Stockholms universitet

Comparison of logistic and ridge regression in genetic association studies

Hanna Fues

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2015:6
Matematisk statistik
September 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Mathematical Statistics
Stockholm University
Master Thesis **2015:6**
http://www.math.su.se

Comparison of logistic and ridge regression in genetic association studies

Hanna Fues*

September 2015

Abstract

Genetic association studies are used to find regions of the genome that contribute to a specific disease by testing for an association between disease status and genetic variation(s). We use single nucleotide polymorphisms (SNPs) to investigate the genetic variability. SNPs are locations where single nucleotides differ on the DNA in at least 1% of the population. From a statistical point of view these investigations are commonly performed using logistic regression techniques, modeling disease risk as a function of markers, i.e. SNPs, but the need for a penalized regression approach arises when many markers are correlated. Furthermore, in genetic association studies one often has the situation where the number of markers studied exceed the number of observations, increasing the need for a penalized approach. In this thesis interest lies in the analysis of a genomic region, containing highly correlated markers, in relation to breast cancer risk. We do so by studying data on 89 050 individuals part of the Breast Cancer Associ- ation Consortium (BCAC) and compare logistic regression and ridge regression techniques, for localizing independent signals among the multiple markers in said genomic region. Furthermore, we use a data-driven method for estimation of the penalization parameter, proposed by Cule and De Iorio [2013], in our ridge regression analyses. We find that both regression approaches give similar results, deeming markers from the same genomic region as being significantly associated with breast cancer risk. Our analysis is the first step in a long chain of events leading to the identification of locations that are associated with breast cancer risk and this thesis gives an important indication as to what region future investigations should be focusing on.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: hannafues@gmail.com. Supervisor: Michael Höhle.

Acknowledgement

I want to thank my supervisor Hatef Darabi for giving me the opportunity to write this thesis at the Department of Medical Epidemiology and Biostatistics at Karolinska Insitutet and for all the support, advice, discussions and pep talks during these five months. Furthermore, I want to thank my supervisor at Stockholm University, Michael Höhle, for valuable feedback and advice troughout the duration of the thesis.

Contents

Chapter 1

Introduction

Each year in Sweden more than 57 000 cancer cases are diagnosed, of these over 8 000 are breast cancers. Approximately 30% of all cancers among Swedish women are breast cancers, making it the most common cancer among women in Sweden [Cancerfonden]. Both non-genetic and genetic risk factors are involved in the initiation and the course of the illness. The genetic component of the disease is reflected by a tendency for it to cluster in families [Lichenstein et al., 2000]. For example women who have a mother or sister who had breast cancer have a two to three times higher risk than other women of developing the disease themselves [Cancerfonden].

The aim of this thesis is to analyze a highly correlated genomic region in relation to breast cancer risk by comparing logistic regression methodology to ridge regression techniques for localizing independent associated signals among multiple markers. The analyses will be done on a breast cancer data set containing information about 89 050 individuals disease status and their genotypes at 3 279 markers along a genomic region. The data originates from 39 case-control studies participating in BCAC [Breast Cancer Association Consortium].

In the first year of the new millennium it was announced that the International Human Genome Project and Celera Genomics Corporation had both completed an initial sequencing of the human genome [International Human Genome Sequencing Consortium, 2001]. Subsequent technological advances concerned with gene sequencing and genotyping initiatives have improved our understanding of population specific genetic variations and enhanced our understanding of complex diseases; often through the use of genetic association studies. Genetic association studies are used to find regions of the genome that contribute to a specific disease by testing for a correlation between disease status and genetic variation(s). These studies have led to the mapping of susceptibility loci for many diseases including breast cancer [Turnbull et al., 2010]; by modeling risk of the disease as a function of investigated loci and assessing association. Large sample sizes have, however, been needed to detect and confirm genetic variants that are associated with modest increase in risk. The Breast Cancer Association Consortium (BCAC) [Breast Cancer Association Consortium, 2006], which was established to conduct collaborative studies, has detected several of the currently known common variants associated with breast cancer risk [Michailidou et al., 2015].

Logistic regression is commonly used for modeling disease risk as a function of a marker. See for example French et al. [2013] where the authors analyzed data from case-control studies with regard to breast cancer, and used, among other, both univariable logistic and multivariable stepwise logistic regression to model disease risk as a function of markers. However, using univariable analyses for genomic data is not optimal when, for example, many genetic variants that contribute to disease are interdependent. It would be more appropriate to use multivariable techniques that allow for the study of combined effects of multiple markers. Furthermore, in genetic association studies, one often has the situation where the number of covariates exceeds the number of observations (the small n, large p problem). In such situations a penalized regression approach, that potentially also can handle multicollinearity between markers, would be more suitable.

As outlined above we will in this thesis analyze breast cancer data, using both logistic regression methods and ridge regression, which is a penalized regression approach. We are interested in comparing the methods in regard to which markers the respective methods localize. Throughout the thesis we assume that the theory of linear regression modeling and its details is known to the reader. For supplementary reading about the linear model consider e.g. Hastie et al. [2010, chapter 3].

The this is structured as follows. In Chapter 2, we first provide a short description of the structure of the human genome. In Chapter 3 some key concepts related to genetic association studies are introduced as background for the applied work in this thesis. Logistic regression modeling techniques for genetic association studies are introduced in Chapter 4, as well as maximum likelihood techniques for estimation of the model parameters. Furthermore, Section 4.3 covers some theory of penalized regression modeling and logistic ridge regression in particular, while Section 4.4 deals with the concept of principal component logistic regression. Principal component regression techniques will come in handy in Section 4.5, where the proposed data-driven approach, by Cule and De Iorio [2013], for estimation of the penalization parameter in ridge regression is explained. In Chapter 5 various aspects of the breast cancer data set we will be analyzing in this thesis are described and an initial univariable analysis is performed. The results of the multivariable logistic regression and ridge regression analyses are presented in Chapter 6 and the results and various other aspects of this thesis are discussed in Chapter 7. Furthermore, Appendix A and Appendix B provide some supplementary theory.

Chapter 2

Biological Background

Most human body cells have in their cell nucleus 23 pairs of chromosomes; 22 pairs of autosomal (non-sex) chromosomes and 1 pair of sex chromosomes (XX or XY). Each pair comprises one copy from each parent respectively. The two chromosomes in one autosomal pair are said to be homologous, meaning that they comprise (nearly) the same information in the same locations [Aerssens et al., 2001].

Each chromosome is made up of two strands of DNA (deoxyribonucleic acid) molecules, that are wound around each other to form a double helix structure, see Figure 2.1 **A**. Each strand of DNA is a single long molecule made up of four different building blocks called nucleotides or bases. The bases are adenine (A), cytosine (C), guanine (G) and thymine (T). The two strands of DNA are held together by binding between the opposing bases, where bonding only occurs between A and T, and C and G respectively; referred to as complementary base pairing. See Figure 2.1 **B** and **C** for the molecular structure of the four bases and their complementary base pairing. The specific sequence of bases encodes the genetic information needed to create proteins; where a stretch of DNA is called a gene if it contains said information. The gene content is then transcribed to create another molecule called RNA. The single-stranded RNA molecule is very similar to DNA and consists of a sequence of bases that is complementary to it's DNA template. This RNA is then transported from the nucleus to the cell cytoplasm where it is translated into a protein [Thomas, 2004]. This flow of genetic information from DNA to RNA to protein, is referred to as the central dogma of molecular biology [Genetic Science Learning Center, University of Utah].

Surprisingly, our DNA has very little variability; more than 99% of the base sequences in the DNA are the same in all humans. The loci, or DNA locations, that vary from person to person are called polymorphic, where the alternate sequences found at a polymorphic locus are called alleles. The term polymorphism is usually only used for those variations that are present in at least 1% of the population. The most common type of polymorphisms in the human genome are single nucleotide polymorphisms (SNPs, pronounced 'snips'), where a single base is substituted for another [Thomas, 2004]. Up to four different substitutions are possible at each SNP location in the genome: one for each nucleotide abbreviated as A, C, G and T [Genetic Science Learning Center, University of Utah], see Figure 2.2.

The two copies of a specific gene, one on each chromosome, inherited from the mother and father respectively, are not always identical. For most genes many alleles exist. An individual's genotype for a specific genetic variation is the combination of the two alleles present on the two homologous chromosomes [Aerssens et al., 2001]. For example, in humans most SNPs are biallelic, i.e. only two alleles are present, denoted by A and a, where $A, a \in \{\text{A, C, G, T}\}$. The possible genotypes in the population are thus AA (homozygote wildtype), Aa (heterozygote) and aa (homozygote variant allele) [Thomas, 2004].

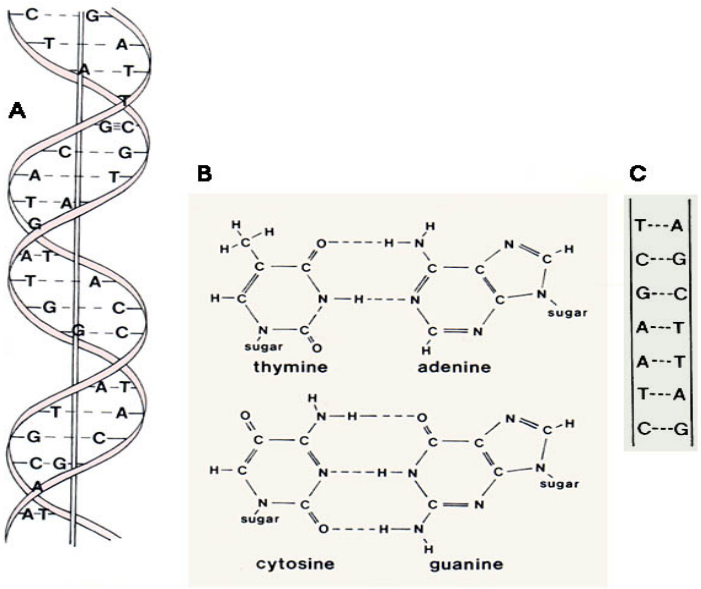

Figure 2.1: **A**: Double helix structure of DNA. **B**: Molecular structure of the four building blocks of DNA and their complementary base pairing. **C**: Unwound DNA showing pairing of the bases. Source: http://www.apsnet.org/edcenter/K-12/TeachersGuide/DNA_Easy/Pages/Background.aspx (retrieved April 2, 2015).

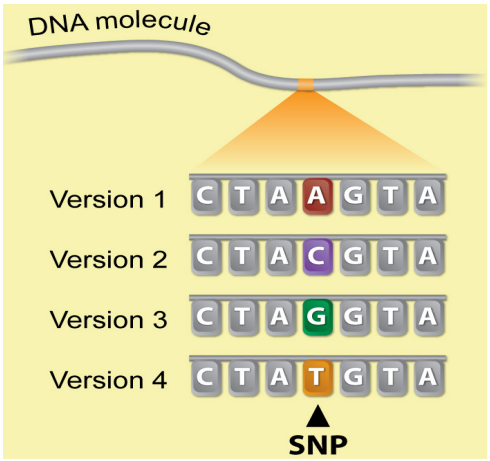

Figure 2.2: SNPs are single nucleotide substitutions of one base for another. Each SNP location can have up to four versions: one for each nucleotide A, C, G and T. Source: Genetic Science Learning Center, University of Utah (retrieved April 1, 2015).

Chapter 3

Genetic Association Studies

In genetic association studies the statistical association between a person's genotype with his/her phenotype, i.e. observable outcome or disease status, is examined [Li, 2008]. This is done to identify those SNPs that contribute to that specific phenotype [Lewis and Knight, 2012].

The aim of this chapter is to introduce genetic association studies and some of its concepts such as linkage disequilibrium, minor allele frequency and Hardy-Weinberg equilibrium. We look at how genetic data from a case-control study can be summarized in a contingency table and how the association between genotype and phenotype can be assessed from it. Furthermore, we introduce the genome-wide association study significance level $\alpha = 5 \times 10^{-8}$ which we will use throughout this thesis.

Genetic association studies are mostly based on case-control studies. Here, a number of cases having the disease of interest are collected together with a number of control individuals. Standard methods for collecting controls are either to use a series of individuals who have been screened as negative for presence of the disease or to use individuals randomly selected from the population, whose disease status is unknown [Lewis and Knight, 2012]. The frequency of genotypes or alleles is compared between cases and controls at each SNP and a significant difference in frequency between the case and control group is suggestive of an association of that SNP with disease [Aerssens et al., 2001].

For a single biallelic SNP tested in a case-control study we can summarize the genotype and allele counts in a 2×3 or 2×2 contingency table, respectively, as seen in Table 3.1.

	Genotype			
	AA	Aa	aa	Total
Cases	a	b	c	n_{case}
Controls	d	e	f	n_{cont}
Total	n_{AA}	n_{Aa}	n_{aa}	n

	Allele		
	A	a	Total
Cases	$2a + b$	$b + 2c$	$2n_{case}$
Controls	$2d + e$	$e + 2f$	$2n_{cont}$
Total	$n_{Aa} + 2n_{AA}$	$n_{Aa} + 2n_{aa}$	$2n$

Table 3.1: Genotype (left) and allele (right) counts observed for a single SNP tested in a case-control study.

Analyses could be done on either the genotype or allele counts as presented in Table 3.1. Sasieni [1997] recommends that SNP data should be analyzed by genotype rather than by alleles, since

analyses of the 2×2 table above are only valid when the combined case-control population is in Hardy-Weinberg equilibrium (see definition in Section 3.1.3). In other words, analyses of genotype data are more robust to departures from Hardy-Weinberg equilibrium.

There are a number of different methods available to asses the association between genotype and phenotype. One is to use Pearson's chi square test to asses departure from the null hypothesis that cases and controls have the same distribution of genotype counts. On the 2×3 table of genotype counts, the test statistic has a chi-squared distribution with two degrees of freedom [Lewis and Knight, 2012].

Furthermore, logistic regression is commonly used to model the disease risk of each individual given their genotype, see detailed description in Section 4.2. More recently, penalized regression approaches, such as ridge regression, have been utilised to enable the analysis of several, possibly correlated, SNPs at once [Cule and De Iorio, 2013].

In genome-wide association studies (GWAS) usually a large number of statistical tests are computed, one for each SNP, which requires a definition of a genome-wide threshold of significance that accounts for test multiplicity and guards against false positive results that will occur by chance when performing a large number of tests at a standard significance level α. When estimating a threshold of significance, false discoveries can be controlled by using, for example, the simple and conservative Bonferroni correction $\alpha^* = \alpha/p$, where p is the number of tested SNPs. The Bonferroni correction's assumption of independent tests will, however, be violated if there is a dependence between the SNPs. To take into account the dependence between SNPs, permutation procedures and estimation of an effective number of tests, which is the equivalent number of independent SNPs along the genome, have been proposed, then choosing the significance level as $\alpha^* = \alpha/k$, where k is the number of effective tests [Jannot et al., 2015]. For example, Pe'er et al. [2008] used permutation procedures to show that one million tests is the effective number of tests genome-wide in Europeans, based on the data collected by the International HapMap Consortium. This yields a significance level $\alpha^* = 5 \times 10^{-8}$, which, as shown by Jannot et al. [2015], has become the standard genome-wide significance threshold in GWAS.

3.1 Concepts in Genetic Association Studies

3.1.1 Linkage disequilibrium

Linkage disequilibrium (LD) is defined as the tendency for two alleles at two loci to be associated with each other in the population more than would be expected by chance. This may be the result of co-inheritance of particular alleles at neighboring loci and can lead to a correlation between SNPs in the population [Ardlie et al., 2002].

If we denote two adjacent biallelic loci by A and B with alleles A and a, and B and b respectively, then each of the $2^2 = 4$ different combinations of alleles are called a haplotype. The classical

definition of the LD coefficient is

$$D = p_{AB} - p_A p_B, \tag{3.1.1}$$

where p_{AB} denotes the observed haplotype frequency for the haplotype that consists of alleles A and B and $p_A p_B$ denotes the expected haplotype frequency, where p_A is the frequency of allele A at the first locus and p_B is the frequency of allele B at the second locus. Hence, D measures the deviation between the haplotype frequency and its expectation under independence [Langefeld and Fingerlin, 2007].

There are two other common measures of LD, namely D' and r^2, that both range between 0 (linkage equilibrium) and 1 (complete LD) [Ardlie et al., 2002]. Thus as D' and r^2 approach 1 the correlation between the loci increases. We have that [Thomas, 2004]

$$D' = \frac{|D|}{D_{max}} \qquad D_{max} = \begin{cases} \min(p_A p_b, p_a p_B) & \text{if } D \geq 0 \\ \min(p_A p_B, p_a p_b) & \text{if } D < 0. \end{cases} \tag{3.1.2}$$

The measure r^2 is the square of Pearson's correlation coefficient for the two loci, i.e.

$$r^2 = \frac{D^2}{p_A p_B p_a p_b}. \tag{3.1.3}$$

Only when $D' = 1$ and the allele frequencies at the two loci are identical, will r^2 be 1 [Langefeld and Fingerlin, 2007] and this is referred to as perfect LD [Ardlie et al., 2002].

3.1.2 Minor allele frequency

The frequency of the allele that is present in no more than 50% of a population, i.e. the less common allele at a variable site, is called the minor allele frequency (MAF) [Foulkes, 2009]. As an example consider the SNP for which AA is present in 75% of the population, Aa in 20% and aa in 5%. The frequency for the A and a allele respectively, is then estimated as

$$\text{freq}(A) = \frac{(2 \cdot 75 + 20)\%}{2} = 85\% \tag{3.1.4}$$

$$\text{freq}(a) = \frac{(2 \cdot 5 + 20)\%}{2} = 15\%. \tag{3.1.5}$$

In this case the minor allele frequency, i.e. the frequency of the a-allele, is 15% [Foulkes, 2009].

Theorem. *When the genotypes G are encoded as 0, 1 and 2 for the number of a-alleles present, i.e.*

$$G = \begin{cases} 0 & \textit{if genotype } AA, \\ 1 & \textit{if genotype } Aa, \\ 2 & \textit{if genotype } aa, \end{cases}$$

and 'a' is the minor allele, the minor allele frequency at a specific SNP can be calculated as

$$\text{MAF} = \frac{\sum_{i=1}^{n} G_i}{2n} = \frac{\overline{G}}{2},$$

where n is the number of individuals we have genotyped.

