NTNU – Trondheim
Norwegian University of
Science and Technology

Statistical methods for detecting genotype-phenotype association in the presence of environmental covariates

Marit Runde

Master of Science in Physics and Mathematics
Submission date: June 2013
Supervisor: Mette Langaas, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Preface

This thesis completes my master's degree in Industrial Mathematics at the Norwegian University of Science and Technology (NTNU). The work has been carried out at the Department of Mathematical Sciences during the spring of 2013.

I would like to give special thanks to my supervisor, Associate Professor Mette Langaas, for great guidance and feedback during the work of this thesis. I have appreciated our weekly meetings and your helpful inputs and advices. I would also like to thank Associate Professor Øyvind Bakke for valuable comments and help when needed.

Trondheim, June 2013

Abstract

This thesis shows how statistical methods based on logistic regression models can be used to analyze and interpret biological data. In genome-wide association studies, the aim is to detect association between genetic markers and a given phenotype. This thesis considers a situation where the phenotype is the absence or presence of a common disease, the genetic marker is a biallelic single nucleotide polymorphism (SNP), and environmental covariates are available. The main goal is to study and compare four statistical methods (Score test, Likelihood ratio test, Wald test and Cochran-Armitage test for trend) which, by using different approaches, test the hypothesis about whether there is an association or not between the disease and the genetic marker. The methods are applied to simulated datasets in order to measure their test size and statistical power, and to compare them. Interaction between the genetic marker and the environmental effect is also considered, and strategies for simulating cohort and case-control data with genotype and environmental covariates are studied.

The power simulations show that methods based on logistic regression models are appropriate for detecting genotype-phenotype association, but when the environmental effect is moderate, a simpler method (Cochran-Armitage test for trend) which does not require model fitting at all, is adequate. When an interaction effect is included in the model, the hypothesis testing becomes more complex. Several possible approaches to this problem are discussed.

Sammendrag

Denne masteroppgaven viser hvordan statistiske metoder basert på logistiske regresjonsmodeller kan brukes til å analysere og tolke biologiske data. I genetiske assosiasjonsstudier er hensikten å finne assosiasjon mellom genetiske markører og en gitt fenotype. Denne oppgaven ser på en situasjon hvor fenotypen er hvorvidt en vanlig sykdom er tilstede eller ikke, den genetiske markøren er en biallelisk enkel nukleotide polymorfisme, og miljøkovariater er tilgjengelig. Hovedmålet er å studere og sammenligne fire statistiske metoder (Score test, Likelihood ratio test, Wald test og Cochran-Armitage test for trend) som, ved å bruke ulike tilnærminger, tester hypotesen om hvorvidt det finnes en assosiasjon eller ikke mellom sykdommen og den genetiske markøren. Metodene er brukt på simulerte datasett for å beregne deres teststørrelse og -styrke, og for å kunne sammenligne dem. Interaksjon mellom den genetiske markøren og miljøeffekten er også studert, samt stategier for simulering av kohort og case-control data med genotype- og miljøkovariater.

Styrkesimuleringene viser at metoder basert på logistiske regresjonsmodeller er hensiktsmessige for å oppdage genotype-fenotype-assosiasjon, men når miljøeffekten er moderat, vil en enklere metode (Cochran-Armitage test for trend), som ikke krever modelltilpassning i det hele tatt, være tilstrekkelig. Når en interasjonseffekt er inkludert i modellen, så vil hypotesetesingen bli mer sammensatt. Flere mulige tilnærminger til dette problemet er diskutert.

Contents

1 Introduction **1**
1.1 Objective . . . 1
1.2 Genotype and phenotype . . . 2
1.3 Single nucleotide polymorphisms . . . 2
1.4 Hardy-Weinberg equlibrium . . . 3
1.5 Genome-wide association studies . . . 3
1.6 Candidate gene study . . . 4
1.7 The common disease common variant hypothesis . . . 4
1.8 Epidemiology and study design . . . 4
1.9 Interaction . . . 5
1.10 Statistical software . . . 5
1.11 Structure of the report . . . 5

2 Statistical methods **7**
2.1 The exponential family of distributions . . . 7
2.2 The logistic regression model . . . 9
2.3 Maximum likelihood estimation in logistic regression . . . 10
2.4 Odds ratio . . . 13
2.5 Normal data and logistic regression . . . 14
2.6 The Score test . . . 15
2.7 The Likelihood ratio test . . . 16
2.8 The Wald test . . . 18
2.9 Graphically representation of the Score test, the Likelihood ratio test and the Wald test . . . 18
2.10 The Cochran-Armitage test for trend . . . 19

