	1.0247369	10.80608	0.0010116729	0.0028702560

Figure 4.13 presents the Manhattan plot for the results from the analysis of the full sample using `mlreg`. As for the Manhattan plot for the analysis of the reduced sample in Figure 4.9, the red line illustrates the local significance level α_{loc}, and the green line intersect the 10th most significant SNP. From the plot it is clear that none of the SNPs are above the red line, which means that there are no SNPs that have a p-value lower than α_{loc}. Thus, none of the SNPs are significant.

The genomic control inflation factor from fitting a linear model using `mlreg` for the full sample, is estimated to be $\hat{\lambda} = 1.205695$. According to the theory of genomic control in Section 3.3, this estimate indicates stratification. The result also shows that the genomic control inflation factor was somewhat smaller for the reduced sample, which was $\hat{\lambda} = 1.16864$, and this is an indication of that the full sample should be analyzed with consideration of confounders. However, the estimate of the inflation factor for the reduced sample is not acceptable.

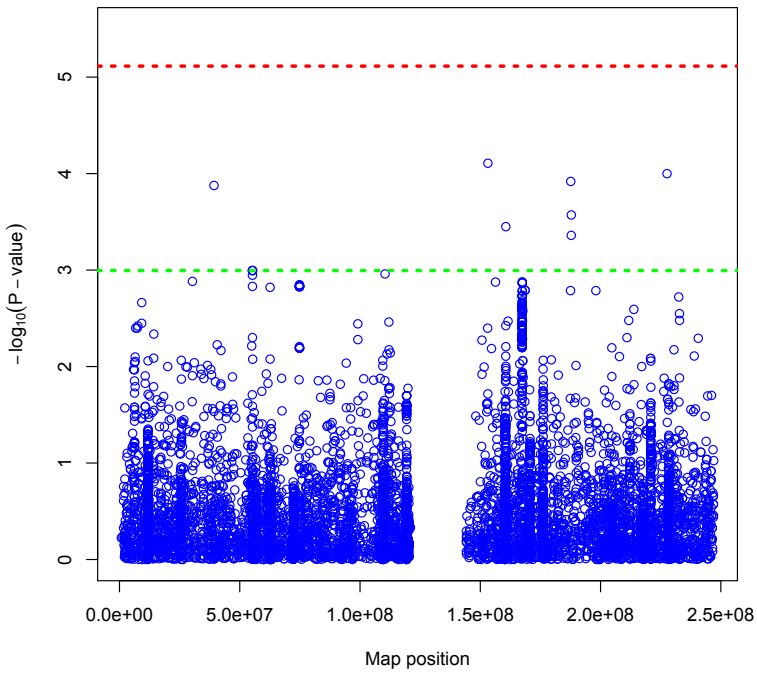

Figure 4.13: Manhattan plot for the single SNP analysis of the full sample using `mlreg`. The red line represents the local significance level α_{loc}, while the green line intersects the 10th most significant SNP.

To further compare the results from the analyses using the `mlreg` function, histograms of the χ_1^2-statistics of the reduced and full sample are presented in Figure 4.14, in the left panel and right panel, respectively. The histograms show that the χ_1^2-statistics of the reduced sample follow the theoretical distribution a little better than the statistics of the full sample.

Figure 4.15 shows Q-Q plots of the observed χ_1^2-statistics versus the expected χ_1^2-statistics for the reduced and full sample in the left and right panel, respectively. The black lines show the theoretical slope without any stratification, while the red lines represent the fitted slope for the data, $\hat{\lambda}$. From the plots we can see that for the reduced sample the red line falls slightly closer to the black line, than for the full sample. However, the results from the Q-Q plots are that either the data sets don't fit the expected distribution perfectly, or the test statistics are correlated (or there are many SNPs with association to the trait). The genomic control inflation factors also indicate that the data sets don't fit the model well. Nonetheless, based on Figure 4.11 and Table 4.8 it is reasonable to think that the deviations from the theoretical distribution are arising from correlated test statistics.

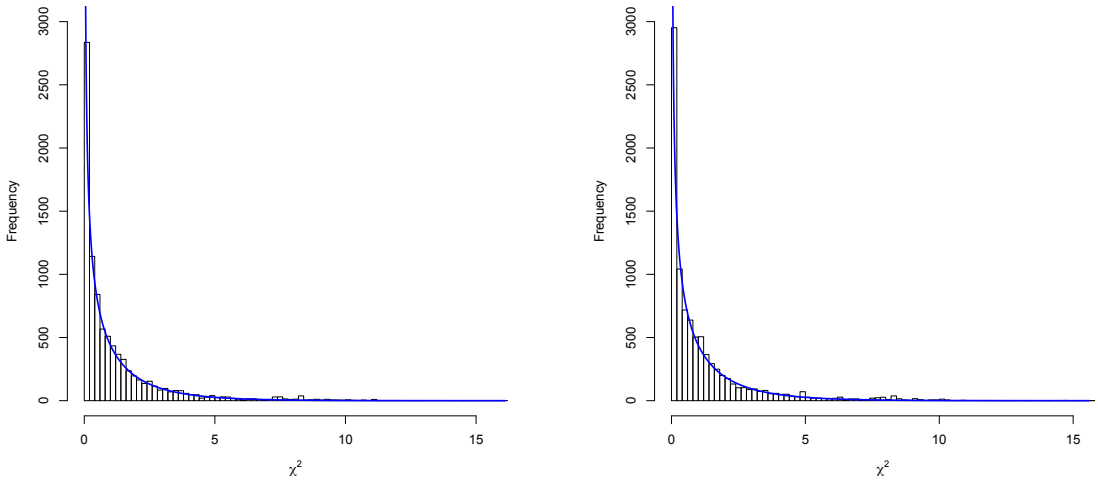

Figure 4.14: Histograms of the χ_1^2-statistics from the analyses of the reduced (left panel) and full sample (right panel) using `mlreg`. The blue lines represent a theoretical χ^2 distribution with 1 degree of freedom.

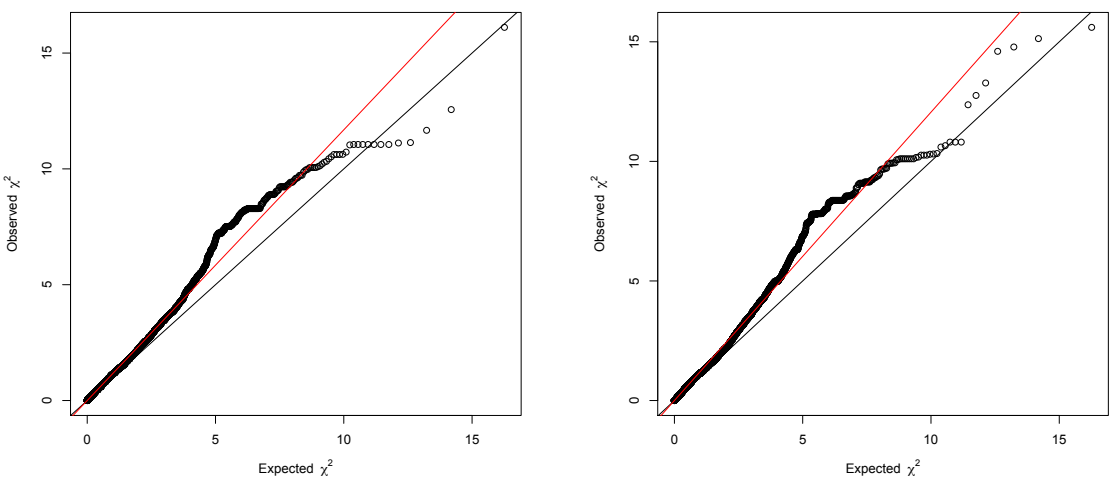

Figure 4.15: The observed χ_1^2-statistics plotted against the expected χ_1^2-statistics, from the analyses of the reduced (left panel) and the full sample (right panel).