Proof.

$$\begin{aligned}\text{MAF} = \text{freq}(a) &= \frac{2 \cdot n_{aa} + 1 \cdot n_{Aa} + 0 \cdot n_{AA}}{2n} \\ &= \frac{2 \cdot \sum_{i=1}^{n} \mathbf{I}(G_i = aa) + 1 \cdot \sum_{i=1}^{n} \mathbf{I}(G_i = Aa) + 0 \cdot \sum_{i=1}^{n} \mathbf{I}(G_i = AA)}{2n} \\ &= \frac{\sum_{i=1}^{n} G_i}{2n} = \frac{\overline{G}}{2},\end{aligned}$$

where $\mathbf{I}(\cdot)$ is the indicator function taking value 1 if its condition is fulfilled and value 0 otherwise. We see that $\sum_{i=1}^{n} \mathbf{I}(G_i = aa)$ counts the number of individuals with genotype aa, thus multiplying this with 2 yields the sum of all $G_i = 2$, while $\sum_{i=1}^{n} \mathbf{I}(G_i = Aa)$ yields the sum of all $G_i = 1$ and $0 \cdot \sum_{i=1}^{n} \mathbf{I}(G_i = AA)$ yields the sum of all $G_i = 0$.

□

3.1.3 Hardy-Weinberg Equilibrium

The Hardy-Weinberg equilibrium (HWE) states that under specific assumptions (random mating, no new mutations, no migration or selection) the allele and genotype frequencies in a population, after enough time, will remain constant from generation to generation, see Langefeld and Fingerlin [2007]. That is, all subsequent generations will have the same genotype frequencies unless there are violations of these assumptions. Furthermore, the genotype probabilities of AA, Aa and aa are under HWE p^2, $2pq$ and q^2 respectively, where p and $q = 1 - p$ are allele probabilities for A and a respectively. In practice, HWE is highly robust to departures from all of the assumptions except for random mating relative to the locus of interest [Langefeld and Fingerlin, 2007].

Lewis and Knight [2012] state that controls in a case-control study should be in HWE, provided the population they were selected from fulfills the above assumptions. Furthermore, departure from HWE in controls also commonly implies genotyping errors or a population substructure and a test for a statistical difference between the expected and observed genotype counts in each SNP separately can thus be performed in the beginning of the analysis as a quality control check, discarding the SNPs not in HWE. For this test the chi-squared goodness-of-fit test statistic is calculated for each SNP by summing over the possible genotypes

$$X^2 = \sum_{G \in \{AA, Aa, aa\}} \frac{(\text{observed count}_G - \text{expected count}_G)^2}{\text{expected count}_G},$$

and here X^2 has an asymptotically chi-squared distribution with $(3-1)-1=1$ degree of freedom. The expected genotype probabilities are calculated as in (3.1.4) and (3.1.5) and the expected counts as $n_{cont}p^2$ (AA), $2n_{cont}pq$ (Aa) and $n_{cont}q^2$ (aa), where n_{cont} is the total number of controls in that specific SNP [Langefeld and Fingerlin, 2007].

Chapter 4

Theory & Methods

In this chapter we introduce the methods we will be using in this thesis, and the theory behind them. We start by introducing the logistic regression model in Section 4.2. The model is introduced in general and applied to the analysis of case-control studies to estimate the covariate (SNP) effect on the prospective odds ratio, as described in Chapter 3. Furthermore, in Section 4.3 we introduce penalized regression models, in particular the logistic ridge regression model, which are useful in data situations where the covariates are highly correlated or when the number of covariates exceeds the number of observations. In these situations, which both arise commonly in genetic association studies, the parameters in the model cannot be uniquely estimated. If instead a penalized regression approach is utilized, (biased) parameter estimates can be obtained. Maximum likelihood (ML) techniques for estimation of the model parameters are discussed and the ML estimators in the logistic and ridge regression models are derived. Methods for estimating the penalization parameter in ridge regression are discussed in Section 4.3.2, in particular the method suggested by Cule and De Iorio [2013] since we will be utilizing it. Their approach is based on principal component regression and hence the principal component logistic regression model is also introduced in Section 4.4.

Furthermore, in Section 4.6, we introduce two methods used in meta-analysis for pooling p-values from several different studies: Fisher's method and Stouffer's weighted Z-score method.

4.1 Prospective vs. retrospective odds ratio

Returning to the 3×2 contingency table, described in Chapter 3, of genotype counts given disease status resulting from a case-control study, we write the assumed conditional probabilities for the events as in Table 4.1b. Here, for example, $P(x = 2 \mid y = 1) = \phi_{12}$ and $P(x = 2 \mid y = 0) = \phi_{02}$ denote the probabilities of having genotype 2 when being a case or control, respectively. These probabilities are estimated by their observed proportions, a/n_{case} and b/n_{cont} respectively.

In the following we show the equivalence of the prospective and retropective odds ratios, following Lachlin [2011, chapter 5]. Throughout the thesis, genotype AA, denoted by 0, will serve as the reference genotype group, meaning that when we talk about the odds of genotype $x = 1$ or 2, it

			y		
			Case	Control	
Genotype			1	0	Total
	aa	2	a	b	n_{aa}
x	Aa	1	c	d	n_{Aa}
	AA	0	e	f	n_{AA}
		Total	n_{case}	n_{cont}	n

(a) Table of observed genotype counts.

			y	
			Case	Control
Genotype			1	0
	aa	2	ϕ_{12}	ϕ_{02}
x	Aa	1	ϕ_{11}	ϕ_{01}
	AA	0	$1-\phi_{12}-\phi_{11}$	$1-\phi_{02}-\phi_{01}$
		Total	1	1

(b) Table of assumed conditional probabilities.

Table 4.1: Contingency table of (a) observed genotype counts given samples of n_{case} cases and n_{cont} controls and (b) the corresponding assumed conditional probabilities.

is always in relation to genotype 0. The retrospective OR of genotype aa given disease is

$$\mathrm{OR}_{retro,20} = \frac{P(x=2 \mid y=1)/P(x=0 \mid y=1)}{P(x=2 \mid y=0)/P(x=0 \mid y=0)} = \frac{\phi_{12}/(1-\phi_{12}-\phi_{11})}{\phi_{02}/(1-\phi_{02}-\phi_{01})} \mathrel{\widehat{=}} \frac{af}{be}, \tag{4.1.1}$$

where '$\widehat{=}$' in the last step means 'estimated as'. $\mathrm{OR}_{retro,20}$ is thus the odds of having genotype aa when diseased vs. not diseased. Interest is however, in the prospective OR of being diseased given genotype aa:

$$\mathrm{OR}_{20} = \frac{P(y=1 \mid x=2)/P(y=1 \mid x=0)}{P(y=0 \mid x=2)/P(y=0 \mid x=0)}.$$

To obtain the OR above one also needs to account for the prevalence of the disease in the population from which the cases and controls arose: $P(y=1)=\delta$. In that case we have, using Bayes theorem, that

$$\begin{aligned}
P(y=1 \mid x=2) &= \frac{P(x=2 \mid y=1)P(y=1)}{P(x=2)} \\
&= \frac{P(x=2 \mid y=1)P(y=1)}{P(x=2 \mid y=1)P(y=1) + P(x=2 \mid y=0)P(y=0)} \\
&= \frac{\phi_{12}\delta}{\phi_{12}\delta + \phi_{02}(1-\delta)} \\
P(y=1 \mid x=0) &= \frac{P(x=0 \mid y=1)P(y=1)}{P(x=0)} \\
&= \frac{P(x=0 \mid y=1)P(y=1)}{P(x=0 \mid y=1)P(y=1) + P(x=0 \mid y=0)P(y=0)} \\
&= \frac{(1-\phi_{12}-\phi_{11})\delta}{(1-\phi_{12}-\phi_{11})\delta + (1-\phi_{02}-\phi_{01})(1-\delta)}.
\end{aligned}$$

Together this yields

$$\begin{aligned}
\mathrm{OR}_{02} &= \frac{P(y=1 \mid x=2)/P(y=1 \mid x=0)}{P(y=0 \mid x=2)/P(y=0 \mid x=0)} \\
&= \frac{P(x=2 \mid y=1)/P(x=0 \mid y=1)}{P(x=2 \mid y=0)/P(x=0 \mid y=0)} \\
&= \mathrm{OR}_{retro,02}.
\end{aligned} \tag{4.1.2}$$

Equation (4.1.2) shows that the retrospective and prospective OR are equivalent, i.e. the odds of having genotype aa when diseased compared to when not diseased is the same as the odds of disease when having genotype aa compared to when having genotype AA. The same can be showed for when $x = 1$, i.e. the genotype is Aa. Thus we have showed that the OR of genotype status given disease from a retrospectively sampled case-control study also provides an estimate of the prospective OR of disease given genotype status.

This suggests that logistic regression can be applied to the analysis of case-control studies to estimate and asses the covariate effects on the prospective OR.

4.2 Logistic regression

The theory in this chapter can be found in Agresti [2013], unless otherwise indicated.

Let $\mathbf{X}$ be an $n \times (p+1)$ matrix of covariates, with rows $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$, setting $x_{i0} = 1$ for an intercept term, and let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be a random vector of binary outcomes, i.e. $Y_i \in \{0, 1\}$. In our case, $\mathbf{X}$ is the covariate matrix with genotype information for each individual i, $i = 1, \ldots, n$, at each SNP j, $j = 1, \ldots, p$, meaning that $x_{ij} \in \{0, 1, 2\}$ depending on the number of a-alleles in individual i's genotype at SNP j. Furthermore, $\mathbf{Y}$ is the vector of phenotypes where 1 denotes a case and 0 a control.

In this setup, where Y_i is a Bernoulli distributed random variable with success probability $\pi(\mathbf{x}_i)$, the logistic regression model is given as

$$\pi(\mathbf{x}_i) = \mathrm{P}(Y_i = 1 \mid \mathbf{x}_i) = \frac{\exp\{\mathbf{x}_i\boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_i\boldsymbol{\beta}\}} = \frac{\exp\{\sum_{j=0}^{p} \beta_j x_{ij}\}}{1 + \exp\{\sum_{j=0}^{p} \beta_j x_{ij}\}}, \tag{4.2.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the vector of coefficients. Equivalently, the log odds has the linear relationship

$$\text{logit}\left(\pi(\mathbf{x}_i)\right) = \log\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right) = \mathbf{x}_i\boldsymbol{\beta} = \sum_{j=0}^{p} \beta_j x_{ij}. \tag{4.2.2}$$

When it comes to interpreting the coefficients, we have that $\exp(\beta_j)$, where $j = 1, \ldots, p$, is the multiplicative effect on the odds for disease obtained from a 1-unit increase in the jth covariate when keeping the levels of all other covariates fixed.

4.2.1 Maximum likelihood estimation of the logistic regression model

The likelihood of the above described logistic regression model, where the Y_i's are conditionally independent given the $\mathbf{x}_i$'s and $Y_i = y_i \sim \text{Bernoulli}(\pi(\mathbf{x}_i))$, $i = 1, \ldots, n$, is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1 - y_i}, \tag{4.2.3}$$

where $\pi_i = \pi(\mathbf{x}_i)$, as in (4.2.1). The log likelihood is thus

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)\}. \tag{4.2.4}$$

Note that the density of $\boldsymbol{y}$ characterized by 4.2.3 belongs to an exponential family with canonical statistic

$$\boldsymbol{t}(\boldsymbol{y}) = \left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i x_{i1}, \ldots, \sum_{i=1}^{n} y_i x_{ip} \right),$$

canonical parameter $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ and $a(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[1 + \exp\{x_i \boldsymbol{\beta}\}\right]^{-1}$.

This implies that the logistic regression model belongs to the class of generalized linear models (GLMs). The link function, that connects the mean value, $E[Y_i] = \pi_i$, to the linear predictor, $\eta_i = g(\pi_i) = \mathbf{x}_i \boldsymbol{\beta}$, is the canonical logit link, i.e. $g(\pi_i) = \log\left[\pi_i/(1 - \pi_i)\right] = \text{logit}(\pi_i)$, as in (4.2.2).

We use maximum likelihood estimation to get an estimate of $\boldsymbol{\beta}$. Differentiating $l(\boldsymbol{\beta})$ and setting equal to zero, yields the likelihood equations

$$\sum_{i=1}^{n} \mathbf{x}_i' \left[y_i - \widehat{\pi}_i\right] = \mathbf{X}'[\boldsymbol{y} - \widehat{\boldsymbol{\pi}}] = \mathbf{0}, \tag{4.2.5}$$

where $\widehat{\pi}_i = \exp\{\mathbf{x}_i \widehat{\boldsymbol{\beta}}\} / \left[1 + \exp\{\mathbf{x}_i \widehat{\boldsymbol{\beta}}\}\right]$ is the probability obtained from plug in of the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\pi}}$ is the vector of $\widehat{\pi}_i$'s. Since these $p + 1$ equations are non-linear in $\boldsymbol{\beta}$ they require an iterative solution, for example the Newton-Raphson algorithm or Fisher scoring, which are asymptotically equivalent for the logistic regression model.

For the estimation of the regression coefficients and the covariance matrix of $\widehat{\boldsymbol{\beta}}$, the observed information matrix, or equivalently the Fisher information matrix, since they are identical here due to the canonical logit link, is required. We have that the Fisher information matrix is

$$I(\boldsymbol{\beta}) = E\left[-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right] = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \pi_i (1 - \pi_i) = \mathbf{X}'\mathbf{W}\mathbf{X}, \tag{4.2.6}$$

where $\mathbf{W} = \text{diag}[\pi_i(1 - \pi_i)]$ is the diagonal weight matrix. We recall that $\text{Var}(Y_i) = \pi_i(1 - \pi_i)$, so $\mathbf{W}$ is the variance-matrix of $\mathbf{Y}$.

Given a starting value $\boldsymbol{\beta}^{(0)}$, the Fisher scoring algorithm iteratively updates

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{(t+1)} &= \widehat{\boldsymbol{\beta}}^{(t)} + I^{-1}(\widehat{\boldsymbol{\beta}}^{(t)}) U(\widehat{\boldsymbol{\beta}}^{(t)}) \\ &= \widehat{\boldsymbol{\beta}}^{(t)} + \left[\mathbf{X}'\widehat{\mathbf{W}}^{(t)}\mathbf{X}\right]^{-1} \mathbf{X}' \left[\boldsymbol{y} - \widehat{\boldsymbol{\pi}}_i^{(t)}\right], \quad t = 0, 1, 2, \ldots \end{aligned} \tag{4.2.7}$$

where $I(\cdot)$ and $U(\cdot)$ are the expected Fisher information matrix and score function respectively. The iterations will be stopped once a convergence criterion is met, for example the relative convergence criterion $\|\widehat{\boldsymbol{\beta}}^{(t+1)} - \widehat{\boldsymbol{\beta}}^{(t)}\| / \|\widehat{\boldsymbol{\beta}}^{(t)}\| \leq \varepsilon$, where $\varepsilon > 0$ is a pre-defined number and $\|\boldsymbol{\beta}\|$

denotes the length of vector $\boldsymbol{\beta}$, then $\widehat{\boldsymbol{\beta}}^{(t)}$ is the maximum likelihood estimator [Fahrmeir et al., 2013].

The ML estimators have an asymptotic normal distribution with variance-covariance matrix equal to the inverse of the Fisher information matrix, which is estimated by plug in of the ML estimator $\widehat{\boldsymbol{\beta}}$, i.e.

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = \left[\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right]^{-1}. \tag{4.2.8}$$

The existence of the ML estimates depends on the configuration of the sample points in the observation space. There are three mutually exclusive and exhaustive categories: complete separation, quasi-complete separation and overlap [Albert and Anderson, 1984]. Complete separation refers to the situation when a hyperplane can pass through the space of explanatory variables such that on one side of that hyperplane $Y = 1$ for all observations whereas $Y = 0$ for all observations on the other side. When at least one observation from each response group is exactly on the hyperplane we have quasi-complete separation. When neither of these separations exists there is an overlap of observations and the ML estimates exist and are unique [Agresti, 2013].

4.2.2 Forward selection

Having covered the theory of logistic regression models in the previous sections, we move on to model selection. Forward selection is a variable selection procedure that starts with only an intercept in the model and then adds variables sequentially. At each stage it selects the variable giving the greatest improvement in fit according to some criterion. The procedure stops when further additions do not significantly improve the fit [Agresti, 2013].

The inclusion criterion we will be looking at is the Akaike information criterion (AIC), which balances the goodness of fit of the model, as measured by the log likelihood, with its complexity, and the model with minimum AIC value is the preferred one. Altogether, the AIC of a model is, in the logistic regression setting, defined as

$$\mathrm{AIC} = -2\left(l(\widehat{\boldsymbol{\beta}}) - \dim(\boldsymbol{\beta})\right),$$

where $l(\widehat{\boldsymbol{\beta}})$ is the maximized log likelihood. Denoting the maximized likelihood of the model with L and the dimension of the parameter vector, i.e. the number of parameters in the model, with p, the information criterion can be written as $\mathrm{IC} = -2\log(L) + kp$, where setting $k = 2$ yields AIC.

If we consider two models, M_0 with p parameters and M_1 with $p+1$ parameters, we will choose M_1 over M_0 if $\mathrm{AIC}_1 < \mathrm{AIC}_0$. With L_i denoting the maximized likelihood of model $i = 0, 1$, we

have that

$$\begin{aligned} & \mathrm{AIC}_1 < \mathrm{AIC}_0 \\ \iff & -2\log(L_1) + k(p+1) < -2\log(L_0) + kp \\ \iff & \underbrace{-2\log(L_0/L_1)}_{=X^2} > k. \end{aligned} \tag{4.2.9}$$

We recognise the left hand side as the likelihood ratio (LR) test statistic X^2, also called deviance, which has a χ^2-distribution with $(p+1)-p=1$ degree of freedom. The probability of X^2 taking a value larger than $k=2$ is approximately 16%, i.e. $\mathrm{P}(X^2 > 2) \approx 0.16$. Thus AIC is equivalent to a LR test at significance level $\alpha \approx 0.16$, when assessing a single variable at a time. The significance level of the test can be adjusted by changing the value of k. For example choosing $k = 10.83$, yields a LR test at $\alpha \approx 10^{-3}$, while $k = 29.72$ yields a LR test at GWAS significance level $\alpha \approx 5 \times 10^{-8}$. We will denote the forward selection procedure as forward likelihood ratio inclusion at significance level α.