3 Statistical models including genotype and environmental covariates **21**
3.1 The simplest model . . . 21
3.2 Model including a linear genotype covariate . . . 22
3.3 Model including a linear genotype covariate and a linear environmental covariate . . . 24
3.4 Model including an interaction term . . . 26

3.5 The genotype variable and the environmental variable as factors . . 29

4 Statistical test theory **33**
4.1 Hypothesis testing . . . 33
4.2 Statistical power and test size . . . 34
4.3 Estimating power by simulation . . . 34

5 Power study **37**
5.1 Procedure for data simulation . . . 37
5.2 Data simulation and hypothesis testing . . . 39
5.3 Comparison of the performance of the Score test, the LRT and the Wald test . . . 40
5.4 Performance of the CATT and the Score test . . . 41
5.5 Effect of parameter values . . . 47
5.5.1 Intercept parameter . . . 47
5.5.2 Minor allele frequency . . . 47
5.5.3 Distribution of the environmental variable . . . 48
5.5.4 Size of the effects . . . 49
5.5.5 Level of significance . . . 50
5.5.6 Sample size . . . 50
5.6 Simulation inspired by the TOP study . . . 51
5.7 Data simulation and hypothesis testing when an interaction effect is present . . . 53

6 Simulation of case-control data **61**
6.1 Method to simulate case-control data when the genotype is the only covariate . . . 61
6.2 Method to simulate case-control data including a genotype covariate and an environmental covariate . . . 62
6.3 Comments . . . 64

7 Discussion and conclusion **67**
7.1 Different approaches . . . 67
7.2 Conclusion . . . 68

Bibliography **71**

A R code **73**
A.1 Data simulation . . . 73
A.2 Hypothesis testing without assuming an interaction effect . . . 74
A.3 Hypothesis testing when assuming an interaction effect . . . 75

Chapter 1

Introduction

1.1 Objective

In traditional medicine, patients have been treated based on how the majority of previous patients with similar symptoms and diagnosis have responded to various treatment or treating methods. This works for those who respond well to the treatment the majority of the population need to recover, but for those who rather need the treatment that only works for a minority, the majority treatment might give adverse consequences. Medical scientists are constantly in search for new and better methods, procedures and treatments which can improve the way diseases and conditions are treated. As they have been able to identify and map the human genome and how genomic variation is linked to pharmacological properties, a new branch within the field of medical research has been developed. This branch is called personalized medicine.

Personalized medicine is about being able to predict susceptibility to disease, improve detection of disease, give more effective and customized medical treatment, and predict and prevent side effects of drugs based on each individual unique genetic makeup. To be able to do this, we need techniques to detect which elements that have impact on the risk of developing different diseases. Such elements may be environmental factors like smoking and obesity, or it may be clinical factors like variations in the DNA, or what is most likely, a combination.

We are in this thesis going to consider some statistical methods for this purpose, and the majority of these are constructed around the logistic regression model. By using simulated datasets for a given genetic marker, we will test how well these methods detect association between a disease and the genotype when we also have data for the environmental influence available. Some may find it appropriate to adjust for the environmental factors, but we choose to merge them into one variable and include this in the model. The environmental factors may of course

differ depending on the disease we are considering. If we can find reliable statistical methods that detect the association between genotype and disease, this will lead to a better understanding of different diseases. It will also be possible to give future patients better customized medical treatment.

We start out by defining and explaining some main concepts within the fields of biology and epidemiology that will be useful for this study.

1.2 Genotype and phenotype

The genome is the heredity information of an individual. It includes both coding and non-coding sequences of DNA. Each individual inherits two copies of each DNA sequence, one from its mother and one from its father. The DNA sequences are organized in what is called chromosomes. Each individual inherits 23 chromosomes from each parent, giving 23 pairs of chromosomes. The chromosomes in a pair are homogeneous, which means they carry genes for the same characteristics and have identical structure, thus genes that belong together is placed in the same position (locus) in the pair of chromosomes. In each position on the chromosome we find what is called alleles. The combination of alleles at the corresponding locus in a pair of chromosomes is together referred to as the genotype.

If we consider a situation where there are two possible alleles at a given position in a population under study, 'a' and 'A', we have three possible genotypes; 'aa', 'aA' and 'AA'. We will in this thesis denote 'aa' as genotype 0, 'aA' as genotype 1 and 'AA' as genotype 2. Here, 'A' is assumed to be the allele with the lowest frequency in the population we are considering. In our study we are looking at statistical methods to test if the genotype has any influence on the probability of developing a given disease.