4.7 Analysis using linear mixed models

Based on the analyses in Section 4.6, there are no substantial indications of the need to analyze the full sample by taking into consideration the confounding components, except the high values of the genomic inflation factor. However, since we based on estimated kinship coefficients know that the full sample consists of several close relatives, it is necessary to analyze the full sample utilizing the `GRAMMAR` method to account for genetic substructure.

The first step of the analysis is to fit a linear mixed model to the trait $VO_{2_{max}}$ and the environmental covariates age and activity level, using the estimated kinship matrix to give Φ. This is done by first applying the `polygenic` function, which gives the inverse of the covariance matrix and estimates of the residuals of the trait. The environmental residuals are the residuals in which both the effect of the covariates and the estimated polygenic effect are factored out. The environmental residuals are plotted in Figure 4.16, which shows that the residuals are centered around a horizontal line. To total time for the `polygenic` analysis was 8.6 minutes on a computer of the type Intel Core i7-4770 (quad core, 3.4 GHz).

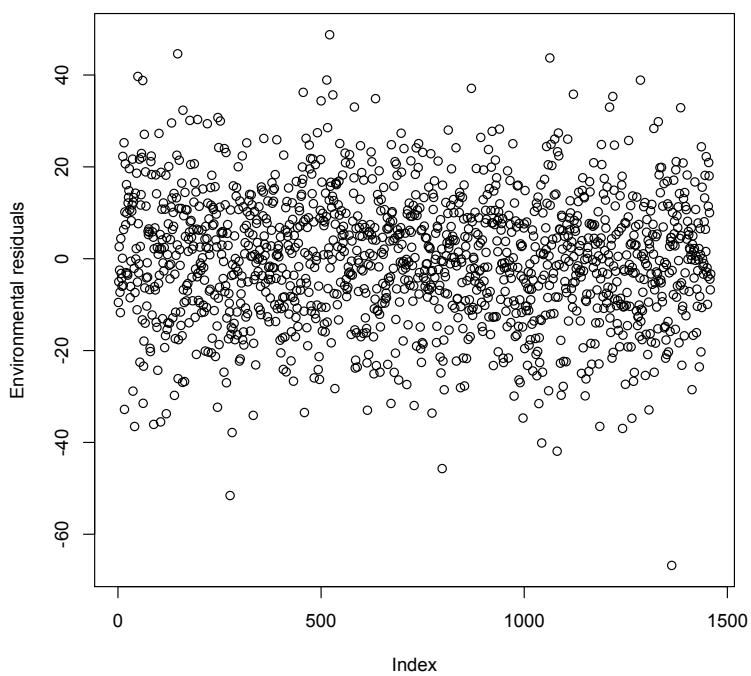

Figure 4.16: Environmental residuals from fitting a linear mixed model to the trait and the environmental covariates, by applying `plygenic`.

The results from fitting a linear mixed model to the trait and the environmental covariates also give the maximum likelihood estimates:

$$\begin{aligned}\hat{\beta}_0 &= 182.7785472\\ \hat{\beta}_{\text{age}} &= -1.1318525\\ \hat{\beta}_{\text{act}} &= 2.8573193.\end{aligned} \tag{4.3}$$

As stated in Section 3.4.4, there are three different approaches of the GRAMMAR method. The summary of the top 10 results sorted by p-values, "P1df", of the ordinary GRAMMAR, GRAMMAR-GC and GRAMMAR-gamma tests are found in Table 4.10, 4.11 and 4.12, respectively. The results are the same when sorting each table by p-values of adjusted χ^2_1-statistics, "Pc1df".

Table 4.10: Summary for fitting a linear mixed model to the full sample using ordinary `GRAMMAR`, sorted by p-values P1df.

	Position	n	effB	se_effB	chi2.1df	P1df	Pc1df
rs10921875	187604620	1446	-2.004059	0.5137388	15.21725	0.0000958243	0.0000958242
rs155902	187843884	1446	-1.936447	0.5180952	13.96986	0.0001857653	0.0001857653
rs1218592	153156109	1443	2.198052	0.6124970	12.87858	0.0003323635	0.0003323635
rs516084	187822397	1446	-1.887392	0.5268318	12.83454	0.0003402784	0.0003402784
rs499689	227632355	1446	-1.845267	0.5182980	12.67533	0.0003705110	0.0003705110
rs12036597	39302307	1446	-3.075682	0.8695578	12.51083	0.0004045999	0.0004045999
rs2274165	156327667	1446	2.994822	0.9095253	10.84208	0.0009921942	0.0009921942
rs560145	232500695	1445	2.751356	0.8708699	9.98130	0.0015813784	0.0015813784
rs1935660	187509749	1446	-1.627586	0.5196101	9.81144	0.0017342933	0.0017342933
chr1:160611918	160611918	1446	1.745662	0.5690435	9.41086	0.0021570445	0.0021570445

Table 4.11: Summary for fitting a linear mixed model to the full sample using `GRAMMAR`-GC, sorted by p-values P1df.

	Position	n	effB	se_effB	chi2.1df	P1df	Pc1df
rs10921875	187604620	1446	-2.004059	0.5137388	15.21725	0.0000958243	0.0000509804
rs155902	187843884	1446	-1.936447	0.5180952	13.96986	0.0001857653	0.0001038149
rs1218592	153156109	1443	2.198052	0.6124970	12.87858	0.0003323635	0.0001939168
rs516084	187822397	1446	-1.887392	0.5268318	12.83454	0.0003402784	0.0001988803
rs499689	227632355	1446	-1.845267	0.5182980	12.67533	0.0003705110	0.0002179159
rs12036597	39302307	1446	-3.075682	0.8695578	12.51083	0.0004045999	0.0002395166
rs2274165	156327667	1446	2.994822	0.9095253	10.84208	0.0009921942	0.0006274166
rs560145	232500695	1445	2.751356	0.8708699	9.98130	0.0015813784	0.0010346520
rs1935660	187509749	1446	-1.627586	0.5196101	9.81144	0.0017342933	0.0011423613
chr1:160611918	160611918	1446	1.745662	0.5690435	9.41086	0.0021570445	0.0014435612

Table 4.12: Summary for fitting a linear mixed model to the full sample using `GRAMMAR`-gamma, sorted by p-values P1df.