4.3 Penalized regression

In the situation when the number of covariates exceeds the number of observations in a regression model, i.e. the small n, large p problem, or when the covariates are highly collinear, the parameters in the model cannot be uniquely estimated. In the logistic regression model this is due to the (near) singularity of the matrix $\mathbf{X'WX}$ when estimating the maximum likelihood estimates of the model, see Equation (4.2.7), making the inversion of this matrix impossible. By applying a penalty to the diagonal of $\mathbf{X'WX}$, its inversion is made possible but this introduces high bias in the coefficient estimates of the regression model, whereas the ordinary logistic regression coefficient estimates are unbiased. On the other hand, the penalization yields coefficient estimates with lower variances than in the unpenalized model. Weighting bias and variance of the coefficient estimates against each other in the penalized and unpenalized models is referred to as the bias-variance trade-off and can be described by for example the mean squared error (MSE); where one is interested in having a lower MSE in the penalized model than in the unpenalized [Fahrmeir et al., 2013].

If we consider an arbitrary parametric model with parameter vector $\boldsymbol{\beta}$ and log likelihood function $l(\boldsymbol{\beta})$, then the penalized likelihood estimator of $\boldsymbol{\beta}$ maximizes the penalized log likelihood

$$l_P(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda\, \mathrm{Pen}(\boldsymbol{\beta}), \tag{4.3.1}$$

where $\mathrm{Pen}(\cdot) > 0$ is a penalty function and $\lambda > 0$ is a penalization parameter controlling the strength of the penalty term [Fahrmeir et al., 2013].

The three methods commonly mentioned in relation to penalized regression are Lasso (least absolute shrinkage and selection operator), ridge regression and Elastic net. In Lasso regression, proposed by Tibshirani [1996], the penalty function is the L_1-norm, i.e. $\mathrm{Pen}(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$,

while in ridge regression [Hoerl and Kennard, 1970a,b] it is the L_2-norm $\text{Pen}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$. The Elastic net penalty [Zou and Hastie, 2005] is a combined penalty of lasso and ridge regression penalties, where a parameter determines how much weight should be given to either of the two. In this thesis we will focus on the use of ridge regression in the setting of logistic regression models.

4.3.1 Logistic ridge regression

The penalized method we will be using in this thesis is ridge regression (RR). Here the penalty function is the squared L_2-norm:

$$\text{Pen}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2 = \boldsymbol{\beta}'\boldsymbol{\beta}.$$

The degree of penalization depends on the parameter λ in Equation (4.3.1). Increasing λ yields greater shrinkage of the parameter estimates towards zero [Agresti, 2013]. Note however, that ridge regression does not perform variable selection, meaning that parameter estimates are only shrunk towards zero and never exactly equal to zero.

For the logistic ridge regression model, the penalized log likelihood to be maximized, takes the form

$$\begin{aligned} l_{RR}(\boldsymbol{\beta}) &= l(\boldsymbol{\beta}) - \lambda\, \boldsymbol{\beta}'\boldsymbol{\beta} \\ &= \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)\} - \lambda\, \boldsymbol{\beta}'\boldsymbol{\beta}, \end{aligned} \tag{4.3.2}$$

where π_i is as in (4.2.1). Typically, we do not want to penalize the intercept in RR, meaning that β_0 should be excluded from the penalty function. In the linear ridge regression setting this can be achieved by centering the response and all covariates, so that the mean of $\mathbf{Y}$ is zero and the mean of each covariate $X_j,\ j = 1, \ldots, p$ is zero, which results in $\hat{\beta}_0 = 0$. In the logistic regression model this is not possible, instead we could modify the penalty of the logistic ridge regression to be

$$\text{Pen}(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{K}\boldsymbol{\beta}, \tag{4.3.3}$$

where $\boldsymbol{K} = \text{diag}(0, 1, \ldots, 1)$ is a $(p+1) \times (p+1)$ penalty matrix that excludes the intercept but remains the identity matrix for the rest of the coefficient vector [Fahrmeir et al., 2013].

Maximum penalized likelihood estimates, $\widehat{\boldsymbol{\beta}}_{RR}$, are obtained by maximizing (4.3.2) with penalty given by (4.3.3), using an iterative algorithm, for example Newton-Raphson.

The score function, i.e. the first derivative of $l_{RR}(\boldsymbol{\beta})$, is

$$\begin{aligned} U_{RR}(\boldsymbol{\beta}) &= U(\boldsymbol{\beta}) - 2\lambda\boldsymbol{K}\boldsymbol{\beta} \\ &= \mathbf{X}'\left[\boldsymbol{y} - \boldsymbol{\pi}\right] - 2\lambda\boldsymbol{K}\boldsymbol{\beta}. \end{aligned} \tag{4.3.4}$$

The matrix of negative second derivatives of $l_{RR}(\boldsymbol{\beta})$, i.e. the Fisher information, is

$$\begin{aligned} I_{RR}(\boldsymbol{\beta}) &= I(\boldsymbol{\beta}) + 2\lambda \boldsymbol{K} \\ &= \mathbf{X}'\mathbf{W}\mathbf{X} + 2\lambda \boldsymbol{K}. \end{aligned} \tag{4.3.5}$$

In the above equations, $U(\boldsymbol{\beta})$ and $I(\boldsymbol{\beta})$ are the score function and Fisher information matrix respectively, from the usual logistic regression model as explained in Section 4.2.1 and $\mathbf{W}$ is a diagonal matrix with elements $\pi_i(1-\pi_i),\ i = 1, \ldots, n$.

Given a starting value $\widehat{\boldsymbol{\beta}}_{RR}^{(0)}$, the Newton-Raphson algorithm iteratively updates

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{RR}^{(t+1)} &= \widehat{\boldsymbol{\beta}}_{RR}^{(t)} + I_{RR}^{-1}\left(\widehat{\boldsymbol{\beta}}_{RR}^{(t)}\right) U_{RR}\left(\widehat{\boldsymbol{\beta}}_{RR}^{(t)}\right) \\ &= I_{RR}^{-1}\left(\widehat{\boldsymbol{\beta}}_{RR}^{(t)}\right)\left[\widehat{\boldsymbol{\beta}}_{RR}^{(t)} I_{RR}\left(\widehat{\boldsymbol{\beta}}_{RR}^{(t)}\right) + U_{RR}\left(\widehat{\boldsymbol{\beta}}_{RR}^{(t)}\right)\right] \\ &= I_{RR}^{-1}\left(\widehat{\boldsymbol{\beta}}_{RR}^{(t)}\right)\left[U\left(\widehat{\boldsymbol{\beta}}_{RR}^{(t)}\right) + \widehat{\boldsymbol{\beta}}_{RR}^{(t)} I\left(\widehat{\boldsymbol{\beta}}_{RR}^{(t)}\right)\right] \\ &= \left[\mathbf{X}'\widehat{\mathbf{W}}^{(t)}\mathbf{X} + 2\lambda \boldsymbol{K}\right]^{-1} \mathbf{X}'\widehat{\mathbf{W}}^{(t)}\left[\mathbf{X}\widehat{\boldsymbol{\beta}}_{RR}^{(t)} + \left(\widehat{\mathbf{W}}^{(t)}\right)^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\pi}})^{(t)}\right], \qquad t = 0, 1, 2, \ldots \end{aligned} \tag{4.3.6}$$

until a preselected convergence criterion is fulfilled and the maximum penalized likelihood estimator is then $\widehat{\boldsymbol{\beta}}_{RR}^{(t)}$. The variance-covariance matrix is then estimated as

$$\begin{aligned} \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}_{RR}) &= \widehat{\mathrm{Var}}\left(\left[\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X} + 2\lambda \boldsymbol{K}\right]^{-1} \mathbf{X}'\widehat{\mathbf{W}}\left[\mathbf{X}\widehat{\boldsymbol{\beta}}_{RR} + \widehat{\mathbf{W}}^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\pi}})\right]\right) \\ &= \left[\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X} + 2\lambda \boldsymbol{K}\right]^{-1} \widehat{\mathrm{Var}}\left(\mathbf{X}'\widehat{\mathbf{W}}\left[\mathbf{X}\widehat{\boldsymbol{\beta}}_{RR} + \widehat{\mathbf{W}}^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\pi}})\right]\right)\left[\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X} + 2\lambda \boldsymbol{K}\right]^{-1} \\ &= \left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X} + 2\lambda \boldsymbol{K}\right)^{-1}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X} + 2\lambda \boldsymbol{K}\right)^{-1}. \end{aligned} \tag{4.3.7}$$

If $\lambda = 0$ in (4.3.7), we see that we get the expression for the estimated variance-covariance matrix in Equation (4.2.8), in the usual logistic regression model. Furthermore, if $\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}$ is singular and thus cannot be inverted, the matrix $\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X} + 2\lambda \boldsymbol{K}$, we get by adding λ to the diagonal of $\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}$, can.

4.3.2 Choice of penalization parameter

There are a number of different ways the penalization parameter λ can be chosen. When ridge regression was introduced by Hoerl and Kennard [1970a,b], the authors interest was in finding a penalization parameter for which the mean squared error (MSE) for the coefficients in a linear ridge regression model was smaller than the MSE of the respective ordinary least squares (OLS)

estimates. Based on this, Hoerl et al. [1975] proposed the penalization parameter

$$\lambda_{HKB} = \frac{p\hat{\sigma}^2}{\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}}, \tag{4.3.8}$$

where $\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/(n-p)$, with p the number of covariates and $\widehat{\boldsymbol{\beta}}$ the OLS estimates.

Schaefer et al. [1984], introduced the 'Ridge type' estimator in the logistic ridge regression situation, with penalization parameter

$$\lambda_{SRW} = \frac{p}{\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}}, \tag{4.3.9}$$

and, following the approach of Hoerl and Kennard [1970a,b], showed that, when the covariates are collinear, it will result in coefficient estimates with smaller mean squared error than the maximum likelihood estimates. In (4.3.9), p is the number of covariates and $\widehat{\boldsymbol{\beta}}$ are the maximum likelihood estimates of the logistic regression coefficients.

The two penalization parameters above, however, are not defined when $p > n$, since neither the ordinary least squares regression coefficients nor the maximum likelihood estimates of the logistic regression model are defined in this case. As a consequence Cule and De Iorio [2013] propose a data driven method for estimating λ, based on (4.3.8) and (4.3.9), that is also valid when $p > n$. This is the method we will use to estimate the penalization parameter when doing our analyses and it is more thoroughly described in Section 4.5. The proposed method by Cule and De Iorio [2013] is closely related to principal component regression modeling and therefore the theory of principal component analysis and regression will be briefly discussed in Section 4.4.

Furthermore, visual methods have been suggested for the selection of λ. For example Hoerl and Kennard [1970a,b] introduced the 'Ridge trace', a plot of the ridge regression coefficient estimates against λ as it increases from zero, and propose choosing λ corresponding to the region on the plot at which estimates no longer change significantly as λ increases further. Cule et al. [2011] introduced the 'p-value trace', which plots the p-values of the regression coefficients against λ as λ increases from zero, and enables the visualization of the change in p-values with increasing shrinkage.

Choosing λ using cross-validation (CV) based methods is also common, here the data is partitioned into a training set for parameter estimation and an evaluation set to asses the quality of the model. In k-fold-cross-validation the data set is randomly split into k subsets (the folds) of similar size. First off the first subset is used as an evaluation set and the remaining $k-1$ subsets are together used as a training set. This is repeated k times, with each of the folds used once as evaluation set [Fahrmeir et al., 2013].

As an example, the package `penalized` [Goeman, 2010] in `R` uses likelihood CV to compare the predictive ability of different values of the penalization parameter. The CV log likelihood is calculated with default leave-one-out-CV, which is the same as k-fold-CV when $k = n$, and

n being the number of observations. Denoting the log likelihood by $l(\boldsymbol{\beta})$ and letting $l_{(-i)}(\boldsymbol{\beta})$ denote the log likelihood with the ith individual eliminated, the contribution of individual i to the log likelihood is $l_i(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - l_{(-i)}(\boldsymbol{\beta}),\ i = 1, \ldots, n$. Maximizing $l_{(-i)}(\boldsymbol{\beta})$ yields the estimates $\widehat{\boldsymbol{\beta}}_{(-1)}$. The cross validated log likelihood is then calculated as

$$\text{cvl} = \sum_{i=1}^{n} l_i \left(\widehat{\boldsymbol{\beta}}_{(-i)} \right),$$

see Van Hoewelingen and Le Cessie [1990]. The function `optL2` can be used to optimize the CV log likelihood with respect to the penalization parameter and is best used in combination with `profL2`, which profiles the CV log likelihood between two specified λ-values. The drawback with using leave-one-out-CV based methods is that it becomes computationally infeasible with increased size of the data [Cule and De Iorio, 2013].

4.4 Principal component regression

Principal component analysis (PCA) is a technique used for explaining a set of correlated variables by a reduced number of uncorrelated variables having maximal variance, where the uncorrelated variables are called principal components (PCs) [Aguilera et al., 2006].

Principal component regression refers to the use of principal components as covariates in a regression model instead of the original p variables. Since the PCs are uncorrelated, this is a possible way to deal with multicollinearity in the covariates. If all PCs are included in the regression, the resulting model is equivalent to the original one and the large variances of the estimated regression coefficients due to multicollinearity are still present. If some PCs are excluded from the regression model, retaining only the PCs which explain most of the variation, the variances of coefficient estimates can be greatly reduced, but the estimates are usually biased [Jolliffe, 2002].

For details on principal component analysis and how the principal components are constructed, see Appendix A.1. Furthermore, principal component linear regression is explained in Appendix A.2. Below we explain principal component regression in the logistic regression setting, denoted by PCLR.

4.4.1 Principal component logistic regression

In the logistic regression situation, we have a binary response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and a $n \times (p+1)$ matrix $\mathbf{X}$ of covariates, with rows $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$, where $x_{i0} = 1$ captures the intercept term. As explained in Section 4.2, we model $\pi_i = \pi(\mathbf{x}_i) = \mathrm{P}(Y_i = 1 \mid \mathbf{x}_i)$, or equivalently, the log-odds

$$\text{logit}_i = \text{logit}\left(\pi_i\right) = \sum_{j=0}^{p} x_{ij}\beta_j. \tag{4.4.1}$$

Assume that the covariates $X_1, \ldots, X_p$ have been centred and standardized, i.e. subtracted by their means and divided by their standard deviations. The log-odds (4.4.1) can now be expressed in terms of all PCs as [Aguilera et al., 2006]

$$\text{logit}_i = \text{logit}\,(\pi_i) = \sum_{k=0}^{p} z_{ik}\gamma_k, \tag{4.4.2}$$

where $z_{ik},\ i = 1, \ldots, n,\ k = 0, \ldots, p$, are the elements of the principal component matrix $\mathbf{Z} = \mathbf{XA}$ with $\mathbf{A}$ being the $(p+1) \times (p+1)$ orthogonal matrix

$$\mathbf{A} = \left(\begin{array}{c|c} 1 & \mathbf{0}' \\ \hline \mathbf{0} & A \end{array}\right)$$

where the columns of A are the eigenvectors of the matrix $\mathbf{X}'\mathbf{X}$, denoted by $\mathbf{a}_j,\ j = 1, \ldots, p$. Furthermore, $\mathbf{0}$ is a $p \times 1$ vector of zeros and $\gamma_k = \sum_{j=0}^{p} a_{jk}\beta_j,\ k = 0, \ldots, p$. For details on the construction of the PCs we refer to Appendix A.1.

In matrix form, where $\mathbf{L} = (\text{logit}_1, \ldots, \text{logit}_n)'$ denotes the $n \times 1$ vector of log-odds, we have

$$\mathbf{L} = \mathbf{X}\boldsymbol{\beta} = \mathbf{ZA}'\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\gamma}. \tag{4.4.3}$$

Therefore, the estimates of the ordinary logit model can be obtained from the estimates of the PCLR model as $\widehat{\boldsymbol{\beta}} = \mathbf{A}\widehat{\boldsymbol{\gamma}}$.

Assume that the PCs are ordered such that the first PC has the largest variance, the second PC the second largest variance and so on. In a reduced PCLR we model, in terms of the first s PCs,

$$\pi_{i,(s)} = \frac{\exp\{\sum_{j=0}^{s} z_{ij}\gamma_j\}}{1 + \exp\{\sum_{j=0}^{s} z_{ij}\gamma_j\}}, \quad i = 1, \ldots, n, \tag{4.4.4}$$

where the subscript (s) indicates how many PCs were used in the PCLR model. Equivalently, we can express the above in matrix form in terms of the vector of log-odds $\mathbf{L}_{(s)} = (\text{logit}_{1,(s)}, \ldots, \text{logit}_{n,(s)})$ with components $\text{logit}_{i,(s)} = \text{logit}(\pi_{i,(s)})$ as

$$\mathbf{L}_{(s)} = \mathbf{Z}_{(s)}\boldsymbol{\gamma}_{(s)} = \mathbf{XA}_{(s)}\boldsymbol{\gamma}_{(s)} = \mathbf{X}\boldsymbol{\beta}_{(s)}. \tag{4.4.5}$$

Thus the maximum likelihood estimation of this PCLR model will provide an estimation of the original parameters $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}_{(s)} = \mathbf{A}_{(s)}\widehat{\boldsymbol{\gamma}}_{(s)}$.

The main difference between principal component regression in the linear setting and the logistic setting is that in the latter

$$\widehat{\boldsymbol{\gamma}}_{(s)} = (\widehat{\gamma}_{0,(s)}, \widehat{\gamma}_{1,(s)}, \ldots, \widehat{\gamma}_{s,(s)})' \neq (\widehat{\gamma}_0, \widehat{\gamma}_1, \ldots, \widehat{\gamma}_s)',$$

i.e. the estimator $\widehat{\boldsymbol{\gamma}}_{(s)}$ in terms of the first s PCs is not the vector of the first s components of the estimator $\widehat{\boldsymbol{\gamma}}$ in terms of all the PCs.

4.5 Automatic choice of the ridge parameter - an approach proposed by Cule and De Iorio (2013)

Cule and De Iorio [2013] propose a data-driven method to estimate the penalization parameter which is also valid when there are more covariates than observations in the model which often is the case with genetic data. Their motivation behind the proposed method is to choose a penalization parameter so that the ridge regression model performs at least as well as, and often better than, a principal component regression model with the same degrees of freedom in prediction problems. Cule and De Iorio [2013] have implemented this method in the `R` package `ridge`, for both continuous and binary outcomes. We will make use of the package in our analyses, using the method in the logistic regression setting with binary outcomes. In Section 4.5.1 the proposed method is summarized in the logistic regression setting, following the article by Cule and De Iorio [2013]. For a summary of the method when applied in linear regression settings, see Appendix A.3. In Section 4.5.2 we explain how the significance of each parameter estimate in the ridge regression model is assessed. In Section 4.5.3 we give details of some of the `R` functions we will use.