In addition to the genotype influence, we are also interested in whether environmental factors, like smoking habits, diet and age, have an impact on this probability or not. The disease under study is in a more general context called a phenotype. A phenotype is an observed characteristic of an individual. This characteristic is a result of the genetic makeup and influence from the environment.

1.3 Single nucleotide polymorphisms

Genetic variations may occur within our DNA. Such variants may affect which diseases we develop, how well we respond to medical treatment, which side effects we experience and so on (Human Genome Project Information, 2013). The DNA sequence is made up of the four bases guanine, cytosine, adenine and thymine. A possible situation may be that one base in the sequence is alternated with another.

As an example, we can imagine that a guanine base is replaced by an adenine base in a particular position in the DNA sequence, and thus this sequence is slightly changed. If we consider the population under study, we can calculate the ratio of chromosomes carrying the less common variant to chromosomes with the more common variant. This ratio is called the frequency of the minor allele, and if the frequency is greater than 0.01 in the population under study, the variation is called a single nucleotide polymorphism (SNP) (Human Genome Project Information, 2013). We will throughout this thesis refer to the minor allele frequency as MAF.

1.4 Hardy-Weinberg equlibrium

When assuming that the genotype and the allele frequencies in a large, randomly mating population will remain stable over generations, and that the relationship between genotype and allele frequencies is fixed, we assume what is called the Hardy-Weinberg equilibrium (HWE) (Ziegler & König, 2010, p. 38).

If we define the genotype frequencies in a population as P(genotype 0) $= g_0$, P(genotype 1) $= g_1$ and P(genotype 2) $= g_2$, with $g_0 + g_1 + g_2 = 1$, then the allele frequencies P(a) $= g_0 + \frac{1}{2}g_1 = p$ and P(A) $= g_2 + \frac{1}{2}g_1 = q$, with $p + q = 1$. Next, we consider matings within this population. When assuming that the parental genotypes are independent, we obtain the following frequencies for the different genotypes occurring in the offspring; P(aa) $= p^2$, P(aA) $= 2pq$ and P(AA) $= q^2$. Hence, the genotype frequencies for the offspring can be calculated directly from the allele frequencies in the original population the parents were a part of.

1.5 Genome-wide association studies

The aim of a genome-wide association (GWA) study is to detect common genetic variants that might affect the probability of developing certain diseases or disorders (National Human Genome Research Institute, 2013). GWA studies are usually performed by studying the DNA of individuals with a disease of interest, and then compare to the DNA of individuals not suffering from the disease. By doing this, it is possible to detect if there are any variants associated with the certain disease (Zheng et al., 2012, Chapter 12). The individuals that do have the disease of interest, are called the cases, and those who are not affected by it, are called the controls. The advantage of GWA studies is that the entire genome is studied, not just a few genetic regions.

We will in this thesis consider some statistical methods that are efficient for detecting how a genetic marker and how environmental factors affect the risk of developing a certain disease in GWA studies. The methods we discuss may also be used for each SNP in a GWA study.

1.6 Candidate gene study

In GWA studies, the entire genome is scanned for common genetic variations. In some settings we do have some prior knowledge and an idea of which genes that may play a role in the development of the disease of interest. In such situations, a candidate gene study is a more appropriate approach to test for disease association (Daly & Day, 2001). Here, only the predetermined genes of interest are screened in order to detect association to disease. Typically, the selected genes are analyzed among a group of individuals carrying a certain disease and a group of individuals not carrying this disease. If genetic variations do appear significantly more frequently among the diseased individuals than in the rest of the population, a genetic marker is identified. An important precondition of the candidate gene study is the presence of information from previous studies about where to look for genetic variation. If such information is not available, a GWA study is more appropriate.

1.7 The common disease common variant hypothesis

GWA studies are said to be powerful approaches to identify variants associated with phenotypes under the hypothesis called the common disease common variant (CDCV) hypothesis. CDCV assumes that the genetic risk of a common complex disease (or phenotype) is mainly attributed to a small number of high-frequency genetic variants with moderately small effects. A competing hypothesis, the common disease rare variant (CDRV) hypothesis, suggests that multiple rare variants are the major contributors to genetic susceptibility to such common complex diseases. A brief history of the debate centered around these two hypotheses are given in Schork et al. (2009).