	Position	n	effB	se_effB	chi2.1df	P1df	Pc1df
rs10921875	187604620	1446	-3.009576	0.7317368	16.91612	0.0000390684	0.0000509804
rs155902	187843884	1446	-2.908040	0.7379418	15.52947	0.0000812291	0.0001038149
rs1218592	153156109	1443	3.300904	0.8724016	14.31637	0.0001545157	0.0001939168
rs516084	187822397	1446	-2.834373	0.7503856	14.26741	0.0001585872	0.0001988803
rs499689	227632355	1446	-2.771112	0.7382306	14.09042	0.0001742285	0.0002179159
rs12036597	39302307	1446	-4.618876	1.2385426	13.90755	0.0001920252	0.0002395166
rs2274165	156327667	1446	4.497445	1.2954698	12.05250	0.0005172300	0.0006274166
rs560145	232500695	1445	4.131823	1.2404115	11.09563	0.0008653154	0.0010346520
rs1935660	187509749	1446	-2.444212	0.7400995	10.90681	0.0009581163	0.0011423613
chr1:160611918	160611918	1446	2.621530	0.8105093	10.46150	0.0012188852	0.0014435612

From the tables it is clear that all the methods give the same SNPs as the most significant ones. However, the GRAMMAR-gamma method estimates larger regression coefficients for the SNP covariates than the other methods, and thus also different χ^2_1-statistics and p-values. We will continue using the GRAMMAR-gamma method, because this method gives the most correct results, see Section 3.4.4.

Figure 4.17 displays a Manhattan plot for the single SNP analysis of the full sample using `GRAMMAR`-gamma. The red line illustrates the local significance level α_{loc}, while the green line intersects the 10th most significant SNP. The plot shows that there are no SNPs above the red line, thus no SNPs have a p-value lower than α_{loc}.

The estimated genomic control inflation factor of fitting a linear mixed model to the full model is $\hat{\lambda} = 0.9919442$. The low value of λ indicates that the model correctly accounts for population structure and cryptic relatedness.

A histogram of the χ^2_1-statistics of the analysis is plotted in the left panel of Figure 4.18, and it shows that the statistics follow the theoretical distribution, represented by the blue curve, very well.

Moreover, the right panel of Figure 4.18 is a Q-Q plot of the observed χ^2_1-statistics versus the expected χ^2_1-statistics. The red line represents the fitted slope of the data, while the black line shows the theoretical slope $\hat{\lambda}$. It is clear from the plot that the data follows the theoretical slope quite well, which is an indication that the linear mixed model fits the full sample.

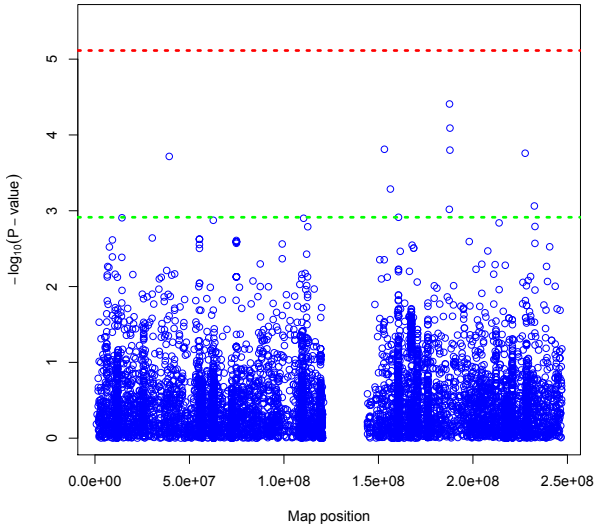

Figure 4.17: Manhattan plot for the single SNP analysis of the full sample using `GRAMMAR`-gamma. The red line represents the local significance level α_{loc}, while the green line intersects the 10th most significant SNP.

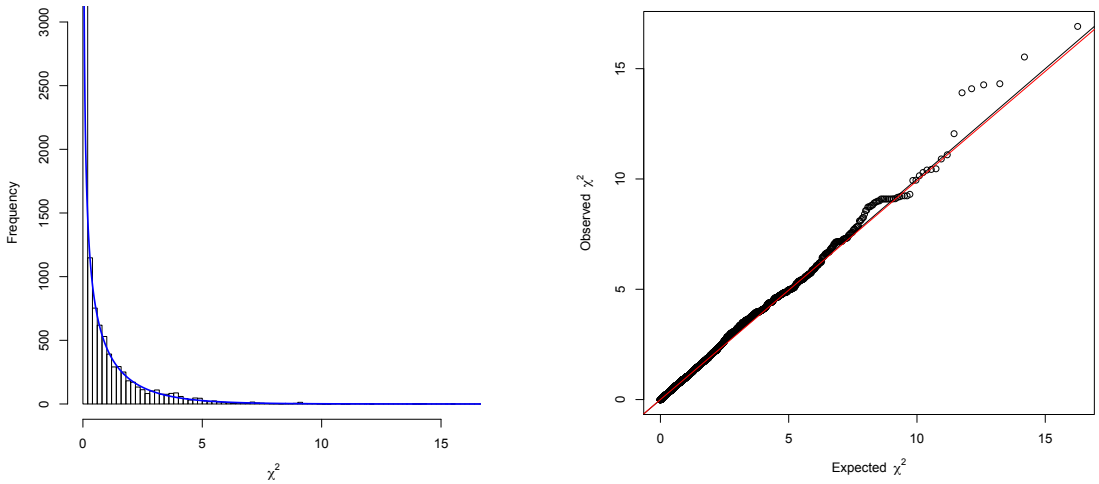

Figure 4.18: The left panel is a histogram of the χ^2_1-statistics from the analysis of the full sample using `GRAMMAR`-gamma. The blue line represents a theoretical χ^2-distribution with 1 degree of freedom. The right panel shows the observed χ^2_1-statistics from the analysis of the full sample using `Grammar`-gamma plotted against the expected χ^2_1-statistics.

4.8 Comparison of the analyses of the reduced and full sample

Analyzing samples of unrelated individuals using a linear model, as done for the reduced sample in this thesis, is the most commonly and easiest approach to analyze GWA data. However, in order to utilize the data from all the individuals in the study, and probably attain higher statistical power, one can analyze the full sample using a linear mixed model approach. This section will consider one of the main purposes of this thesis, which is to compare the results from the two methods to see if they agree.

Table 4.13: Results from `mlreg` analysis of the reduced and the full sample, and `GRAMMAR`-gamma analysis of the full sample. The results show the top 10 SNPs to be associated with $VO_{2_{max}}$ sorted by p-values ("P1df"), and the estimated regression coefficients of each SNP ("effB").