Consider n individuals having been genotyped at p SNPs each. $\mathbf{X}$ is the covariate matrix with genotype information for each individual $i,\ i = 1, \ldots, n$, at each SNP $j,\ j = 1, \ldots, p$, meaning that $x_{ij} \in \{0, 1, 2\}$ depending on the number of a-alleles in individual i's genotype at SNP j. Each $X_j,\ j = 1, \ldots, p$, i.e. each column of $\mathbf{X}$, is centered and standardized, meaning each observation in X_j is subtracted by its column mean and divided by its standard deviation. The sample covariance matrix is given by $\mathbf{S} = \mathbf{X}'\mathbf{X}/(n-1)$, thus if we divide each observation in $\mathbf{X}$ by $\sqrt{n-1}$ we get that $\mathbf{S} = \mathbf{X}'\mathbf{X}$, where $\text{diag}(\mathbf{S}) = \mathbf{1}_p$ and $\mathbf{1}_p$ is a p-length vector of ones. We denote this as $\mathbf{S}$ being in correlation form and the off-diagonal elements describe the correlations between the SNPs, which, if squared, also give the measure r^2 of the linkage disequilibrium as described in Section 3.1.1.

$\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is the response vector of phenotypes, where $Y_i \in \{0, 1\},\ i = 1, \ldots, n$, and 1 denotes a case and 0 a control.

We also define the so called 'hat' or projection matrix $\mathbf{H}$ that relates the fitted outcomes of a regression model to the observed ones, i.e. $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$. The specific form of $\mathbf{H}$ is specified in the next section.

Note, previously we denoted the penalization parameter by λ, but here we denote it by k since λ is reserved for the eigenvalues.

4.5.1 Estimation of the penalization parameter in logistic ridge regression

The first steps of the proposed approach by Cule and De Iorio [2013] are to calculate the principal components (PCs) of $\mathbf{X}$ and the PC logistic regression (PCLR) coefficients as described in Section 4.4.1. Now, for a PCLR using $r = 1, \ldots, t$ PCs, where $t = \min(n, p)$ is the maximum

number of non-zero PCs possible, the penalization parameter k_r is computed as

$$k_r = \frac{r}{\widehat{\boldsymbol{\gamma}}'_{(r)}\widehat{\boldsymbol{\gamma}}_{(r)}}, \tag{4.5.1}$$

where $\widehat{\boldsymbol{\gamma}}_{(r)}$ is the r-dimensional vector of estimated PCLR coefficients. The next step is to calculate the effective degrees of freedom for the variance of the ridge logistic regression model, fitted using k_r as the penalization parameter. The effective degrees of freedom are here defined as $\text{tr}(\mathbf{H}'\mathbf{H})$, where $\mathbf{H}$ is the 'hat' matrix as explained previously. In logistic RR, the hat matrix is $\mathbf{H} = (\mathbf{X}'\mathbf{W}\mathbf{X} + k_r\mathbf{I})^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}$, where $\mathbf{W} = \text{diag}(\hat{\pi}_i(1-\hat{\pi}_i))$ and $\hat{\pi}_i$ are the fitted probabilities. Now the number of PCs to use in the calculation of the penalization parameter, r^*, are chosen such that the difference between r and the calculated $\text{tr}(\mathbf{H}'\mathbf{H})$ is minimized. After choosing r^*, the final step is to use the chosen penalization parameter k_{r^*} in a logistic ridge regression model fitted on the full data set. The fitting of the model is done using the CLG algorithm, a cyclic coordinate descent algorithm for penalized logistic regression. The details of the algorithm can be found in Cule and De Iorio [2013, Supplementary Appendix C].

4.5.2 Assessing significance of parameters

Cule et al. [2011] have developed a test of significance for ridge regression coefficients based on an approximation of the distribution of said coefficients under the null hypothesis $H_0 : \hat{\beta}_{j,RR} = 0$, where $\hat{\beta}_{j,RR}$ is the estimate of the jth regression coefficient in a ridge regression model. The proposed test is based on the Wald test that can be used to asses significance in multiple linear or logistic regression models. The Wald test statistic follows a Student t-distribution under $H_0 : \hat{\beta}_j = 0$, where $\hat{\beta}_j$ is the jth coefficient estimate in the multiple regression model. Asymptotically the Wald test statistic follows a standard normal distribution. The test statistic proposed by Cule et al. [2011] is

$$T_j = \frac{\hat{\beta}_{j,RR}}{\widehat{\text{SE}}(\hat{\beta}_{j,RR})}, \tag{4.5.2}$$

where $\hat{\beta}_{j,RR}$ is the estimate of the jth regression coefficient under the ridge linear or logistic regression model and $\text{SE}(\hat{\beta}_{j,RR})$ is an estimate of the standard error. In Chapter 4, Section 4.3.1, the variances of the coefficient estimates in the logistic ridge regression model are derived. Again it is assumed that under $H_0 : \hat{\beta}_{j,RR} = 0$, $T_j \overset{asy}{\sim} N(0,1)$ and the normal distribution is used to test the significance of the ridge regression coefficients.

Using simulation studies, Cule et al. [2011] compare the above approximate test to a permutation test. The permutation test is viewed as a benchmark since it gives an estimate of the null-distribution of the parameter estimates. The authors show that the performance of their test is comparable to that of a permutation test, but at a much reduced computational cost.

4.5.3 `ridge` package in `R`

Cule and De Iorio [2013] have implemented their method for choosing the penalization parameter in both linear and logistic regression settings in an `R` package called `ridge` which we will be utilizing. Since we will be fitting logistic ridge regression models we describe below the two functions that can be used for this model. For more details on the functions below and other in the package we refer to the package documentation, see Cule [2014].

`logisticRidgeGenotypes`: This function fits logistic ridge regression models for SNP data using the method by Cule and De Iorio [2013] for estimation of the penalization parameter as default. For data sets that are too large to read into `R` directly, this function provides code written in C and takes file paths, to files with the SNP and phenotype informations, as arguments. Furthermore, a 'thinning' file path is also taken as argument. This file should contain information about the SNPs name, which chromosome it is on and which position it has on said chromosome. This file is used to thin the SNP data by SNP position, meaning that at each multiple of a predefined distance, a SNP is removed. Unfortunately there is no closer information as to the size of this distance. The estimation of the penalization parameter is then based on this thinned data set, after which the ridge regression model is fitted on the whole data set. One restriction that this function has is that the genotype information has to be coded as 0, 1, 2. The function returns the coefficient estimates as default and if the argument 'verbose' is set to `TRUE` the estimated penalization parameter is returned to the `R` workspace. Furthermore, if specified, output files containing the fitted coefficients and approximated p-values for each, calculated as explained in Section 4.5.2, are returned.

`logisticRidge`: This function fits a logistic ridge regression model, with the penalization parameter estimated using the method by Cule and De Iorio [2013] as default. In contrast to `logisticRidgeGenotypes` the data is read directly into `R` and the function takes a formula of the form 'response $\sim$ covariates' as argument. In this function a restriction is put on the maximum number of PCs used for the computation of the penalization parameter; the maximum number of PCs t is such that at least 90% of the variation in the data is explained. The output from this function contains among other the fitted coefficients, their standard errors and p-values, the chosen penalization parameter and the number of PCs used to compute it.

4.6 Combining p-values in meta-analysis

In this section we introduce two methods often used in meta-analysis, to combine the results from K hypothesis tests into one: Fisher's method and Stouffer's weighted Z-score method. These will be used to combine the results from several ridge regression models in Chapter 6, the 'Results' part of this thesis.

The two procedures for combining p-values, Fisher's and Stouffer's method, both assume that the p-values come from K independent tests of the same null hypothesis, and test whether they

collectively can reject this common hypothesis [Zaykin, 2011]. In our case this null hypothesis is $H_0 : \beta_j = 0,\ j = 1, \ldots, p$ or in other words there is no association between the jth SNP and disease.

4.6.1 Fisher's method

Fisher's method combines the p-values from K independent tests into one test statistic [Whitlock, 2005]:

$$X^2 = -2 \sum_{k=1}^{K} \log(p_k), \tag{4.6.1}$$

where p_k is the p-value of the kth hypothesis test. Under the null hypothesis, X^2 follows a χ^2-squared distribution with $2K$ degrees of freedom, this can be used to estimate an overall p-value.

One major drawback with using Fisher's method is that it treats small and large p-values asymmetrically, being more sensitive to small p-values. This asymmetry can thus result in bias when combining results from multiple tests of the same null hypothesis [Whitlock, 2005].

Fisher's method is implemented in the `R` package `MADAM`, in the function `fisher.method`, which as argument takes a matrix or data frame containing the p-values from the single tests and returns the value of the test statistic X^2, the number of p-values used to calculate X^2 and an overall p-value.

4.6.2 Stouffer's weighted Z-score method

Stouffer's Z-score method combines the p-values from K independent tests into one test statistic [Whitlock, 2005]

$$Z_S = \frac{\sum_{k=1}^{K} Z_k}{\sqrt{K}}, \tag{4.6.2}$$

where $Z_k = \Phi^{-1}(1-p_k)$, Φ and Φ^{-1} denote the standard normal cumulative distribution function and its inverse respectively, and p_k is the p-value of the kth hypothesis test. Under the null hypothesis, Z_S then follows a standard normal distribution, which can be used to calculate an overall p-value. Furthermore, one can introduce weights to the Z-score, the weighted Z-score then takes the form [Whitlock, 2005]

$$Z_w = \frac{\sum_{k=1}^{K} w_k Z_k}{\sqrt{w_k^2}}, \tag{4.6.3}$$

and, under the null hypothesis, Z_w follows a standard normal distribution. Zaykin [2011] recommends using either the square root of the study size as weight, i.e. $w_k = \sqrt{n_k}$, where n_k is

the total number of individuals in study k, or the inverse of the standard error of the estimated coefficient, i.e. $w_k = 1/\widehat{\sigma}_k,\ k = 1, \ldots, K$.

Note that Stouffer's method assumes that the individual p-values are one-sided, while our resulting p-values are two-sided. Since the test statistic we used for calculating the ridge regression p-values is assumed to be standard normal, we can use the symmetry of the normal distribution to convert our two-sided p-values to one-sided p-values prior to combining, by

$$p_{one-sided} = \begin{cases} p_{two-sided}/2 & \text{if } H_1 : \beta_j > 0,\ j = 1, \ldots, p \\ 1 - p_{two-sided}/2 & \text{if } H_1 : \beta_j < 0,\ j = 1, \ldots, p. \end{cases}$$

Here there are two possible one-sided p-values depending on the alternative hypothesis, i.e. either $H_1 : \beta_j > 0$ or $H_1 : \beta_j < 0$ for $j = 1, \ldots, p$. Due to the symmetry of the normal distribution the direction of the alternative hypothesis can be chosen arbitrarily, as long as it is the same H_1 for all studies. Once these one-sided p-values are combined, the result can be converted back to two-sided as [Whitlock, 2005]

$$p_{two-sided} = \begin{cases} 2p_{one-sided} & \text{if } p_{one-sided} < 1/2 \\ 2(1 - p_{one-sided}) & \text{otherwise.} \end{cases}$$

Chapter 5

Data Description & Initial Analysis

The case-control studies investigated, as mentioned in the introduction, originate from studies participating in BCAC [Breast Cancer Association Consortium], as part of the Collaborative Oncological Gene-Environment Study [COGS] which aims to evaluate genetic variants associated with risk of breast, ovarian and prostate cancer. Genotyping was conducted in the same manner for all studies, using the iCOGS array. It comprises around 200 000 SNPs, where standard quality control was performed and SNPs with a deviation from Hardy-Weinberg equilibrium (HWE) at significance level 10^{-5} were excluded [COGS].

The data analyzed in this thesis was provided by Karolinska Institutet and consists of 89 050 individuals phenotype information, i.e. if the individual had breast cancer or not, and his/her genotype information at each of 3 279 SNPs, as well as the individuals study affiliation, where an individual can belong to one of 39 case-control studies of varying sizes participating in BCAC. The genotype information comes from SNPs across a genomic region that spans over 990 kb (kilo-base = unit of length for DNA equal to 1 000 nucleotides) on a specific chromosome. The 89 050 individuals consist of 42 600 controls and 46 450 cases in total, and are all Europeans. Prior to us receiving the data, missing genotypes had been imputed with Impute2 [Howie et al., 2009] using the worldwide 1000 Genomes Project variant data as reference [The 1000 Genomes Project Consortium, 2010]. Thus the received data set contains both genotyped and imputed SNPs. Our data set contains 2 765 imputed SNPs and thus 514 genotyped SNPs. Other epidemiological variables where available but their information considered as classified and were thus not included in the data set provided for the analysis of this thesis. As mentioned, the 39 case-control studies the data originates from are of varying sizes, the smallest study consisting of around 100 individuals and the largest of around 17 000 individuals, with about 50% cases and 50% controls in each of the 39 studies. Figure 5.1 shows a histogram over the study sizes of the 39 studies, where 28 studies out of the 39 studies consist of under 2 000 individuals each, while 8 studies consist of between 2 000 and 5 000 individuals. Two studies are large, consisting of around 10 000 and 17 000 individuals respectively. Since all individuals are considered to come from the same population, it is assumed that the SNP effects are the same between studies.

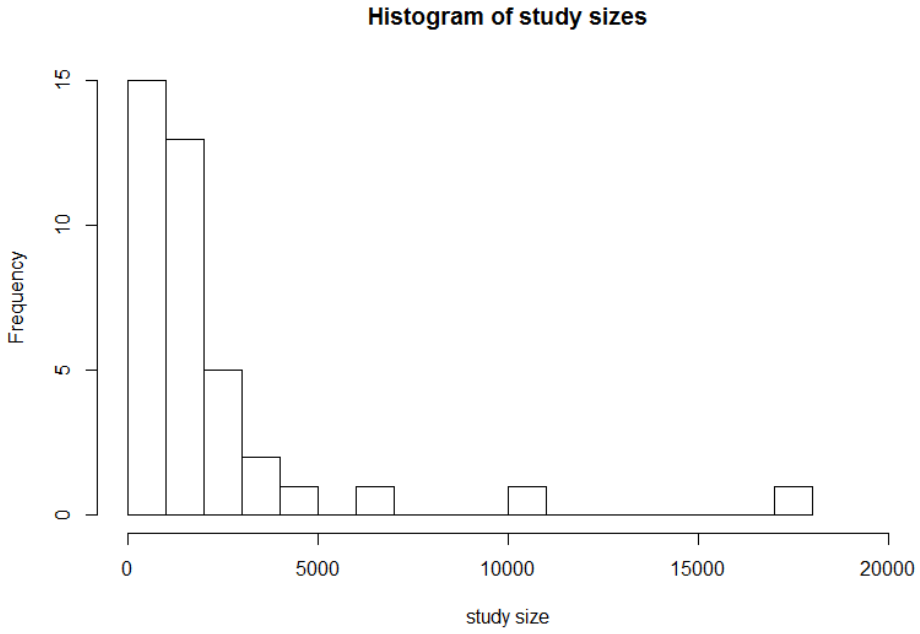

Figure 5.1: Histogram of the study sizes for the 39 case-control studies investigated in this thesis.

5.1 Variable coding

In our data an individual's observed phenotype y_i, $i = 1, \ldots, 89\ 050$, is coded as $y_i = 1$ if individual i is a breast cancer case and $y_i = 0$ if not.

Individual i's genotype at a specific genotyped SNP j, $j = 1, \ldots, 3\ 279$, is in our data set coded as $x_{ij} = 0$ for genotype AA, $x_{ij} = 1$ for genotype Aa or $x_{ij} = 2$ for genotype aa. If individual i's genotype at SNP j has been imputed, the missing genotype is replaced in the style of allele dosage, as defined below, and thus $x_{ij} \in [0, 2]$. Altogether, the design matrix $\mathbf{X}$ is thus the $n \times p$ matrix with observed and imputed genotype information, where $n = 89\ 050$ and $p = 3\ 279$.

Let us take a closer look at the genotype information of an individual provided by the imputation algorithm: Let genotype G be a random variable, describing the genotype of a specific individual at a specific SNP, taking values in $\{0, 1, 2\}$. The imputation algorithm provides probabilities $\mathrm{P}(G = x)$ for each of the three possible genotypes, $x = 0, 1, 2$, at each SNP and for each individual, where as before $x = 0$ denotes genotype AA, $x = 1$ denotes genotype Aa and $x = 2$ denotes genotype aa. One option is to use the most probable genotype, where the genotype is chosen as the one for which the imputed probability is the largest and over a pre-specified threshold $\xi > 0$ [Marchini and Howie, 2010], i.e.

$$G = \underset{\{x \in \{0,1,2\}:\ \mathrm{P}(G=x) \geq \xi\}}{\arg\max} \mathrm{P}(G = x).$$

If the imputed probability is not over the pre-specified threshold, the SNP is removed from the data set. In our data set the genotype information at imputed SNPs is completed in the style

of allele dosage. The allele dosage is also known as the mean genotype and is defined as

$$\mathrm{E}[G] = \sum_{x=0}^{2} x\mathrm{P}(G = x) = \mathrm{P}(G = 1) + 2\mathrm{P}(G = 2),$$

and takes values between 0 and 2 [Marchini and Howie, 2010].

5.2 Correlation between SNPs

In our data the 3 279 SNPs are denoted by SNP1, SNP2,..., SNP3279 and are positioned on the chromosome in that order, making it easy to keep track of which SNPs are in close proximity to each other in reality. Linkage disequilibrium (LD) is, in genetic association studies, used to describe the correlation between SNPs. The most common measure of LD between two SNPs is, as described in Section 3.1.1, the square of the Pearson correlation coefficient, denoted by r^2. Note that r^2 ranges from 0 to 1, where the correlation between the SNPs increases as r^2 approaches 1. Figure 5.2 shows a heatmap of the r^2 measures between all of the 3 279 SNPs in our data set. The SNPs are ordered such that SNP1 is in the bottom left corner of the heatmap and SNP3279 in the upper right corner. We see that there are regions where the correlations between SNPs are very high, i.e. close to 1. This is for example the case in the region from around SNP500 up to SNP850 and from around SNP1900 up to around SNP2600. The heatmap is created using the package `LDheatmap` in `R` [Shin et al., 2006].