`mlreg` reduced sample			`mlreg` full sample		
SNP	effB	P1df	SNP	effB	P1df
rs10921875	-3.001139	0.0000597867	rs1218592	3.302019	0.0000780939
rs1218592	3.158232	0.0003941767	rs499689	-2.753119	0.0001001582
rs155902	-2.567510	0.0006368039	rs10921875	-2.697045	0.0001206390
chr1:74819417	-4.015202	0.0008464761	rs12036597	-4.541835	0.0001327754
chr1:74823928	-4.015720	0.0008552151	rs155902	-2.579056	0.0002685050
chr1:74812647	-3.994218	0.0008853159	chr1:160611918	2.781605	0.0003547482
chr1:74813852	-3.994218	0.0008853159	rs516084	-2.531598	0.0004369021
chr1:74832901	-3.994218	0.0008853159	chr1:55278035	-3.368577	0.0010116729
chr1:74840649	-3.994218	0.0008853159	chr1:55278514	-3.368577	0.0010116729
chr1:74841681	-3.994218	0.0008853159	chr1:55278691	-3.368577	0.0010116729

`GRAMMAR`-gamma full sample		
SNP	effB	P1df
rs10921875	-3.009576	0.0000390684
rs155902	-2.908040	0.0000812291
rs1218592	3.300904	0.0001545157
rs516084	-2.834373	0.0001585872
rs499689	-2.771112	0.0001742285
rs12036597	-4.618876	0.0001920252
rs2274165	4.497445	0.0005172300
rs560145	4.131823	0.0008653154
rs1935660	-2.444212	0.0009581163
chr1:160611918	2.621530	0.0012188852

Table 4.13 shows the results from the `mlreg` analyses of the reduced and the full sample, and the `GRAMMAR`-gamma analysis of the full sample. It is clear that the analysis of the reduced sample using `mlreg` gives approximately equal top SNPs as the analysis of the full sample using `GRAMMAR`-gamma, only that the SNPs from Table 4.8 that are in high linkage disequilibrium are not present in the table of the top 10 most significant SNPs for the full sample using `GRAMMAR`-gamma. However, these SNPs are also identical for the full sample.

The SNP that gives the most significant test result in both methods is rs109218-75, with an estimated regression coefficient of -3.001139 for the reduced sample using `mlreg` and -3.009576 for the full sample using `GRAMMAR`-gamma. Correspondingly, this SNP is listed in the third row of Table 4.13 for the analysis of the full sample using `mlreg`, with an estimated regression coefficient of -2.697045.

The estimates of the regression coefficient for each SNP, $\hat{\beta}_{g_{(k)}}$, are approximately equal for the different methods. When comparing the estimates of Table 4.13 to the estimates in Table 4.10 and 4.11, we see that the estimates from original `GRAMMAR` and `GRAMMAR`-gc for the full sample are different from the estimates from `mlreg` for reduced and full sample and `GRAMMAR`-gamma of full sample.

The genomic inflation factor was estimated for all the different models, using the *median* approach. The analysis of the reduced sample using a linear model resulted in an estimated inflation factor $\hat{\lambda} = 1.16864$, while the corresponding factor for the analysis of the full sample using a linear model was $\hat{\lambda} = 1.205695$. These values indicates presence of stratification, as stated in Section 3.3. The factor is considerably reduced when fitting a linear mixed model to the full sample, where $\hat{\lambda} = 0.9919442$. This value is below 1.05, and thus considered benign according to Price et al. (2010). This shows that the linear mixed model approach is capable of correcting for population structure and cryptic relatedness.

As presented in Section 4.6.2 the `mlreg`-function does not provide the user with estimates of the regression coefficients $\hat{\beta}_0$, $\hat{\beta}_{\text{age}}$ and $\hat{\beta}_{\text{act}}$, and the same is the case for all the `GRAMMAR` approaches. In order to compare the effect of each SNP to the environmental covariates, it is favourable to have these estimates. For the case of analyzing the reduced sample using `mlreg`, it is possible to fit a linear model to the reduced sample using `lm` including a SNP of choice, as seen in Section 4.6.2, to get the estimates of the environmental regression coefficients. However, performing the same step for the full sample using a linear mixed model is complex and time consuming because of the large number of individuals and SNPs, and the corresponding size of the covariance matrix. The `polygenic` step prior to `GRAMMAR` gives the estimates $\hat{\beta}_0$, $\hat{\beta}_{\text{age}}$ and $\hat{\beta}_{\text{act}}$ as seen in Equation (4.3), but these are the estimates from fitting a linear mixed model without a genetic covariate. Comparing these results to the estimates from fitting a linear model to the reduced sample without including any genetic covariates in Listing 4.1 and to the estimates from fitting a linear model to the reduced sample including SNP

rs10921875 in Listing 4.2, we see that $\hat{\beta}_{\text{age}}$ and $\hat{\beta}_{\text{act}}$ are approximately equal for all analyses. The estimates of the intercept, $\hat{\beta}_0$, are similar for the `polygenic` analysis and the `lm` analysis without a genetic covariate, but when including a genetic covariate the estimate of the intercept increases.

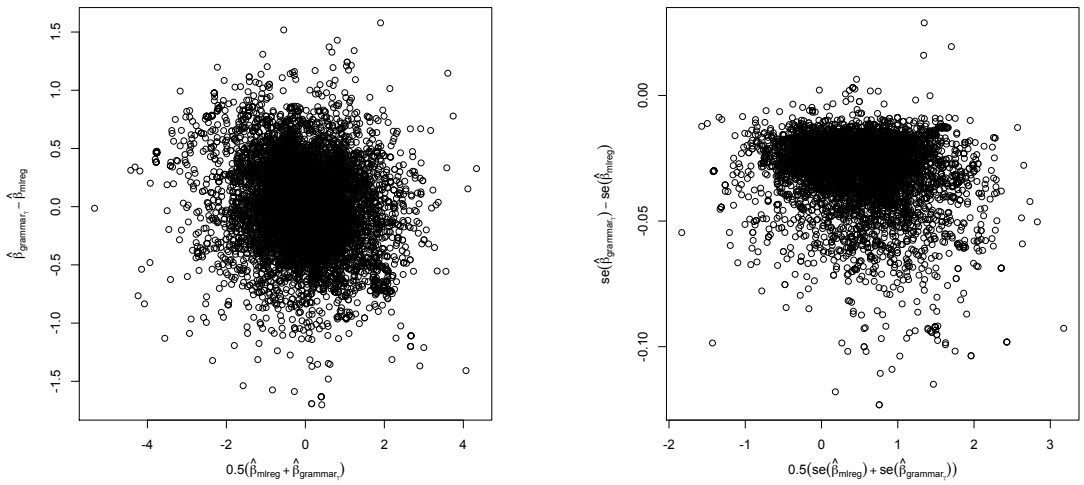

Figure 4.19: Mean-difference plot of $\hat{\beta}_{g_{(k)}}$ estimated from `mlreg` analysis of the reduced sample and `GRAMMAR`-gamma analysis of the full sample in the left panel, and mean-difference plot of standard error of $\hat{\beta}_{g_{(k)}}$ in the right panel, for $k = 1, \ldots, 9069$.