5.3 Single-SNP association with disease

To asses the association of each SNP individually with breast cancer risk we fit for each of the 3 279 SNPs seperately a logistic regression model. Thus for each SNP j, $j = 1, \ldots, 3\,279$, we model

$$\mathrm{logit}(\pi_i) = \beta_0^{(j)} + \beta_1^{(j)} x_{ij}, \tag{5.3.1}$$

where $\pi_i = P(Y_i = 1 \mid x_{ij})$ and $i = 1, \ldots, 89\,050$. The influence of $\beta_1^{(j)}$, i.e. of an association of the jth SNP with disease, is investigated using the likelihood ratio test between the above model and the null model with only an intercept, which has a χ^2-distribution with 1 degree of freedom. Furthermore, the minor allele frequencies among the controls are calculated for each SNP.

Table 5.1 shows the result for the 10 most significant SNPs, which we will denote by 'Top 10', from such an analysis, with estimated coefficients, standard error and likelihood ratio test p-value, their estimated odds ratios (OR) for disease when genotype status is 1 (Aa) compared to 0 (AA), the corresponding 95% confidence interval (CI) and the minor allele frequency (MAF) in controls.

Of the 3 279 SNPs, 326 have a significant association with disease at the significance level 10^{-5},

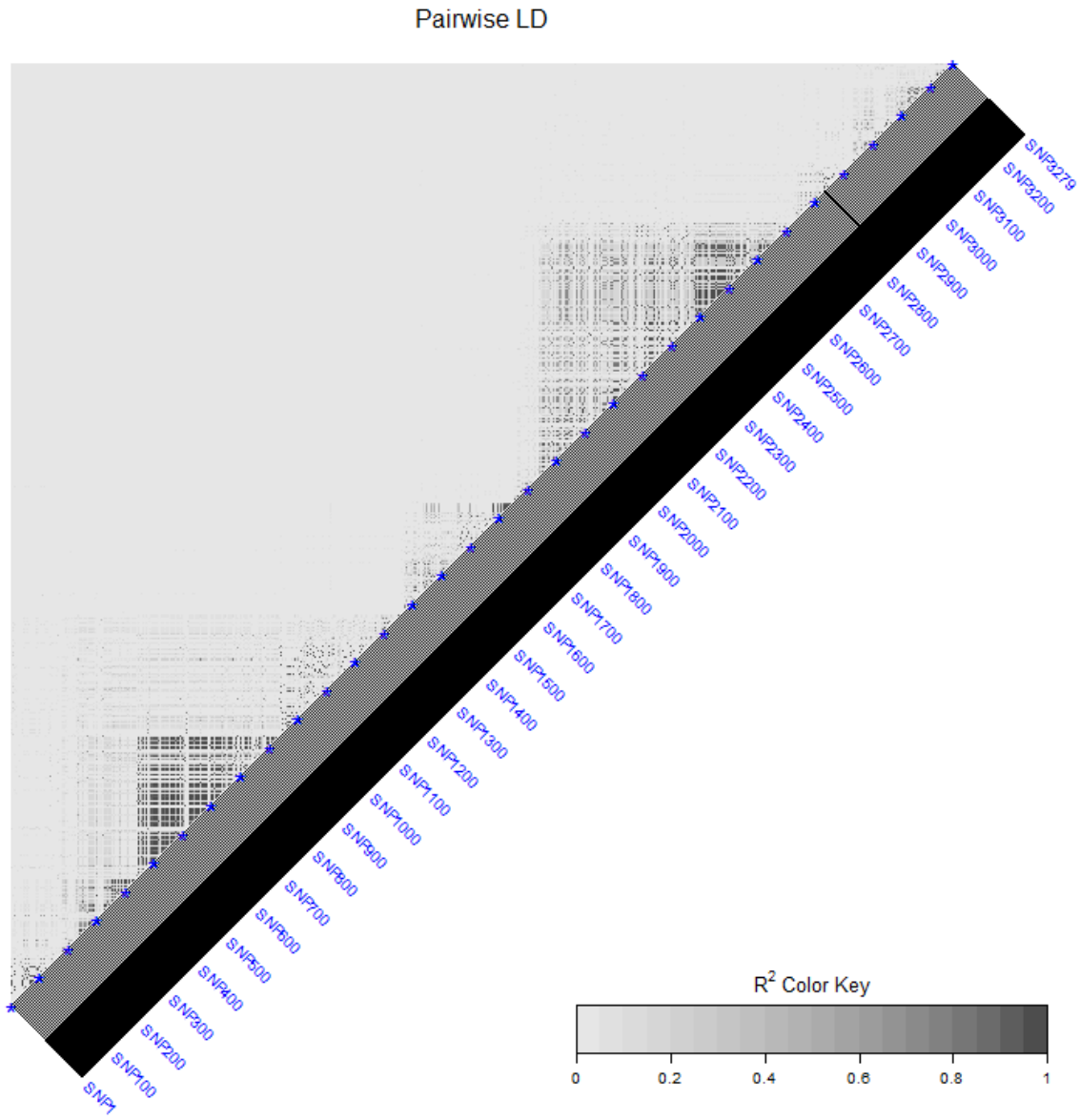

Figure 5.2: Heatmap of pairwise linkage disequilibrium r^2 measurements for the 3 279 SNPs in our data set. The SNPs are ordered such that SNP1 is in the bottom left corner of the heatmap and SNP3279 in the upper right corner.

SNP	LR test p-value	Wald test p-value	$\widehat{\beta_1}$	$\widehat{\text{SE}}$	$\widehat{\text{OR}}$ (95% CI)	MAF
SNP2540	8.42E-15	8.51E-15	-0.083	0.011	0.92 (0.90, 0.94)	0.292
SNP2429	1.24E-14	1.25E-14	-0.083	0.011	0.92 (0.90, 0.94)	0.305
SNP2580	5.01E-14	5.05E-14	-0.078	0.010	0.92 (0.91, 0.94)	0.301
SNP2584	6.43E-14	6.49E-14	-0.079	0.010	0.92 (0.91, 0.94)	0.301
SNP2479	6.74E-14	6.79E-14	-0.078	0.010	0.92 (0.91, 0.94)	0.301
SNP2519	6.74E-14	6.80E-14	-0.078	0.010	0.92 (0.91, 0.94)	0.301
SNP2501	7.12E-14	7.18E-14	-0.078	0.010	0.93 (0.91, 0.94)	0.301
SNP2502	7.17E-14	7.23E-14	-0.078	0.010	0.92 (0.91, 0.94)	0.300
SNP2572	7.47E-14	7.53E-14	-0.079	0.011	0.92 (0.90, 0.94)	0.282
SNP2496	7.66E-14	7.72E-14	-0.078	0.010	0.93 (0.91, 0.94)	0.301

Table 5.1: Univariable logistic regression results for the 10 most significant SNPs.

and all of these SNPs have a minor allele frequency $> 2\%$ among the controls. As can be seen the OR for all of the Top 10 SNPs is close to one, but slightly below, meaning that for example for SNP2540, the odds for disease among the heterozygote (Aa) individuals is 0.92 times the odds for disease among the homozygote wildtype (AA) individuals in that SNP. We see that the Top 10 SNPs all are from the same region, as indiciated by their position numbering (SNP25xx), and their OR estimates and their confidence intervals are all very similar, suggesting strong LD in this region.

We also investigate if inclusion of the individual's study affiliation as covariate in each of the above regression models influences the model significantly. The study affiliation is included as a factor meaning that the model for each SNP $j,\ j = 1, \ldots, 3\,279$, can be written as

$$\text{logit}(\pi_i) = \beta_0^{(j)} + \beta_1^{(j)} x_{ij} + \gamma_2^{(j)} \text{study}_{i,2} + \gamma_3^{(j)} \text{study}_{i,3} + \ldots + \gamma_{39}^{(j)} \text{study}_{i,39}, \tag{5.3.2}$$

where

$$\text{study}_{i,k} = \begin{cases} 1 & \text{if individual } i \text{ belongs to study } k,\ k = 2, \ldots, 39 \\ 0 & \text{otherwise} \end{cases}$$

and study 1 is the reference group. We compare model (5.3.2) to model (5.3.1) by performing a likelihood ratio test between the models, testing the null hypothesis $H_0 : \gamma_2 = \gamma_3 = \ldots = \gamma_{39} = 0$ against the alternative H_1 : at least one $\gamma_k \neq 0, k = 2, \ldots, 39$. The test has a χ^2-distribution with 38 degrees of freedom and is significant, with p-values of magnitude 10^{-16}, for all of the 3 279 tests, indicating that there is a significant difference in effect between the studies for each of the 3 279 SNPs. On the other hand, it turns out that when study affiliation is included as a dummy-coded covariate, the estimated odds ratios either do not differ from the estimates in Table 5.1 at all or the difference is only visible from the third decimal and onwards. Furthermore, the Wald p-values for the β_1 coefficients differ approximately with a factor 10 between the two models for all of the SNPs, the p-value being smaller in the model without study as covariate. This has as consequence that 293 SNPs have a significant association with disease at significance

level 10^{-5} when the univariable analysis of each SNP is done with the inclusion of study as covariate, in contrast to the 326 significant SNPs when not including study as covariate. Since the different case-control studies are, as mentioned before, assumed to be samples of the same population and since the interest at this stage is to identify a SNP or a region of SNPs that are deemed as having an association with disease and not the specific effect estimates, we decide to make further analyses without including an individual's study affiliation as covariate. Hence, the SNPs that will be analyzed in the next chapter using multivariable logistic regression and ridge regression will be the 326 SNPs having a significant association with disease at significance level 10^{-5} when analyzed using the univariable model (5.3.1) without study as covariate.

Figure 5.3 shows a heatmap of pairwise linkage disequilibrium r^2 measurements for the 326 SNPs identified from the above mentioned univariable analyses, i.e. the ones having a univariable p-value below 10^{-5} when modelled using (5.3.1). We see that there seem to be three regions where the SNPs are in very high linkage disequilibrium, indicated by the triangle shapes. Note that not all SNPs are denoted by name in the plot, due to lack of space. The 326 SNPs are sorted according to their number and thus according to their actual positions on the specific chromosome. All significant SNPs are thus between SNP1834 and SNP2712.

Note here that these 326 SNPs are chosen based on their p-values, this is one possible method of filtering SNPs to reduce the initial set of SNPs down to a much smaller group in which it is probable that the SNPs having a true association with disease, or being in linkage disequilibrium (LD) with the true causal SNP, remain. Spencer et al. [2014] mention a number of other methods for filtering SNPs, among other the relative likelihood filter, which includes ranking the SNPs according to their (log) likelihood and retaining those SNPs that have likelihoods within a prespecified ratio of the highest likelihood, and some other filters based on the LD structure of the SNPs.

In summary, the univariable analysis points to an association with disease at SNPs coming from roughly the same region, where the SNPs are highly correlated. Furthermore, the inclusion of individual's study affiliation as a factor in the univariable analyses does yield significant likelihood ratio tests when compared to the model without study affiliation; however the SNP effects were only changed marginally. Therefore, we will disregard study affiliation as covariate in further analyses in this thesis.

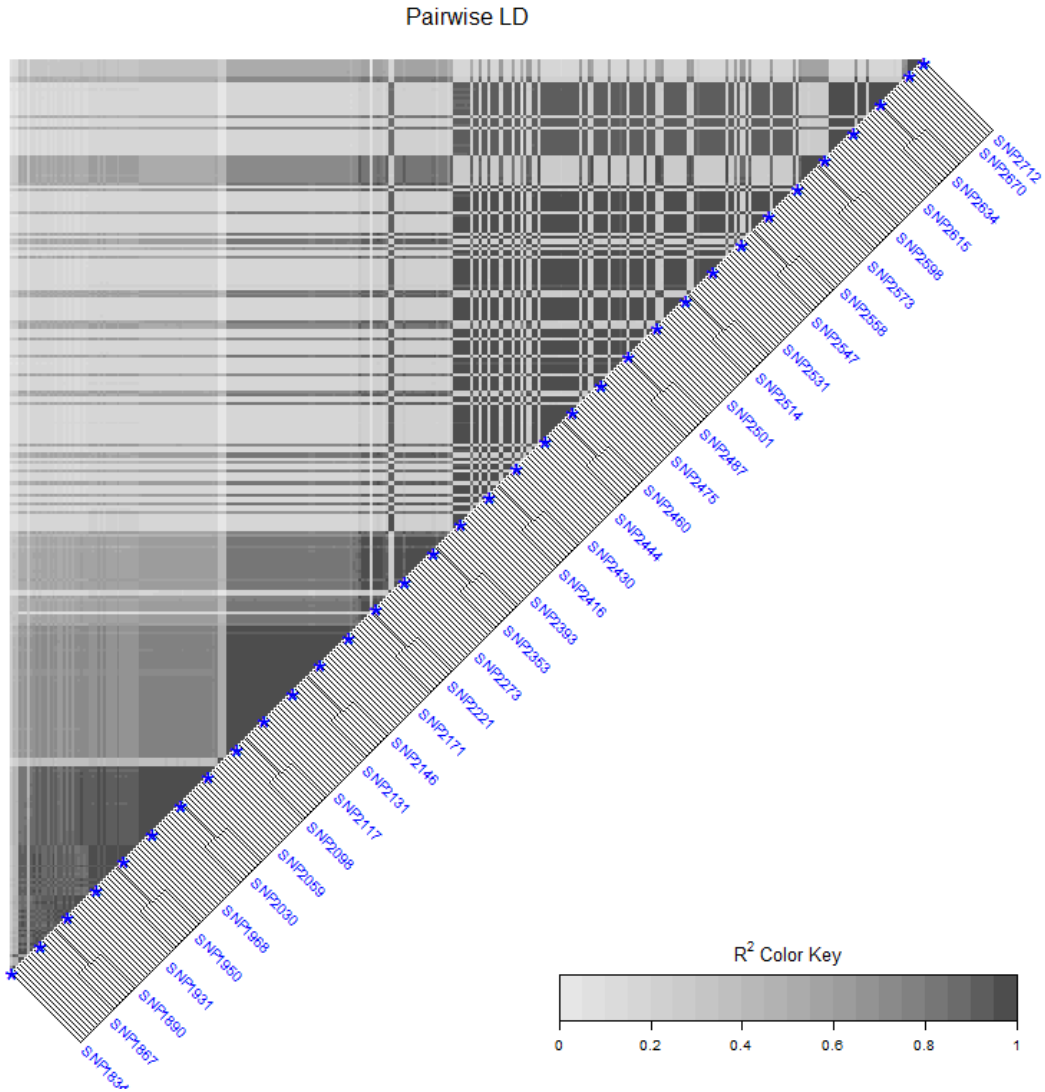

Figure 5.3: Heatmap of pairwise linkage disequilibrium r^2 measurements for the 326 SNPs having a univariable p-value below 10^{-5}. Note that all SNPs are not labelled, some SNP names are displayed to give an indication of which region we are in.

Chapter 6

Results

In this chapter the results of the statistical analysis of the breast cancer data set described in Chapter 5 are presented and described. In Section 6.1, we fit multivariable logistic regression models to our data, using the SNPs as covariates that fulfill two criterions: The first criterion is that the SNP has a minor allele frequency $> 2\%$ among the controls, and the second is that the SNP showed an association with disease at significance level 10^{-5} in the univariable analyses done in Chapter 5. The 326 SNPs that fulfill these two criterions are thus included in a logistic regression model, where variable selection is done with forward selection at three different significance levels, $\alpha = 10^{-3}$, 10^{-7} and GWAS level 5×10^{-8}. In Section 6.2, the same 326 SNPs are included as covariates in a logistic ridge regression model, where the method by Cule and De Iorio [2013] is used for estimation of the penalization parameter. As mentioned before, the data comprises individuals from 39 different case-control studies of varying sizes. The ridge regression model was initially fit on all individuals at once, but due to convergence issues when using the method by Cule and De Iorio [2013], the data has to be split up into 39 subsets, one for each study and a logistic ridge regression model was fitted on each subset, again using the method by Cule and De Iorio [2013] to estimate the penalization parameter for each model. The resulting p-values for the coefficients from each of the 39 models were then combined into one p-value per covariate using Fisher's method and Stouffer's weighted method for combining p-values, see Section 4.6.

The aim of the analysis is to compare the results from the multivariable logistic regression analysis with the results from the logistic ridge regression analysis, by investigating if the same SNPs or SNPs coming from the same genomic regions are selected as having an association with disease.

6.1 Multivariable logistic regression

The 326 SNPs that fulfill the inclusion criterions: Has a minor allele frequency in controls $> 2\%$ and showed an association with disease at significance level 10^{-5} in the univariable analyses done in Section 5.3; are included as covariates in a multivariable logistic regression model where

we choose which of the 326 SNPs to retain trough a forward likelihood ratio inclusion criterion at three different significance levels. Forward selection is done since many of the 326 SNPs are correlated, as we saw in Chapter 5 Figure 5.3, and inclusion of them all would lead to parameter estimates with very high variances. By doing variable selection we want to find the SNPs that are the most 'important', i.e. that are deemed as having the strongest association with disease, that may contribute together or independently to disease.

In the forward selection procedure, explained in Section 4.2.2, a covariate (SNP) is included in the model if the likelihood ratio test of the test $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ rejects the null hypothesis at the given level of significance. We conduct this procedure at three different significance levels $\alpha = 10^{-3}$, 10^{-7} and GWAS level 5×10^{-8}. The model fitting is done with the `step` function in `R`, where the significance level for inclusion can be adjusted by changing the argument `k` of the function. The resulting models include 2 SNPs when using level $\alpha = 10^{-3}$ for inclusion and 1 SNP when using $\alpha = 10^{-5}$ or GWAS level 5×10^{-8}. Table 6.1 and Table 6.2 show summaries of the results.

SNP	p-value	$\widehat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	$\widehat{\text{OR}}$ (95% CI)
SNP2540	4.41E-08	-0.443	0.081	0.64 (0.55,0.75)
SNP2603	7.20E-06	0.354	0.079	1.42 (1.22,1.66)

Table 6.1: Resulting forward selection model, when likelihood ratio inclusion is done at significance level $\alpha = 10^{-3}$.