Figure 4.19 shows mean-difference plots for the results of the `mlreg` and `GRAMMAR`-gamma analyses. The left panel shows a plot of $\hat{\beta}_{g_{(k),\text{grammar}_\gamma}} - \hat{\beta}_{g_{(k),\text{mlreg}}}$ plotted against $\frac{1}{2}(\hat{\beta}_{g_{(k),\text{mlreg}}} + \hat{\beta}_{g_{(k),\text{grammar}_\gamma}})$ for all SNPs $k = 1, \ldots, 9069$, and it shows that the points are centered around zero. The right panel of the figure shows the mean-difference plot of the estimated standard errors from the two methods for all SNPs. It illustrates that the standard error of $\hat{\beta}_{g_{(k),\text{mlreg}}}$ is larger than the standard error of $\hat{\beta}_{g_{(k),\text{grammar}_\gamma}}$ for almost all SNPs.

4.9 Genetic interpretation of the results

The main goal of this thesis is to evaluate the different statistical methods for GWA analysis, but the results from the analysis are as well interesting. The 3 most significant SNPs from Table 4.13 are rs10921875, rs155902 and rs1218592. Even though none of the SNPs have a statistically significant association to the $\text{VO}_{2_{\text{max}}}$ trait based on data on chromosome 1, it is of relevance to investigate the genetic interpretation of the SNPs.

Global MAFs calculated from the 1000 Genomes Project[2] of the SNPs rs109218-75, rs155902 and rs1218592 are 0.3920, 0.4916, and 0.0960, respectively[3]. Based on data from the HUNT study, the MAFs of the SNPs rs10921875, rs155902 and rs1218592 are 0.4208362, 0.4064428 and 0.2445055, respectively. The reason for the difference in MAF of rs1218592 for the 1000 Genomes project and for the HUNT study, may be that the HUNT study only contains data from men from a single population, while the 1000 Genomes project had both male and female participants from different populations all over the world.

The SNPs rs10921875 and rs155902 are located on the same gene, a gene with unknown function for the time being. When finding a SNP associated with a specific trait, it is possible that the SNP can affect the gene it is located on, and in addition genes nearby. It is therefore also necessary to check the surroundings of the SNPs.

The closest gene to the SNP rs10921875 with known function is BRINP3. Polymorphisms that increase expression of this gene have been shown to increase vascular inflammation, and an association of this gene with myocardial infarction has been demonstrated[4]. This is relevant for the $VO_{2_{max}}$ phenotype, since $VO_{2_{max}}$ is an important marker of risk for cardiovascular diseases.

Moreover, the SNP rs1218592 is also located in a gene with unknown function. A gene nearby is KCNN3, that is among other things associated with atrial fibrillation[5]. This is relevant for $VO_{2_{max}}$, as it is shown that athletes have an increased occurrence of atrial fibrillation. The gene is also generally associated to cardiovascular diseases, as the gene BRINP3.

It is standard of GWA studies to perform the first analysis in a discovery cohort, which is then followed by an independent validation cohort including only the most significant SNPs. In a GWA study of $VO_{2_{max}}$ by Bye et al. (2016) similar to the analysis of the reduced sample, the SNP rs1218592 was a candidate SNP from the exploration cohort that failed to be replicated in the validation cohort.

[2] 1000genomes.org (2016). 1000 Genomes | A Deep Catalog of Human Genetic Variation. Available at: http://www.1000genomes.org [Accessed 13 Jun. 2016].

[3] National Center for Biotechnology Information, U.S. National Library of Medicine (2016). Available at: https://www.ncbi.nlm.nih.gov/snp [Accessed 13 Jun. 2016].

[4] National Center for Biotechnology Information, U.S. National Library of Medicine (2016). BRINP3 BMP/retinoic acid inducible neural specific 3. Available at: https://www.ncbi.nlm.nih.gov/gene/339479 [Accessed 13 Jun. 2016].

[5] National Center for Biotechnology Information, U.S. National Library of Medicine (2016). KCNN3 potassium calcium-activated channel subfamily N member 3. Available at: https://www.ncbi.nlm.nih.gov/gene/3782 [Accessed 13 Jun. 2016].

Chapter 5

Discussion and conclusions

5.1 Statistical issues

When performing a GWA analysis there are several confounding factors that can cause correlations between the study participants. In this Master's thesis we have studied statistical methods of analyzing GWA data, and performed a GWA analysis for the SNPs on chromosome 1 of the HUNT $VO_{2_{max}}$ study by applying different statistical models.

The comparison of the models in Section 4.8 showed that the results from using the `GenABEL` function `mlreg` to analyze the reduced sample are similar to the results from the `GRAMMAR`-gamma analysis of the full sample. We would expect the results to be more equal, as the reduced sample are assumed to the a sample of independent individuals. The reduced sample consists of individuals with a pairwise estimated kinship coefficient below 0.1, and when analyzing the sample without a genetic covariate it seemed like the model fitted the data very well, based on examining residual plots. Nonetheless, the estimated genomic control inflation factor from the analysis including a genetic covariate was $\hat{\lambda} = 1.16864$, which indicates the presence of population and family stratification in the data. The genomic control inflation factor is based on results from performing hypothesis tests, so examining the residuals versus the inflation factor is two different approaches. While the residuals from the analysis show a good model fit to the reduced sample, the genomic control inflation factor indicates that maybe the reduced sample is not reduced enough, and we should include only individuals with an estimated kinship coefficient below 0.05 in the reduced sample.

Fitting a linear regression model to the full sample including each genetic covariate, one at a time, did not give very dissimilar results to the reduced sample, rather than that the estimated genomic control inflation factor for the full sample, $\hat{\lambda} = 1.205695$, was higher than for the reduced sample. However, since the estimated kinship matrix for the full sample reveals close relatives in the sample,

the preferred method for analyzing GWA data is fitting a linear mixed model to the full sample, with a scaled estimated kinship matrix as covariance matrix. This ensures that the model captures the correlation and corresponding population and family structure of the data. The estimated genomic control inflation factor from the analysis of the full sample using `GRAMMAR`-gamma was $\hat{\lambda} = 0.9919442$, which is considerably lower than for the other methods.

Genomic control is used to detect confounding components in the data, but we don't apply the method of deflation of test statistics, because it is more important to take the relatedness between the participants into account than to adjust the test statistics by a factor. As presented in Section 3.3, Aulchenko (2014) suggested to use 95% of the least significant SNPs to estimated the inflation factor. If we apply this restriction, the estimated inflation factors for `mlreg` analyses of the reduced sample and full sample, and `GRAMMAR`-gamma for the full sample, are $\hat{\lambda} = 1.03168$, $\hat{\lambda} = 1.077178$ and $\hat{\lambda} = 0.8903894$, respectively. The estimated inflation factors from the `mlreg` analyses are much closer to 1 when using this method, while the estimate from the `GRAMMAR`-gamma analysis is very low. However, this restricted method to estimate the genomic control inflation factor is not applied in this analysis, as none of the SNPs show significant associations to the phenotype.