We see that with forward likelihood ratio inclusion at significance level $\alpha = 10^{-3}$, two SNPs are included in the model, the first one being among the Top 10 significant SNPs from the univariable analyses in Chapter 5. Looking at the linkage disequilibrium (LD) r^2 measurement between these two SNPs we see that they are in very highly LD with an r^2 of 0.98.

SNP2540 has, as in the univariable analysis, an OR below one, but in the model after the forward procedure substantially lower. When keeping the other covariate, SNP2603, fixed, the odds for disease among the individuals with genotype Aa is 0.64 times the odds for disease among the ones with genotype AA at SNP2540, while the odds for disease among individuals with genotype aa is $\exp(2 \times -0.443) = 0.413$ times the odds for individuals with genotype AA. SNP2603, on the other hand, has an OR estimate above one, meaning that the odds for disease among the individuals with genotype Aa is 42% higher than among those with genotype AA at SNP2603, while keeping the other covariate fixed. This also means that the odds for disease for individuals with genotype aa is $\exp(2 \times 0.354) = 2.029$ times the odds for individuals with genotype AA at SNP2603, when keeping the other covariate fixed.

SNP	p-value	$\widehat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	$\widehat{\text{OR}}$ (95% CI)
SNP2540	8.51E-15	-0.083	0.011	0.92 (0.90, 0.94)

Table 6.2: Resulting forward selection model, when likelihood ratio inclusion is done at significance level $\alpha = 10^{-7}$ or GWAS significance level $\alpha = 5 \times 10^{-8}$.

When decreasing the significance level for foward inclusion, to 10^{-7} or GWAS level 5×10^{-8},

only the SNP with the smallest univariable p-value is selected after the forward selection procedure; SNP2540, resulting in the same logistic regression model as described in the context of univariable analyses in Chapter 5, where the odds for disease among individuals with genotype Aa is 0.92 times the odds among individuals with genotype AA at SNP2540.

In Figure 6.1, $-\log_{10}$ of the p-values of the univariable analyses for each of the 3 279 SNP are plotted. Furthermore, the chosen SNPs from the forward selection procedure, at significance level $\alpha = 10^{-3}$, are marked in the figure. Figure 6.2 shows the $-\log_{10}$ of the p-values of the univariable analyses for each of the 3 279 SNPs together with the forward selection SNP chosen at significance level 10^{-7} and GWAS level 5×10^{-8}. All SNPs with p-value above the dotted line were included in the forward selection procedure. The figures also shows which SNPs were imputed. Recall that the numbering of the SNPs describes their position on the specific chromosome in relationship to each other, meaning for example that SNP2000 is positioned in between SNP1999 and SNP2001.

6.2 Logistic ridge regression

Using the 326 SNPs, fulfilling the two criterions as explained above, as covariates, a logistic ridge regression model is fitted using the function `logisticRidgeGenotypes` from the `ridge` `R` package [Cule and De Iorio, 2013]. This function does not read the data directly into `R`, as explained in Section 4.5.3, as it uses an underlying C function as work horse. However, when using `logisticRidgeGenotypes` to fit the ridge regression model using all individuals, the method by Cule and De Iorio [2013] for estimating the penalization parameter does not converge and hence this approach does not give any results. Instead we use the function `logisticRidge` from the same package which performs the same model fitting and estimates the penalization parameter in the same manner, but reads the data set directly into `R`. However, as a consequence we have to split up our data set due to its size. The difference between the two functions with regard to the penalization estimation, is that `logisticRidge` sets the maximum number of principal components r, used in the calculation of the penalization parameter, such that these first r principal components together explain at least 90% of the variation in the data. See Section 4.5 for further details on the estimation of the penalization parameter.

As mentioned before, the data comes from 39 different case-control studies. We thus split the data into these 39 subsets and fit to each subset a logistic ridge regression model using `logisticRidge`, which implements the method of Cule and De Iorio [2013] to estimate the penalization parameter.

Since the method for estimating the penalization parameter by Cule and De Iorio [2013] is applied once for each study, all the resulting ridge regression models have been fitted using different estimated penalization parameters. The estimated penalization parameters vary quite substantially between studies, from $\hat{\lambda} = 3.74$ up to $\hat{\lambda} = 22785.26$, with an overall average value at $\bar{\lambda} = 1382.192$ and standard error $\text{SE}(\hat{\lambda}) = 4406.598$. Due to the different penalization it is hard to compare anything other than the p-values for each coefficient between the separate

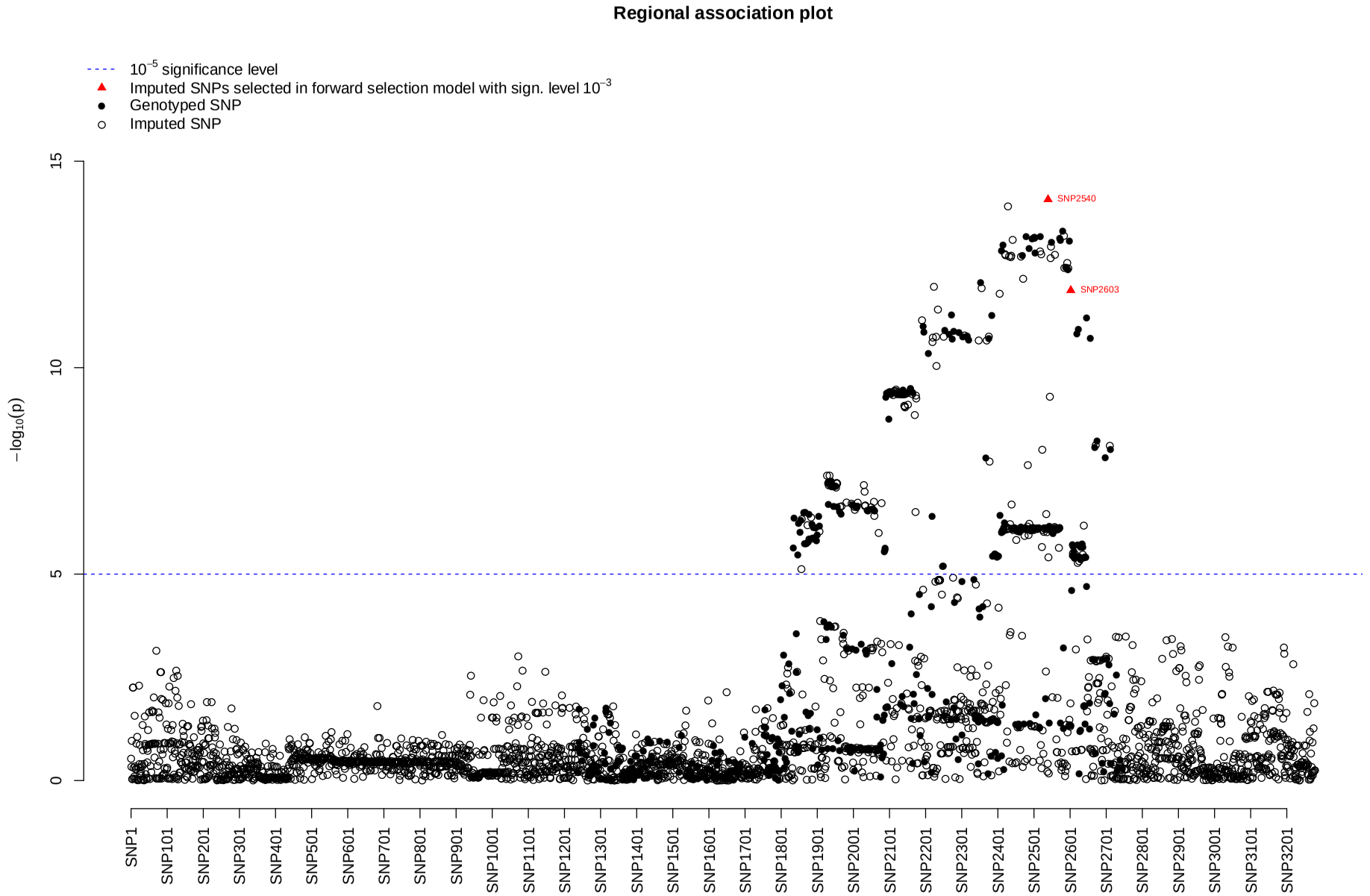

Figure 6.1: $-\log_{10}$ p-values from univariable logistic regression for all SNPs. The two SNPs indicated with a triangle, were included in the resulting model from a forward selection procedure, where significance level 10^{-3} was used for inclusion.

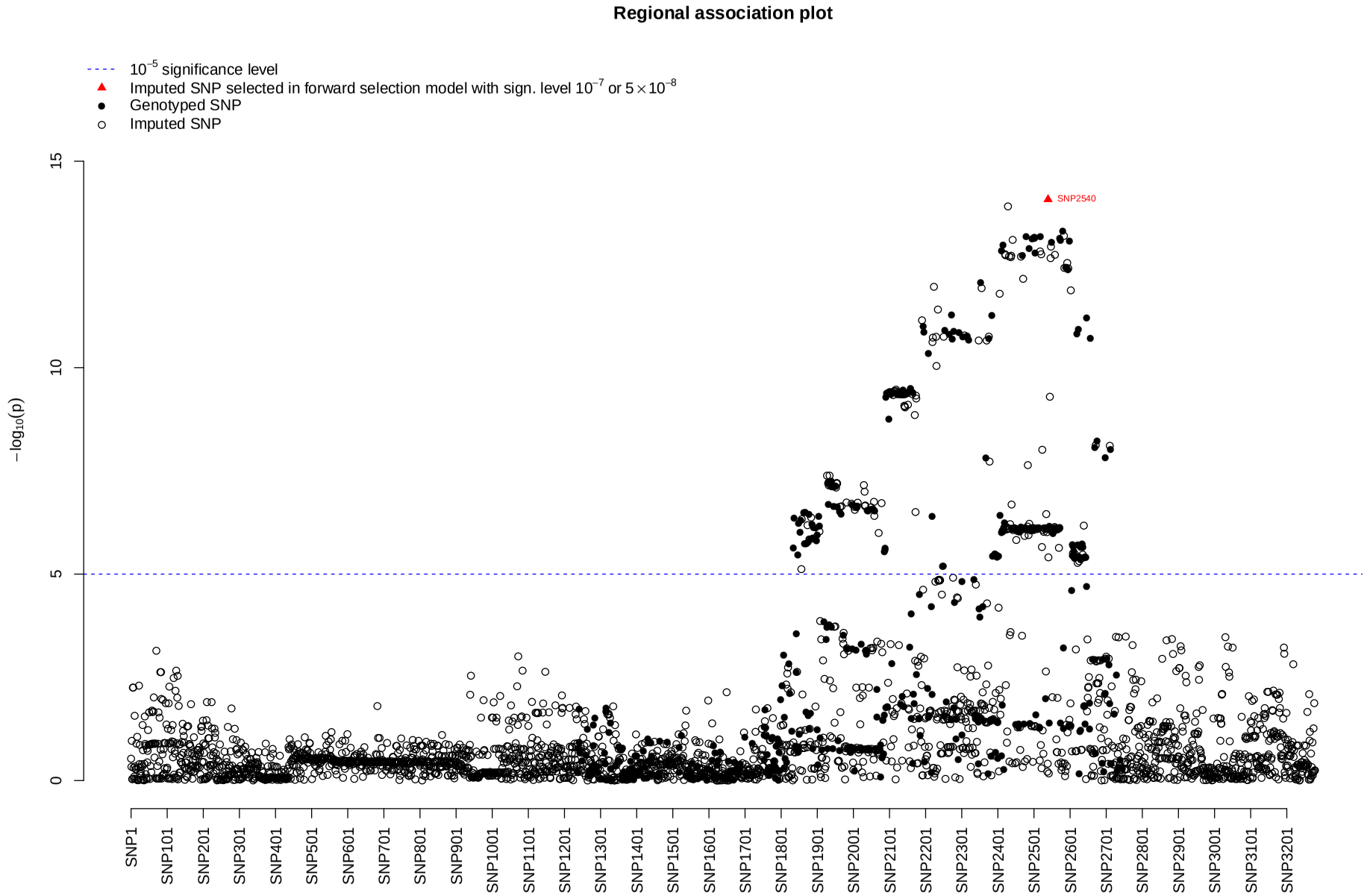

Figure 6.2: $-\log_{10}$ p-values from univariable logistic regression for all SNPs. The SNP indiciated with a triangle was included in the resulting model from a forward selection procedure when using signficance level 10^{-7} or level 5×10^{-8} for inclusion.

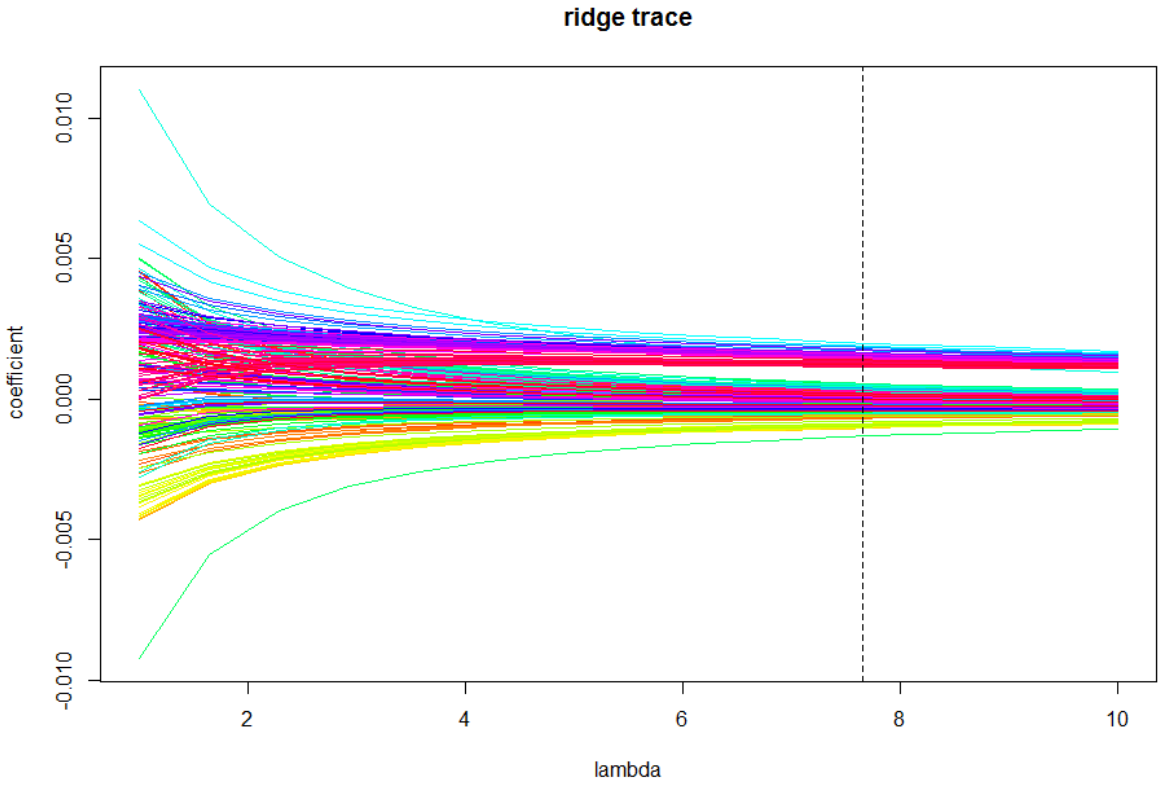

Figure 6.3: Ridge trace plot of the estimated ridge regression coefficients against the penalization parameter λ used. The dashed vertical line indicates the λ chosen by the method by Cule and De Iorio [2013].

models. However, since our primary interest at this stage is to identify which SNPs that show strong association with disease and not their individual effects, this hopefully does not pose an inconvenience. Recall that the penalization parameter controls the degree of shrinkage towards zero of the parameter estimates. To glimpse this shrinkage effect of the ridge penalization, we look closer at one of the 39 studies in our data set, comprising approximately 1000 individuals. Figure 6.3 shows a trace plot of the ridge regression coefficient estimates against the penalization parameter λ as it increases from zero. The vertical dashed line indicates the penalization parameter chosen using the method by Cule and De Iorio [2013]. We see that the coefficients are shrunken towards zero as the penalization parameter increases and the chosen λ is in a region where the coefficient estimates do not change much anymore. Note however the y-axis of the plot, indiciating that the coefficients here are really small regardless the choice of λ.

The ridge regression analysis of the 39 case-control studies took approximately three days to run in `R`.

As we are interested in a result based on all studies combined, we combine the resulting p-values for each coefficient from each of the 39 ridge regression models using two different methods: Fisher's method and Stouffer's weighted Z-score method, as explained in Section 4.6. In the calculation of Stouffer's weighted Z-score we deem weights $w_k = \sqrt{n_k}$, i.e. weights equal to the square root of study size, as being the best choice since, the 39 ridge regression models all have

been fitted with different estimated penalization parameters, making them hard to compare with regard to effect size or their standard error. Furthermore, the sizes of the studies differ quite substantially, from the smallest study with around 100 individuals to the largest with around 17000 individuals, and it thus seems reasonable to take the study sizes into account.

The two procedures for combining p-values, Fisher's and Stouffer's method, both assume that the p-values come from K independent tests of the same null hypothesis, and test whether they collectively can reject this common hypothesis [Zaykin, 2011]. In our case this null hypothesis is $H_0 : \beta_j = 0,\ j = 1, \ldots, p$, or in other words there is no association between the jth SNP and disease in any of the studies and we test against the alternative hypothesis $H_1 : \beta_j \neq 0,\ j = 1, \ldots, p$.

Table 6.3 shows the 10 SNPs with the most significant associations with disease after pooling of the results of the $K = 39$ studies using Fisher's method and the weighted Z-score method, where the weights for each study are chosen as the square root of the study size, i.e. $w_k = \sqrt{n_k}$, where n_k denotes the study size of study $k,\ k = 1, \ldots, K$.