The function `mlreg` use χ^2_1-statistics to compute p-values for each $\hat{\beta}_{g_{(k)}}$. However, as seen in Section 4.6.2, the test statistics of testing each SNP's association in the linear model to the trait are actually $\mathcal{F}_{1,n-p-1}$-distributed, and only asymptotically χ^2_1-distributed.

Moreover, the amount of data in GWA studies is in general enormous, and even though we have decided on a statistical model to analyze the data, we have to apply numerical optimization methods and complex programming tricks to be able to perform the analysis. Using generic functions to perform linear mixed model regression, like `lme4` and `nlme`, is impossible for this enormous amount of data, because of the many tests that have to be performed and the inversion of the covariance matrix. There is also a problem of storage, as there is not enough space to save everything, so only the most important and necessary information is saved. As presented in Section 4.8, the functions in the `GenABEL`-package give priority to the genetic components of the model, and do not give access to the estimates of the environmental components. Another difficulty, and time consuming part of the analysis, with using a package like the `GenABEL`-package, is to process the data into a format that is manageable for the functions of the package, and to understand how to utilize the different functions.

In Section 3.5 we presented a comparison of principal components regression and linear mixed models. The method of including principal components in a linear regression model is not studied in this thesis, because we think applying a linear mixed model is the best method to analyze GWA data with consideration of genetic confounding.

5.2 Discussion of the genetic results

The results from Chapter 4 show that there are no SNPs with statistically significant associations to the phenotype, when using the local significance level $\alpha_{\text{loc}} = 7.694294 \cdot 10^{-6}$ to control the FWER at level 0.05. In spite of this, as presented in Section 4.9, the most significant SNPs from all the methods, rs10921875, rs155902 and rs1218592, are positioned on genes close to other genes that are highly relevant to the $\text{VO}_{2_{\text{max}}}$ trait. The fact that none of the SNPs show statistically significant association to the phenotype is believed to be partially because of the small sample size.

We applied the same local significance level for both the analysis using a linear model and the analysis using a linear mixed model, because we assumed that the distribution of the test statistics is similar for the reduced and the full sample.

It is unknown how many individuals that should be included in a study to obtain statistically significant associations between SNPs and the $\text{VO}_{2_{\text{max}}}$ phenotype, and if it is numerically possible to analyze samples of ten thousands of individuals using linear mixed models. Maybe the best approach to analyze samples of this size is to use linear regression models of a strictly reduced sample.

The results presented in Table 4.8 in Section 4.6.1 showed that there are some highly correlated SNPs. If we should have studied the HUNT $\text{VO}_{2_{\text{max}}}$ data set in more detail, including SNPs from all chromosomes, we would have removed the highly correlated SNPs.

List of references

Amin, N., van Duijn, C. M. and Aulchenko, Y. S. (2007), 'A genomic background based method for association analysis in related individuals', *PLoS ONE* **2**(12), 1–7.

Astle, W. and Balding, D. J. (2009), 'Population structure and cryptic relatedness in genetic association studies', *Statistical Science* **24**(4), 451–471.

Aulchenko, Y. (2014), *GenABEL tutorial.*
URL: *http://www.genabel.org/sites/default/files/pdfs/GenABEL-tutorial.pdf*

Aulchenko, Y. S., de Koning, D.-J. and Haley, C. (2007), 'Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis', *Genetics* **177**(1), 577–585.

Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1), 289–300.

Bingham, N. H. and Fry, J. M. (2010), *Regression, Linear Moedels in Statistics*, Springer-Verlag London.

Bjørnland, T. (2014), Statistical methods for genetic association studies under the extreme phenotype sampling design: Modelling the effects of both common and rare genetic variants, Master's thesis, NTNU.

Bouchard, C., An, P., Rice, T., Skinner, J. S., Wilmore, J. H., Gagnon, J., Pérusse, L., Leon, A. S. and Rao, D. C. (1999), 'Familial aggregation of VO2max response to exercise training: results from the HERITAGE family study', *Journal of Applied Physiology* **87**(3), 1003–1008.

Bye, A. (2008), Gene expression profiling of inherited and acquired maximal oxygen uptake, PhD thesis, NTNU.

Bye, A., Ryeng, E., di Silva, G., Moreira, J. B. N., Stensvold, D. and Wisløff, U. (2016), 'Novel genetic components of aerobic fitness - the HUNT-study (in preparation)', *The Journal of Clinical Investigation* .

Casella, G. and Berger, R. L. (2002), *Statistical Inference*, Thomson Learning.

Chen, W.-M. and Abecasis, G. R. (2007), 'Family-based association tests for genomewide association scans', *American Journal of Human Genetics* **81**(5), 913–926.

Dadd, T., Weale, M. E. and Lewis, C. M. (2009), 'A critical evaluation of genomic control methods for genetic association studies', *Genetic Epidemiology* **33**(4), 290–298.

E. A. Thompson, R. G. S. (1990), 'Pedigree analysis for quantitative traits: Variance components without matrix inversion', *Biometrics* **46**(2), 399–413.

Eu-ahsunthornwattana, J., Miller, E. N., Fakiola, M., Jeronimo, S. M. B., Blackwell, J. M., Cordell, H. J. and Consortium, W. T. C. C. (2014), 'Comparison of methods to account for relatedness in genome-wide association studies with family-based data', *PLoS Genet* **10**(7).

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013), *Regression*, Springer Berlin Heidelberg.

Fletcher, H. and Hickey, I. (2013), *Genetics*, BIOS instant notes, Garland Science.

Goeman, J. J. and Solari, A. (2014), 'Multiple hypothesis testing in genomics', *Statistics in Medicine* **33**(11), 1946–1978.

Halle, K. K., Bakke, Ø., Djurovic, S., Bye, A., Ryeng, E., Wisløff, U., Andreassen, O. A. and Langaas, M. (2016), 'Efficient and powerful familywise error control in genome-wide association studies using generalized linear models', *ArXiv* .

Halliburton, R. (2004), *Introduction to Population Genetics*, Pearson Prentice Hall.

Hoffman, G. E. (2013), 'Correcting for population structure and kinship using the linear mixed model: Theory and extensions', *PLoS One* **8**(10).

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning. with Applications in R*, Springer Texts in Statistics, Springer New York.

Lange, K. (2003), *Mathematical and Statistical Methods for Genetic Analysis*, Statistics for Biology and Health, Springer New York.

Lehmann, E. and Romano, J. P. (2005), *Testing Statistical Hypotheses*, Springer.

Lippert, C. (2013), Linear mixed models for genome-wide association studies, PhD thesis, University of Tübingen Germany.

Østgård, E. T. (2011), Statistical modeling and analysis of repeated measures, using the linear mixed effects model, Master's thesis, NTNU.

Panagiotou, O. A., Evangelou, E. and Ioannidis, J. P. A. (2010), 'Genome-wide significant associations for variants with minor allele frequency of 5overview: A huge review', *American Journal of Epidemiology* **172**(8), 869–889.

Patterson, N., Price, A. L. and Reich, D. (2006), 'Population structure and eigenanalysis', *PLoS Genet* **2**(12), 1–20.

Price, A. L., Zaitlen, N. A., Reich, D. and Patterson, N. (2010), 'New approaches to population stratification in genome-wide association studies', *Nat Rev Genet* **11**(7).

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Ripley, B. (1996), *Pattern recognition and neural networks*, Cambridge University Press.

Speed, D. and Balding, D. J. (2015), 'Relatedness in the post-genomic era: is it still useful?', *Nature Reviews Genetics* **16**(1), 33–44.

Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. and Aulchenko, Y. S. (2012), 'Rapid variance components-based method for whole-genome association analysis', *Nat Genet* **44**(10), 1166–1170.

Thompson, E. A. (1986), *Pedigree Analysis in Human Genetics*, The Johns Hopkins University Press.

Thompson, E. A. (2000), 'Statistical inference from genetic data on pedigrees', *NSF-CBMS Regional Conference Series in Probability and Statistics* **6**, 1–169.

Thornton, T. and McPeek, M. S. (2010), 'Roadtrips: Case-control association testing with partially or completely unknown population and pedigree structure', *American Journal of Human Genetics* **86**(2), 172–184.

VanLiere, J. M. and Rosenberg, N. A. (2008), 'Mathematical properties of the r^2 measure of linkage disequilibrium', *Theoretical Population Biology* **74**(1), 130 – 137.

Walpole, R. E., Myers, R. H., Myers, S. L. and Ye, K. (2012), *Proability and Statistics for Engineers and Scientists*, 9. edn, Pearson.

Appendix A

R-code for use of the GenABEL package

In order to use the functions in the GenABEL-package we had to convert PLINK data into GenABEL-package format. First, we converted the PLINK *.ped* files into PLINK *.tped* files, and PLINK *.fam* files into PLINK *.tfam* files. Then we used the GenABEL function convert.snp.tped to convert genotypic data from *.tped* format to GenABEL-package binary *.raw* format:

```
convert.snp.tped(paste(filelocation,"filname.tped",sep=""),
   paste(filelocation,"filname.tfam",sep=""),"genotype_data.
   raw")
genotype_data=paste(filelocation,"genotype_data.raw", sep="")
```

The covariates are in *.txt* format, and we want to have this in *.dat* format:

```
covariates=read.table(paste(filelocation,"filname.txt",
   sep=""),sep="",header=TRUE,dec=",")
write.table(covariates,"phenotype_data.dat",col.names=TRUE,
   row.names=FALSE)
phenotype_data=paste(filelocation,"phenotype_data.dat",sep=""
   )
```

The data can then be loaded into the GenABEL-package. The function load.gwaa.data loads data (genotypes and phenotypes) from files to a gwaa.data object:

```
data=load.gwaa.data(phenotype_data,genotype_data,makemap=F)
```

The gwaa.data class contains objects holding all GWA analysis data - phenotypes and genotypes. The class uses slots to give genotype and phenotype data, so in order to get information about the genotype data we use: data@gtdata, and correspondingly for the phenotype data: data@phdata. data@phdata is a data frame containing information about the phenotype, and for the HUNT data set it gives information about family ID, ID names of the individuals, age, sex, activity level and $VO_{2_{max}}$.

```
data@phdata$age
data@phdata$activity_level
data@phdata$vo2max
```

The object data@gtdata is an object of the snp.data class. This class gives information about number and ID names of SNPs and individuals, chromosome and strand data of the SNPs.

```
data@gtdata@nids
data@gtdata@nsnps
table(chromosome(data))
```

Quality check of the data is performed by using the function check.marker:

```
qc=check.marker(data, call=0.90, perid.call=0.90, p.level=-1,
    fdrate = 0.2, maf=0.05)
summary(qc)
data_qc=data[qc$idok, qc$snpok]
nids(data_qc)
nsnps(data_qc)
summary(data_qc@phdata)
descriptives.marker(data_qc)
descriptives.trait(data_qc)
```

The kinship matrix is estimated using the function ibs;

```
kinshipmatrix=ibs(data_qc[,autosomal(data_qc)], w="freq")
```

R-code for fitting a linear model to the reduced sample:

```
age=data_reduced@phdata$age
act=data_reduced@phdata$act
vo2max=data_reduced@phdata$vo2max

lm_nogen=lm(vo2max~age+act)
summary(lm_nogen)
```

R-code for extracting chromosome 1 data for the reduced sample, and fitting a linear model including genetic covariates to the data:

```
chr1data_reduced=data_reduced@gtdata@snpnames[data_
   keep@gtdata@chromosome=="1"]
data_reduced_chr1 <- data_reduced[,chr1data_reduced]

age=data_reduced_chr1@phdata$age
act=data_reduced_chr1@phdata$act
vo2max=data_reduced_chr1@phdata$vo2max

lm_reduced_allsnps=mlreg(vo2max~age+act, data_reduced_chr1,
   trait="gaussian")
summary(lm_reduced_allsnps)
estlambda(lm_reduced_allsnps[,"P1df"], plot=TRUE, method="
   median")
plot(lm_reduced_allsnps)
```

The data for the top SNP is extracted, and a linear model is fitted to the data:

```
top_SNP=as.double.snp.data(data_reduced_chr1@gtdata[,"
   rs10921875"])
lm_top_SNP=lm(vo2max~age+act+top_SNP)
summary(lm_top_SNP)
```

R-code for extracting chromosome 1 data for the full sample, and fitting a linear model including genetic covariates to the data:

```
chr1data_full=data_qc@gtdata@snpnames[data_
   qc@gtdata@chromosome=="1"]
data_full_chr1=data_qc[,chr1data_full]

age=data_full_chr1@phdata$age
```

```
act=data_full_chr1@phdata$act
vo2max=data_full_chr1@phdata$vo2max

lm_full_allsnps=mlreg(vo2max~age+act, data_full_chr1, trait="
   gaussian")
summary(lm_full_allsnps)
estlambda(lm_full_allsnps[,"P1df"], plot=TRUE, method="median
   ")
plot(lm_full_allsnps)
```

R-code for fitting a linear mixed model including genetic covariates to the chromosome 1 data of the full sample, using the different GRAMMAR methods:

```
poly <- polygenic(formula=vo2max~age+act, kinship.matrix=data
   _kinship, data=data_full_chr1, llfun="polylik")

grammar_gc=grammar(poly, data_full_chr1, method="gc")
summary(grammar_gc)
estlambda(grammar_gc[,"P1df"], plot=TRUE, method="median")

grammar_raw=grammar(poly, data_full_chr1, method="raw")
summary(grammar_raw)
estlambda(grammar_raw[,"P1df"], plot=TRUE, method="median")

grammar_gamma=grammar(poly, data_full_chr1, method="gamma")
summary(grammar_gamma)
estlambda(grammar_gamma[,"P1df"], plot=TRUE, method="median")

plot(grammar_gamma)
```