Fisher's method		Stouffer's weighted method	
SNP	p-value	SNP	p-value
SNP2429	7.97E-05	**SNP2429**	6.53E-14
SNP2416	2.85E-04	SNP2540	1.93E-13
SNP2548	3.04E-04	**SNP2416**	8.33E-13
SNP2518	3.09E-04	**SNP2572**	1.01E-12
SNP2421	3.13E-04	**SNP2575**	1.07E-12
SNP2438	3.25E-04	**SNP2518**	1.13E-12
SNP2572	3.45E-04	**SNP2548**	1.19E-12
SNP2465	3.58E-04	**SNP2421**	1.22E-12
SNP2575	3.59E-04	SNP2550	1.39E-12
SNP2422	3.71E-04	**SNP2422**	1.53E-12

Table 6.3: Overall p-values for the 10 most significant SNPs after combination of the $K = 39$ logistic ridge regression models using Fisher's and Stouffer's weighted method. Stouffer's method is used with weights $w_k = \sqrt{n_k}$, where $k = 1, \ldots, K$ and n_k is the study size of study k. The SNPs highlighted with bold text are among the 10 most significant regardless of which method was used to combine. The SNP in the highlighted cell had the highest association with disease in the forward logistic regression models.

We see that SNP2429 has the highest significance after combination, regardless of which method is used. When using Stouffer's method with weights $w_r = \sqrt{n_r}$, SNP2540 is ranked second. This SNP is also the one that had the smallest univariable p-value, as calculated in Section 5.3, and which was included in each of the multivariable logistic regression models with forward selection. All in all, when using weights equal to the square root of the study sizes, 326, 77 and 75 SNPs have a significant association with disease at significance levels 10^{-3}, 10^{-7} and GWAS level 5×10^{-8} respectively. Note that at significance level 10^{-3}, all SNPs included in the model are significantly associated with disease.

In Figure 6.5 and Figure 6.4 $-\log_{10}$ of the univariable p-values are plotted, as estimated in Section 5.3, and the 10 most significant SNPs of the ridge regression analysis after combination of the p-values with Fisher's and Stouffer's method respectively, are marked. We see that the ten most significant SNPs from the ridge regression analysis are all in the same region of the plot, meaning they are also in the same genomic region.

Note that 8 of the 10 most significant SNPs in Table 6.3 are the same regardless of which of the two methods was used for combination. Since we just saw that all of these are from the same genomic region we investigate the correlation between them. Table 6.4 shows the matrix of r^2 measures between these 8 SNPs - recall that r^2 is a measure of linkage disequilibrium between 0 and 1 and approaches 1 when the correlation between SNPs increases, see Section 3.1.1. We see that all elements of the matrix are close to one or equal to one, meaning that the SNPs are very highly correlated, suggesting that they are detecting the same disease association. SNP2429 is the SNP with the 'smallest' r^2 with the other SNPs, but has still high values at around 0.94.

	SNP2416	SNP2421	SNP2422	SNP2429	SNP2518	SNP2548	SNP2572	SNP2575
SNP2416	1							
SNP2421	0.99	1						
SNP2422	0.99	1	1					
SNP2429	0.93	0.94	0.94	1				
SNP2518	1	0.99	0.99	0.93	1			
SNP2548	0.99	1	1	0.94	0.99	1		
SNP2572	1	0.99	0.99	0.93	1	0.99	1	
SNP2575	0.99	1	1	0.94	0.99	1	0.99	1

Table 6.4: Matrix of r^2 measures for linkage disequilibrium between the 8 SNPs that are among the 10 most significant in Table 6.3 regardless of method used to combine the ridge regression results.

In Section 6.1 we saw that SNP2540 and SNP2603 were included in a logistic regression model when applying forward likelihood ratio inclusion at significance level 10^{-3}. SNP2540 was, when changing the significance level of the forward inclusion criterion to 10^{-7} or 5×10^{-8}, the only SNP included. In the ridge regression analysis SNP2540 was again among the most significant SNPs when combining the p-values with Stouffer's method with weights equal to the square root of the study sizes. In summary, SNP2540 is found to be the only SNP that has high significance in both the multivariable logistic regression analysis with forward inclusion and the ridge regression analysis. Furthermore, as seen in the regional association plots 6.1, 6.2, 6.4 and 6.5, SNPs from the same genomic region, that are in high linkage disequilibrium with each other, are all deemed as having an association with disease, in both the logistic regression models and ridge regression models. This gives an indiciation that this particular genomic region should be more thoroughly investigated in future analyses.

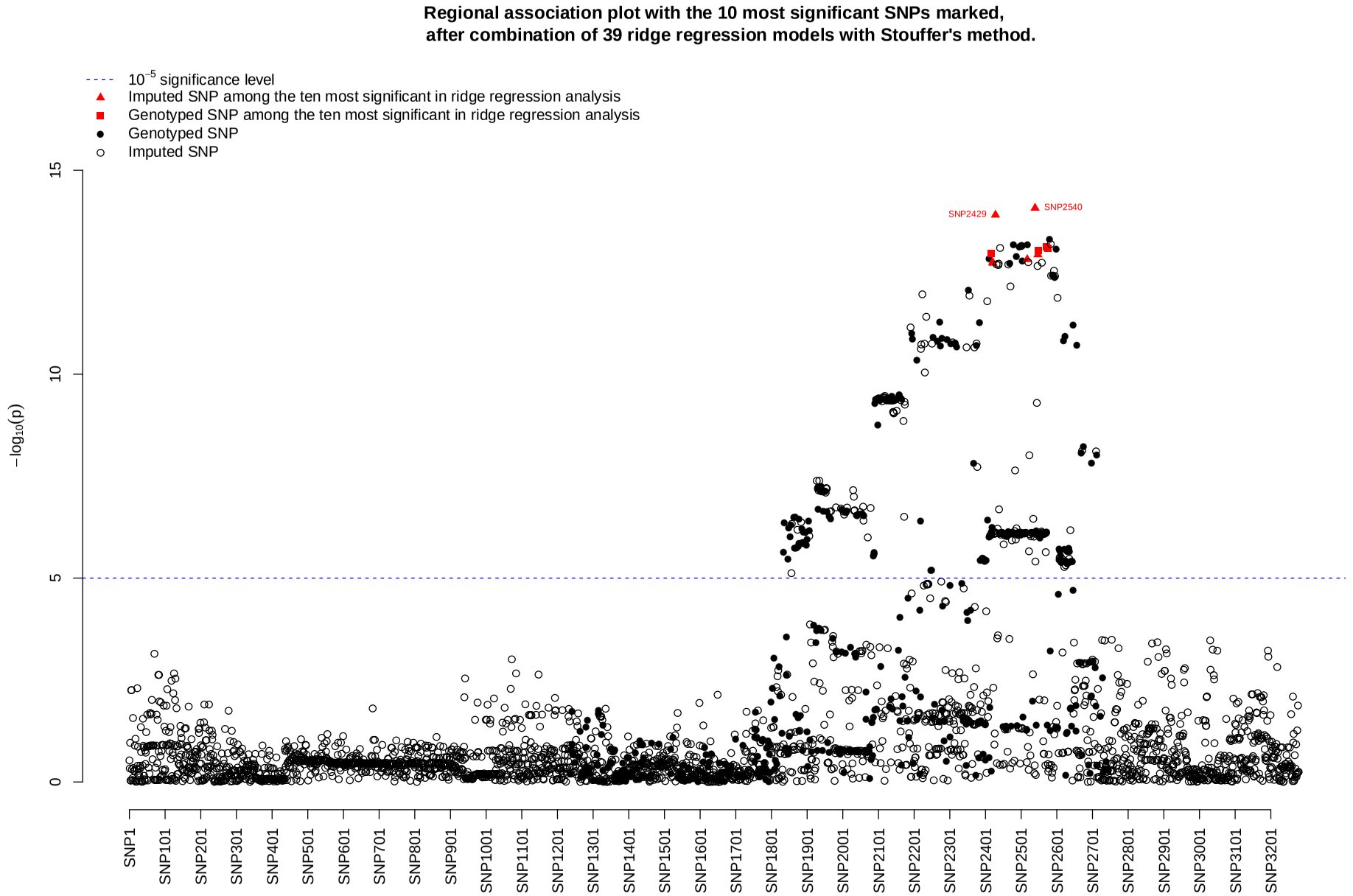

Figure 6.4: $-\log_{10}$ p-values from univariable logistic regression for all SNPs. The ten most significant SNPs in the ridge regression analysis after combination of p-values using Stouffer's method with weights $w_r = \sqrt{n_r}$ are indicated.

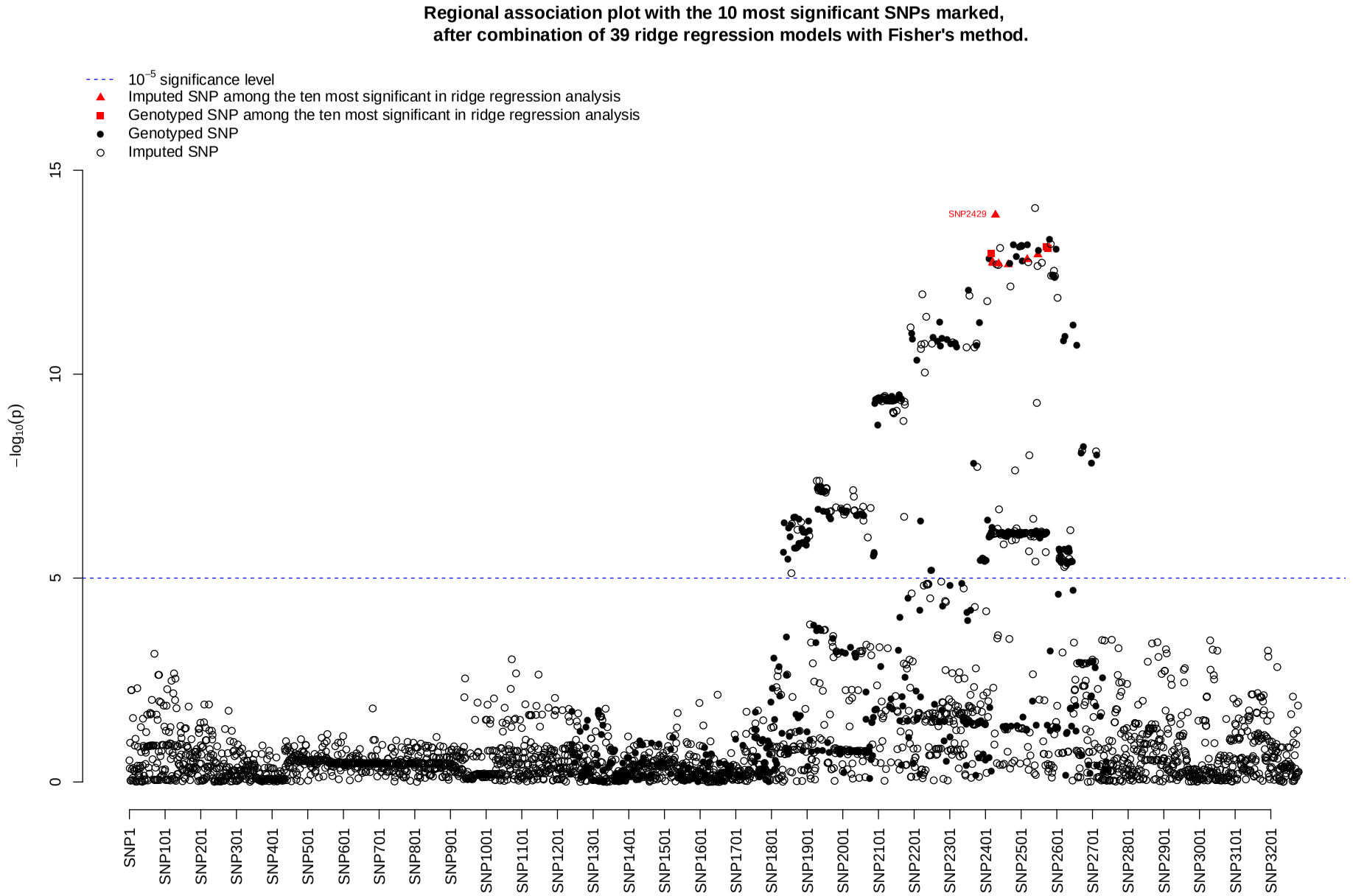

Figure 6.5: $-\log_{10}$ p-values from univariable logistic regression for all SNPs. The ten most significant SNPs in the ridge regression analysis after combination of p-values using Fisher's method are indicated.

Chapter 7

Summary & Discussion

In this thesis we have analyzed breast cancer data containg the disease status and genotype information for 89 050 european individuals, coming from 39 case-control studies. The aim of the analysis was to compare logistic regression models to ridge regression models in terms of localization of independent associated signals among multiple SNPs. In the initial univariable analyses done in Chapter 5, SNP2540 and SNP2429 are identified as the two SNPs most strongly associated with breast cancer risk. All in all, 326 SNPs are found to have a significant association with disease at significance level 10^{-5}. Using these 326 SNPs as covariates in a multivariable logistic regression model where variable selection was done using forward selection with a likelihood ratio test inclusion criterion, SNP2540 and SNP2603 where identified, see Chapter 6. Using the 326 SNPs again as covariates in 39 ridge regression models, one for each study, and then combining the results, identifies SNPs all coming from the same region as being significantly associated with disease, among them again being SNP2429 and SNP2540. Thus the analyses done in this thesis using univariable, multivariable and ridge logistic regression all point to the same SNP or region of SNPs as having an association with disease, namely SNP2540 and the area around it.

These analyses are the initial steps in a long chain of analyses with the purpose to localize regions that harbour SNPs with an association to breast cancer risk. Once the genomic regions, or SNPs, that have a significant association with disease are identified the next step is to focus solely on that region and explore the functional aspects of the SNPs in said region on a cellular level. Furthermore, one could include other epidemiological variables, associated with various aspects of the disease in question, as covariates in the regression analysis. One other important aspect is adjusting for population substructure, particularly when dealing with a big data originating from several studies as we have in this work. In genetic association studies one often includes principal components (PCs) as covariates to take into account the population substructure. Here it is important to include PCs based on the whole genome or at least a larger genomic area than we are studying in this thesis since one otherwise only adjusts for the population structure in the region we are currently investigating, see for example Michailidou et al. [2013].

In this thesis we used ridge regression as a penalized logistic regression approach. This was partly done since there was an interest in exploring the `ridge` package in `R` which implements the data-

driven method by Cule and De Iorio [2013] for estimation of the penalization parameter and then fits a ridge regression model using this penalization parameter. According to the package documentation, the function `logisticRidgeGenotypes` should have been able to analyze a very large amount of SNP data, since it uses an underlying C function as work horse. When conducting the ridge regression analysis on our data, we used genotype information for 89 050 individuals at 326 SNPs, thus yielding approximately 29 million data points. As explained in Section 6.2, this analysis did not yield any results since the algorithm for estimating the penalization parameter did not converge. The function `logisticRidge` in the package does the same kind of penalization parameter estimation and model fitting as `logisticRidgeGenotypes`, the difference being that it reads the data directly into `R`. Another difference with regard to the estimation of the penalization parameter, is that the maximum number of PCs t used to estimate the parameter is such that that the t first PCs together explain at least 90% of the variation. Whereas in `logisticRidgeGenotypes` the maximum number of PCs used is $t = \min(n, p)$, where n is the number of individuals and p is the number of SNPs in the analyzed data. It is unfortunate that the analysis of all individuals at once did not work, since the ability to analyze a large data set at once was what was attractive about the `ridge` package and a large data set is often the case in genetic association studies. An interesting next step here would be to investigate the code part of `logisticRidgeGenotypes` written in C to determine what exactly the problem is when fitting.

Another difference between `logisticRidge` and `logisticRidgeGenotypes` is that the latter only accepts genotype information coded as 0, 1, 2 for the number of a-alleles present, while the former also allows genotype information on the form of allele dosage as explained in Chapter 5. Thus, when the analysis was done with `logisticRidgeGenotypes` the imputed genotypes where given as the most probable genotype, based on the probabilities for each of the three possible genotypes that the imputation algorithm provides. Since this approach did not converge, the analysis was, as explained, done with `logisticRidge` and the imputed genotypes where given as allele dosage. It seems strange that two functions from the same package doing the same type of model fitting, should accept differently coded input. It seems like `logisticRidgeGenotypes` is quite limited by only allowing genotype information on the form 0, 1, 2.

As mentioned briefly in Section 4.3, there are other penalized regression methods possible as well, such as Lasso and Elastic net. One important difference between these methods and ridge regression is that they actually perform variable selection, meaning they can estimate a regression coefficient to be exactly zero. In ridge regression the estimated regression coefficients are only shrunken towards zero, but never actually set equal to zero. Both Lasso and Elastic net are implemented in `R`, for example in the `penalized` package where the penalization parameter(s) is chosen using likelihood-based cross-validation. Furthermore, the ridge penalization is also implemented in this package. It would have been interesting to compare the models we got using the method by Cule and De Iorio [2013] for estimation of the penalization parameter and the ones where cross-validation was used to estimate the penalization parameter, perhaps with regard to the size of the penalization parameter as well as for example the predictive ability of the models or which SNP region they deem as significant.

In this thesis the aim was to analyze a genomic region by utilizing both logistic regression and logistic ridge regression techniques to identify loci that are significantly associated with breast cancer risk. We found a promising subregion with an association to breast cancer risk, which was the same regardless of regression technique utilized. The analyses done in this thesis are important first steps in locating specific locations contributing to breast cancer risk and the subregion found should be subject to further investigations.

Appendix A

Supplementary Theory

In Chapter 4 we introduce the theory and methods used in this thesis. Since we analyze a binary response we focus in that chapter on different logistic regression settings. For completeness sake we, in this appendix, give some background and extend the theory from Chapter 4 to include also the linear regression setting. We start by introducing the theory behind principal component analysis (PCA) and how the principal components (PCs) are constructed in Section A.1. The PCs can then be used as covariates in principal component regression (PCR), which is introduced in the linear regression setting in Section A.2. Here we also explain the relationship between PCR and the singular value decomposition. Moving on to ridge regression, we outline in Section A.3 how the method by Cule and De Iorio [2013] for estimation of the penalization parameter is utilized in the linear ridge regression setting.

A.1 Principal Component Analysis

Principal component analysis (PCA) is a technique used for explaining a set of correlated variables by a reduced number of uncorrelated variables having maximal variance, where the uncorrelated variables are called principal components (PCs) [Aguilera et al., 2006]. For the derivation of the principal components we follow Jolliffe [2002, chapter 1].

Consider a vector of p random variables, $\mathbf{x} = (x_1, \ldots, x_p)'$. PCA seeks the uncorrelated linear combinations $z_1, \ldots, z_p$ of these, that have maximum variance. $z_1 = \boldsymbol{\alpha}_1'\mathbf{x}$ is the linear function with the vector of coefficients $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{1p})'$ that maximises $\text{Var}(z_1) = \text{Var}(\boldsymbol{\alpha}_1'\mathbf{x}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{x}$, under the constraint $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$. Without this constraint it is clear that $\text{Var}(z_1)$ can be increased by simply multiplying $\boldsymbol{\alpha}_1$ by some constant.

The standard approach for maximizing $\text{Var}(\boldsymbol{\alpha}_1'\mathbf{x}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$ subject to $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$ is the technique of Lagrange multipliers. We want to maximize the Lagrangian

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 - 1), \tag{A.1.1}$$

where λ is the Lagrange multiplier. Differentiating with respect to $\boldsymbol{\alpha}_1$ yields

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1 = 0 \tag{A.1.2}$$

or

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I}_p)\boldsymbol{\alpha}_1 = 0, \tag{A.1.3}$$

which we recognize as the characteristic polynomial and thus λ is an eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_1$ is the corresponding eigenvector. The quantity to be maximized is

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1'\lambda\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = \lambda, \tag{A.1.4}$$

so λ must be as large as possible. Therefore $\boldsymbol{\alpha}_1$ is the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}$, denoted by λ_1, and $\mathrm{Var}(z_1) = \mathrm{Var}(\boldsymbol{\alpha}_1'\mathbf{x}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda_1$.

Generally, $z_j = \boldsymbol{\alpha}_j\mathbf{x}$ is called the jth PC of $\mathbf{x}$ with $\mathrm{Var}(z_j) = \lambda_j$, the jth largest eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_j$ is the corresponding eigenvector, where $j = 1, \ldots, p$. Furthermore, $\mathrm{Cov}(z_j, z_k) = 0$ for $j \neq k,\ j, k = 1, \ldots, p$. A proof when $j = 2$ can be found in Jolliffe [2002, page 5-6].

The vectors $\boldsymbol{\alpha}_j$ are commonly referred to as the vectors of coefficients or 'loadings' for the jth PC.

A.2 Principal component linear regression

We follow Jolliffe [2002, chapter 8]. The linear regression model with response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $n \times p$ design matrix $\mathbf{X}$, whose (i, j)th element is the value of the jth covariate for the ith observation, is written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{A.2.1}$$

where $\boldsymbol{\beta}$ is the vector of p regression coefficients and $\boldsymbol{\epsilon}$ is a vector of independent error terms, each with the same variance. Here we assume that $\mathbf{Y}$ is centered about its mean, i.e. each value of $\mathbf{Y}$ has been subtracted by the mean of $\mathbf{Y}$. Furthermore we assume that each column $X_j,\ j = 1, \ldots, p$, of $\mathbf{X}$ has been centered and standardized, meaning each value of X_j has been subtracted by its mean and divided by its standard deviation. This way $\mathbf{X}'\mathbf{X}$ is proportional to the correlation matrix of $\mathbf{X}$. This means that, if $\mathbf{X}$ consists of genotype information at p SNPs, the off-diagonal elements of $\mathbf{X}'\mathbf{X}$ would describe the correlation between the SNPs. Squaring the elements of $\mathbf{X}'\mathbf{X}$ thus gives us a measure of the degree of linkage disquilibrium between the SNPs, as explained in Equation (3.1.3).

Now the values of the PCs for each observation are given by

$$\mathbf{Z} = \mathbf{X}\mathbf{A}, \tag{A.2.2}$$

where the (i, j)th element of $\mathbf{Z}$ is the value of the jth PC for the ith observation, $i = 1, \ldots, n, j = 1, \ldots, p$, and $\mathbf{A}$ is the $p \times p$ orthogonal matrix whose jth column is the jth eigenvector of the correlation matrix $\mathbf{X}'\mathbf{X}$. Since $\mathbf{A}$ is orthogonal we can write

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}\mathbf{A}\mathbf{A}'\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{A.2.3}$$

where $\boldsymbol{\gamma} = \mathbf{A}'\boldsymbol{\beta}$, meaning we have simply replaced the covariates by their PCs in the regression model. Now, $\widehat{\boldsymbol{\gamma}}$ can be calculated as the algebraic solutions to the normal equations for model (A.2.3) as

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}. \tag{A.2.4}$$

Using only a reduced set of PCs in the principal component regression (PCR) model, any large variances for coefficient estimates caused by multicollinearities can be reduced, even though these estimators usually are biased. The reduced PCR model takes the form

$$\mathbf{Y} = \mathbf{Z}_s\boldsymbol{\gamma}_s + \boldsymbol{\epsilon}_s, \tag{A.2.5}$$

where $s < p$ so $\boldsymbol{\gamma}_s$ is a vector of s elements that are a subset of elements of $\boldsymbol{\gamma}$, $\mathbf{Z}_s$ is an $n \times s$ matrix whose columns are the corresponding columns of $\mathbf{Z}$ and $\boldsymbol{\epsilon}_s$ is the appropriate error term.

A.2.1 Connection to singular value decomposition

As mentioned above, if we assume that each $X_j,\ j = 1, \ldots, p$, has been centered and standardized, $\mathbf{X}'\mathbf{X}$ is proportional to the correlation matrix of $\mathbf{X}$. Utilizing the singular value decomposition (B.2.1) and spectral theorem (B.1.1) respectively, we get that

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{A}' \tag{A.2.6}$$

$$\mathbf{X}'\mathbf{X} = \mathbf{A}\mathbf{D}^2\mathbf{A}', \tag{A.2.7}$$

where the p columns of $\mathbf{U}$ are eigenvectors of $\mathbf{X}\mathbf{X}'$, the p columns of $\mathbf{A}$ are the eigenvectors of $\mathbf{X}'\mathbf{X}$ and $\mathbf{D}^2$ is the diagonal matrix with eigenvalues of $\mathbf{X}'\mathbf{X}$ in descending order in the diagonal.

We saw previously that the PCs are given by $\mathbf{Z} = \mathbf{X}\mathbf{A}$, where the columns of $\mathbf{A}$ correspond to the eigenvectors of $\mathbf{X}'\mathbf{X}$. By this, (A.2.6) and (A.2.7), we see that the PCs are given by $\mathbf{U}\mathbf{D}$ and the coefficients of the linear combinations, a.k.a. 'loadings', are given by the columns of $\mathbf{A}$: $\mathbf{U}\mathbf{D}$=$\mathbf{X}\mathbf{A}$ [Jolliffe, 2002].

Furthermore, using the above we can rewrite the PCR coefficient estimates in equation (A.2.4) as

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} = ((\mathbf{U}\mathbf{D})'\mathbf{U}\mathbf{D})^{-1}\mathbf{Z}'\mathbf{Y} = \mathbf{D}^{-2}\mathbf{Z}'\mathbf{Y}. \tag{A.2.8}$$

A.3 Choosing the penalization parameter in linear ridge regression - a summary of the method by Cule and De Iorio [2013]

In Section 4.5 we described the method proposed by Cule and De Iorio [2013] for estimation of the penalization parameter in the logistic ridge regression setting. For the completeness of this thesis we here provide the method in the linear ridge regression setting.

Consider n individuals having been genotyped at p SNPs each. $\mathbf{X}$ is the covariate matrix with genotype information for each individual $i,\ i = 1, \ldots, n$, at each SNP $j,\ j = 1, \ldots, p$, meaning that $x_{ij} \in \{0, 1, 2\}$ depending on the number of a-alleles in individual i's genotype at SNP j. Each $X_j,\ j = 1, \ldots, p$, i.e. each column of $\mathbf{X}$, is centered and standardized, meaning each observation in X_j is subtracted by its column mean and divided by its standard deviation. The sample covariance matrix is given by $\mathbf{S}= \mathbf{X}'\mathbf{X}/(n-1)$, thus if we divide each observation in $\mathbf{X}$ by $\sqrt{n-1}$ we get that $\mathbf{S} = \mathbf{X}'\mathbf{X}$, where $\text{diag}(\mathbf{S})= \mathbf{1}_p$ and $\mathbf{1}_p$ is a p-length vector of ones. We denote this as $\mathbf{S}$ being in correlation form and the off-diagonal elements describe the correlations between the SNPs, i.e. the linkage disequilibrium as described in Section 3.1.1.

Furthermore, in the linear regression setting, $\mathbf{Y}= (Y_1, \ldots, Y_n)'$ is the vector of phenotype measurements that have been centered, i.e. the mean of $\mathbf{Y}$ is subtracted from each $Y_i,\ i = 1, \ldots, n$.

We also define the so called 'hat' or projection matrix $\mathbf{H}$ that relates the fitted outcomes of a regression model to the observed ones, i.e. $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$. The specific form of $\mathbf{H}$ is described below.

Note, previously the penalization parameter was denoted by λ, here we denote it by k since λ is reserved for the eigenvalues.

The first step of the proposed approach is to calculate the PCs of $\mathbf{X}$ as $\mathbf{Z}=\mathbf{X}\mathbf{A}$ and the PCR coefficients as $\widehat{\boldsymbol{\gamma}} = \boldsymbol{\Lambda}^{-1}\mathbf{Z}'\mathbf{Y}$, see Equation (A.2.2) and (A.2.8), where the columns of $\mathbf{A}$ are the eigenvectors of $\mathbf{S}$ and $\boldsymbol{\Lambda}$ is the diagonal matrix with eigenvalues of $\mathbf{S}$, denoted by $\lambda_j,\ j = 1, \ldots, t$, as diagonal elements, in descending order and at most $t = \min(n, p)$ of them are non-zero. For $r = 1, \ldots, t$ the penalization parameter k_r is calculated as

$$k_r = \frac{r\hat{\sigma}_r^2}{\widehat{\boldsymbol{\gamma}}_r'\widehat{\boldsymbol{\gamma}}_r}, \tag{A.3.1}$$

where

$$\hat{\sigma}_r^2 = \frac{(\mathbf{Y} - \mathbf{Z}_r\widehat{\boldsymbol{\gamma}}_r)'(\mathbf{Y} - \mathbf{Z}_r\widehat{\boldsymbol{\gamma}}_r)}{n-r} \tag{A.3.2}$$

and $\widehat{\boldsymbol{\gamma}}_r$ is the vector of the first r PCR coefficients and $\mathbf{Z}_r$ are the first r columns of $\mathbf{Z}$. The number of PCs to use in the calculation of k_r is then chosen as the r^* that minimizes the difference between r and the effective degrees of freedom for variance in the model calculated using k_r. The effective degrees of freedom for variance are $\text{tr}(\mathbf{H}'\mathbf{H})$, where $\mathbf{H}$ is the 'hat' matrix as described above. In the linear ridge regression setting we have that $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X} + k_r\mathbf{I}_p)^{-1}\mathbf{X}'$.

Thus the ridge estimates in the linear ridge regression setting are calculated, using k_{r^*}, as

$$\widehat{\boldsymbol{\beta}}_{RR,k_{r^*}} = (\mathbf{X}'\mathbf{X} + k_{r^*}\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}. \tag{A.3.3}$$

Appendix B

Matrix Decompositions

In this chapter we state the Spectral theorem, Section B.1, and the Singular value decomposition, B.2, which can both be utilized when doing principal component linear regression as explained in Appendix A.2.

B.1 Spectral theorem

The spectral decomposition of a $n \times n$ symmetric matrix $\mathbf{A}$ is given by [Johnson and Wichern, 1998]

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i u_i u_i' = \mathbf{U}\mathbf{\Lambda}\mathbf{U}', \tag{B.1.1}$$

where $\mathbf{\Lambda}$ is the diagonal matrix with eigenvalues of $\mathbf{A}$, λ_i, $i = 1, \ldots, n$, as diagonal elements in descending order and u_i are the associated normalized eigenvectors and form the columns of $\mathbf{U}$. Thus $\mathbf{U}$ is an orthogonal matrix, i.e. $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of dimension n.

B.2 Singular value decomposition

The singular value decomposition of a $n \times p$ matrix $\mathbf{A}$ is given by [Johnson and Wichern, 1998]

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}', \tag{B.2.1}$$

where $\mathbf{D}$ is the $p \times p$ diagonal matrix with so called singular values $\sigma_j = \sqrt{\lambda_j}$, $j = 1, \ldots, p$, as diagonal elements in descending order, where λ_j denote the eigenvalues of $\mathbf{X}\mathbf{X}'$ (and $\mathbf{X}'\mathbf{X}$). $\mathbf{U}$ is the orthogonal $n \times p$ matrix where the columns form the eigenvectors of $\mathbf{X}\mathbf{X}'$ and $\mathbf{V}$ is the orthogonal $p \times p$ matrix where the columns form the eigenvectors of $\mathbf{X}'\mathbf{X}$.

Bibliography

J. Aerssens, M. Armstrong, R. Gilissen, and N. Cohen. The human genome: An introduction. *The Oncologist*, 6:100–109, 2001.

A Agresti. *Categorical Data Analysis*. John Wiley & Sons, 2013.

A.M. Aguilera, M. Escabias, and M.J. Valderrama. Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50:1905–1924, 2006.

A. Albert and J.A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 1:1–10, 1984.

K.G. Ardlie, L. Kruglyak, and M. Seielstad. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet*, 3:299–309, 2002.

Breast Cancer Association Consortium. URL http://www.cogseu.org/.

Breast Cancer Association Consortium. Commonly studied single-nucleotide polymorphisms and breast cancer: Results from the breast cancer association consortium. *J Natl Cancer Inst*, 98(19):1382–96, 2006.

Cancerfonden. URL www.cancerfonden.se.

COGS. Collaborative Oncological Gene-environment Study. URL http://www.cogseu.org/,http://www.nature.com/icogs/.

E. Cule and M. De Iorio. Ridge regression in prediction problems: Automatic choice of the ridge parameter. *Genetic Epidemiology*, 37:704–714, 2013.

E. Cule, P. Vineis, and M. De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12:372, 2011. URL http://www.biomedcentral.com/1471-2105/12/372.

Erika Cule. Package 'ridge', 2014. URL http://cran.r-project.org/web/packages/ridge/ridge.pdf.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: Models, Methods and Applications*. Springer, 2013.

A.S. Foulkes. *Applied Statistical Genetics with R: For Population-based Association Studies*. Springer, 2009.

J.D. French, M. Ghoussaini, S.L. Edwards, K.B. Meyer, K. Michailidou, S. Ahmed, S. Khan, and M.J. Maranian et al. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin d1 expression through long-range enhancers. *The American Journal of Human genetics*, 92:489–503, 2013.

Genetic Science Learning Center, University of Utah. URL http://learn.genetics.utah.edu.

J.J. Goeman. L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, 52:70–84, 2010.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition edition, 2010.

A. E. Hoerl, R.W. Kennard, and K.F. Baldwin. Ridge regression: Some simulations. *Communications in Statistics*, 4(2):105–123, 1975.

A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67,69–82, 1970a,b.

B.N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, 2009. doi: 10.1371/journal.pgen.1000529.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

Anne-Sophie Jannot, Georg Ehret, and Thomas Perneger. $P < 10^{-8}$ has emerged as a standard of statistical significance for genome-wide association studies. *Journal of Clinical Epidemiology*, 68(4):460 – 465, 2015. ISSN 0895-4356. doi: http://dx.doi.org/10.1016/j.jclinepi.2015.01.001.

R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis, Fourth Edition*. Prentice-Hall, Inc., 1998.

J.T. Jolliffe. *Principal Component Analysis, Second Edition*. Springer, 2002.

John M. Lachlin. *Biostatistical Methods: The Assessment of Relative Risks, Second Edition*. John Wiley & Sons, Inc, 2011.

C.D. Langefeld and T.E. Fingerlin. Association methods in human genetics. *Methods in Molecular Biology*, 404:431–460, 2007.

C.M. Lewis and J. Knight. Introduction to genetic association studies. *Cold Spring Harbor Protoc.*, 2012. doi: 10.1101/pdb.top068163.

W. Li. Three lectures on case-control genetic association analysis. *Briefings in Bioinformatics*, 9 (1):1–13, 2008. doi: 10.1093/bib/bbm058.

P. Lichenstein, N.V. Holm, P.K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, and E. Pukkala et al. Environmental and heritable factors in the causation of cancer − Analyses of Cohorts from Twins from Sweden, Denmark, and Finland. *The New England Journal of Medicine*, 343:78–95, 2000.

J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet.*, 11(7):499–511, 2010. doi: 10.1038/nrg2796.

K. Michailidou, Hall P., A. Gonzalez-Neira, M. Ghoussaini, J. Dennis, R.L. Milne, and M.K. Schmidt et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet.*, 45(4):353–361, 2013. doi: 10.1038/ng.2563.

K. Michailidou, J. Beesley, S. Lindstrom, S. Canisius, J. Dennis, M.J. Lush, and M.J. Maranian et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet.*, 47:373–380, 2015. doi: 10.1038/ng.3242.

I. Pe'er, R. Yelensky, D. Altshuler, and M.J. Daly. Estimation of multiple testing burden for genomwide association studies of nearly all common variants. *Genetic Epidemiology*, 32:381–385, 2008.

P.D. Sasieni. From genotype to genes: Doubling the sample size. *Biometrics*, 53:1253–1261, 1997.

R. Schaefer, L. Roi, and R. Wolfe. A ridge logistic estimator. *Communications in Statistics - Theory and Methods*, 13(1):99–113, 1984.

J.H. Shin, S. Blay, B. McNeney, and J. Graham. Ldhmodels: An r function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of Statistical Software*, 16:Code Snippet 3, 2006.

A.V. Spencer, A. Cox, and K. Walters. Comparing the efficacy of snp filtering methods for identifying a single causal snp in a known association region. *Annals of Human Genetics*, 78: 50–61, 2014.

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

D.C. Thomas. *Statistical Methods in Genetic Epidemiology.* Oxford University Press, Inc., 2004.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(1): 267–288, 1996.

C. Turnbull, S. Ahmed, J. Morrison, D. Pernet, A. Renwick, M. Maranian, and S. Seal et al. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.*, 43(6):504–507, 2010.

J.C. Van Hoewelingen and S. Le Cessie. Predictive value of statistical models. *Statistics in Medicine*, 9:1303–1325, 1990.

M.C. Whitlock. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, 18:1368–1373, 2005.

D.V. Zaykin. Optimally weighted Z-test is a powerful method for combining probabilities in meta.analysis. *Journal of Evolutionary Biology*, 24:1836–1841, 2011.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67(Part 2):301–320, 2005.