R-code for calculating the MAF of a SNP:

```
find_maf=function(snpname){
  genotypes=as.double.snp.data(data_qc@gtdata[,snpname])
  maf=sum(na.omit(genotypes))/(2*length(na.omit(genotypes)))
}
```

Appendix B

R-output

B.1 Quality control

```
> qc = check.marker(data, call=0.90, perid.call = 0.90, p.
   level = -1, fdrate = 0.2, maf = 0.05)
Excluding people/markers with extremely low call rate...
196725 markers and 1488 people in total
13 people excluded because of call rate < 0.1
692 markers excluded because of call rate < 0.1
Passed: 196033 markers and 1475 people

RUN 1
196033 markers and 1475 people in total
87954 (44.86694%) markers excluded as having low (<5%) minor
   allele frequency
4297 (2.191978%) markers excluded because of low (<90%) call
   rate
5868 (2.993374%) markers excluded because they are out of HWE
    (FDR <0.2)
0 (0%) people excluded because of low (<90%) call rate
Mean autosomal HET is 0.3364606 (s.e. 0.007519463)
0 people excluded because too high autosomal heterozygosity (
   FDR <1%)
Mean IBS is 0.7280533 (s.e. 0.008871439), as based on 2000
   autosomal markers
16 (1.084746%) people excluded because of too high IBS
   (>=0.95)
In total, 102649 (52.36312%) markers passed all criteria
In total, 1459 (98.91525%) people passed all criteria
```

```
RUN 2
102649 markers and 1459 people in total
172 (0.1675613%) markers excluded as having low (<5%) minor
   allele frequency
0 (0%) markers excluded because of low (<90%) call rate
0 (0%) markers excluded because they are out of HWE (FDR
   <0.2)
0 (0%) people excluded because of low (<90%) call rate
Mean autosomal HET is 0.3367832 (s.e. 0.00752099)
0 people excluded because too high autosomal heterozygosity (
   FDR <1%)
Mean IBS is 0.7302109 (s.e. 0.008353749), as based on 2000
   autosomal markers
0 (0%) people excluded because of too high IBS (>=0.95)
In total, 102477 (99.83244%) markers passed all criteria
In total, 1459 (100%) people passed all criteria

RUN 3
102477 markers and 1459 people in total
0 (0%) markers excluded as having low (<5%) minor allele
   frequency
0 (0%) markers excluded because of low (<90%) call rate
0 (0%) markers excluded because they are out of HWE (FDR
   <0.2)
0 (0%) people excluded because of low (<90%) call rate
Mean autosomal HET is 0.3367832 (s.e. 0.00752099)
0 people excluded because too high autosomal heterozygosity (
   FDR <1%)
Mean IBS is 0.7278335 (s.e. 0.008325251), as based on 2000
   autosomal markers
0 (0%) people excluded because of too high IBS (>=0.95)
In total, 102477 (100%) markers passed all criteria
In total, 1459 (100%) people passed all criteria
```

B.2 Descriptive summary tables

For the markers that passed the quality control

```
> data_qc = data[qc$idok, qc$snpok]
> descriptives.marker(data_qc)
```

```
$`Minor allele frequency distribution`
   X<=0.01 0.01<X<=0.05 0.05<X<=0.1 0.1<X<=0.2  X>0.2
No       0            0       17345      26409  58723
Prop     0            0       0.169      0.258  0.573

$`Cumulative distr. of number of SNPs out of HWE, at
   different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05   all X
No          0        0     517     4630 102477
Prop        0        0   0.005    0.045      1

$`Distribution of proportion of successful genotypes (per
   person)`
     X<=0.9 0.9<X<=0.95 0.95<X<=0.98 0.98<X<=0.99 X>0.99
No        0           0            0            0   1459
Prop      0           0            0            0      1

$`Distribution of proportion of successful genotypes (per SNP
   )`
     X<=0.9 0.9<X<=0.95 0.95<X<=0.98 0.98<X<=0.99 X>0.99
No        0         262          778         1070 100367
Prop      0       0.003        0.008         0.01  0.979

$`Mean heterozygosity for a SNP`
[1] 0.3373553

$`Standard deviation of the mean heterozygosity for a SNP`
[1] 0.1288184

$`Mean heterozygosity for a person`
[1] 0.3367832

$`Standard deviation of mean heterozygosity for a person`
[1] 0.00752099
```

For the markers on chromosome 1 that passed the quality control

```
> descriptives.marker(data_all_chr1)
$`Minor allele frequency distribution`
     X<=0.01 0.01<X<=0.05 0.05<X<=0.1 0.1<X<=0.2    X>0.2
```

```
No          0             0    1479.000   2320.000 5270.000
Prop        0             0       0.163      0.256    0.581

$`Cumulative distr. of number of SNPs out of HWE, at
   different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 all X
No           0        0  44.000 344.000  9069
Prop         0        0   0.005   0.038     1

$`Distribution of proportion of successful genotypes (per
   person)`
     X<=0.9 0.9<X<=0.95 0.95<X<=0.98 0.98<X<=0.99 X>0.99
No        0           0            0            0   1459
Prop      0           0            0            0      1

$`Distribution of proportion of successful genotypes (per SNP
   )`
     X<=0.9 0.9<X<=0.95 0.95<X<=0.98 0.98<X<=0.99   X>0.99
No        0      26.000        95.00      119.000 8829.000
Prop      0       0.003         0.01        0.013    0.974

$`Mean heterozygosity for a SNP`
[1] 0.3402256

$`Standard deviation of the mean heterozygosity for a SNP`
[1] 0.1281584

$`Mean heterozygosity for a person`
[1] 0.3399593

$`Standard deviation of mean heterozygosity for a person`
[1] 0.02427776
```

B.3 Results from t-tests

```
> t.test(data_keep@phdata$act, data_remove@phdata$act, var.
   equal = TRUE)

   Two Sample t-test

data:  data_keep@phdata$act and data_remove@phdata$act
```

```
t = 0.61713, df = 1445, p-value = 0.5372
alternative hypothesis: true difference in means is not equal
    to 0
95 percent confidence interval:
 -0.3137388  0.6017587
sample estimates:
mean of x mean of y
 3.427453  3.283443

> t.test(data_keep@phdata$vo2max, data_remove@phdata$vo2max,
   var.equal = TRUE)

   Two Sample t-test

data:  data_keep@phdata$vo2max and data_remove@phdata$vo2max
t = -0.14428, df = 1456, p-value = 0.8853
alternative hypothesis: true difference in means is not equal
    to 0
95 percent confidence interval:
 -4.278374  3.692135
sample estimates:
mean of x mean of y
 136.3968  136.6899

> t.test(data_keep@phdata$age, data_remove@phdata$age, var.
   equal = TRUE)

   Two Sample t-test

data:  data_keep@phdata$age and data_remove@phdata$age
t = 0.034046, df = 1457, p-value = 0.9728
alternative hypothesis: true difference in means is not equal
    to 0
95 percent confidence interval:
 -1.934152  2.002478
sample estimates:
mean of x mean of y
 49.48768  49.45351
```