In our work, we are going to fit a model based on logistic regression to each SNP. Hence, we will assume the CDCV hypothesis. If each model was fitted to several SNPs together, the CDRV hypothesis would have been appropriate.

1.8 Epidemiology and study design

Epidemiology is a wide field of research, and it seems to be more definitions of epidemiology than there are epidemiologists. Rothman et al. (2008) (p. 32) use the following definition; 'Epidemiology is the study of the distribution of health-related states and events in populations.' Thus, epidemiology is all about systematically looking for patterns, frequencies, causes and effects of health-related conditions in a population. Rothman et al. (2008) (p. 88) divide the possible study designs for

observational studies into four main groups. These are; 'cohort study − in which all subjects in a source population are classified according to their exposure status and are followed over time to ascertain disease incidence; case-control studies − in which cases arising from a source population and a sample of the source population are classified according to their exposure history; cross-sectional studies − in which one ascertains exposure and disease status as of a particular time; and at last, ecologic studies − in which the units of observations are groups of people.'

Since most GWA studies are based on case-control designs, we will in our work use simulated data in a case-control setting.

1.9 Interaction

The term interaction has in the present context a number of different interpretations. According to Wang et al. (2010), there are no universally accepted definition in neither biology nor statistics. Generally, one can say that the term implies that objects or factors in a study do not act independently. It is common to distinguish between statistical interaction and biological interaction. Statistical interaction is used by statisticians to describe departure from additivity in statistical models. Wang et al. (2010) define biological interaction as the joint action of two or more factors, thinking of the physical interaction between molecules.

We will in this thesis use the term interaction in a statistical setting to describe departure from additivity. That is, if an explanatory variable in a model will influence the response variable differently depending on the value of another explanatory variable, then we state that an interaction effect is present.

1.10 Statistical software

All statistical analyses in this thesis were performed using the statistical software R, R Development Core Team (2013). The packages *statmod* by Gordon Smyth (2013), and *Rassoc* by Yong Zang, Wingkam Fung and Gang Zheng (2009) were used.

1.11 Structure of the report

We have now given a brief introduction to some basic terms and concepts in the fields of biology and epidemiology. We will continue in Chapter 2 with reviewing some statistical methods we later will use to detect association between genotype and disease. Among these methods, there are three which are based on likelihood

estimation, and one which does not require any kind of model estimation. In Chapter 3 we present some possible statistical models and corresponding hypotheses and test statistics. Further, in Chapter 4 we will address hypothesis testing, and how to estimate statistical power by simulation. Chapter 5 starts with a power study where we evaluate and compare the performance of the four methods introduced in Chapter 2. We will also perform a case study inspired by a research project called the Thematically Organized Psychosis (TOP) study. At the end of Chapter 5 we include an interaction effect in the simulated datasets, and look at the performance of the methods in this setting. In Chapter 6 we outline a method to simulate case-control data, and finally, in Chapter 7, we sum up and discuss our findings before drawing conclusions.

Chapter 2

Statistical methods

The content of this chapter is mainly based on Chapter 2 in McCullagh & Nelder (1989), Chapter 4, 5 and 7 in Dobson & Barnett (2008), Chapter 1 and 5 in Agresti (1996), Smyth (2003) and Langaas & Bakke (2013).

2.1 The exponential family of distributions

Probability distributions that can be written on the form

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right), \tag{2.1}$$

where $a(\phi)$, $b(\theta)$ and $c(y,\phi)$ are known functions, are said to belong to the exponential family of distributions (McCullagh & Nelder, 1989, Chapter 2). The parameter of interest, θ, is called the canonical parameter, while ϕ is regarded as a nuisance parameter. The distributions belonging to this family, share some properties that make them very useful for statistical analysis.

It can be shown that the expected value and the variance of a random variable Y, which belongs to an exponential family, are given by

$$\mathrm{E}(Y) = b'(\theta) \quad \text{and} \quad \mathrm{Var}(Y) = a(\phi)b''(\theta),$$

respectively. Here, b' denotes the first derivative of b with respect to θ, and b'' denotes the second derivative of b with respect to θ. The log-likelihood function is given by

$$l(\theta,\phi;y) = \ln(f(y;\theta,\phi)) = \frac{y\theta - b'(\theta)}{a(\phi)} + c(y,\phi).$$

To find an expression for what is called the score vector, denoted U, we differentiate the log-likelihood function with respect to θ. This gives

$$U = \frac{\partial l(\theta, \phi; Y)}{\partial \theta} = \frac{Y - b'(\theta)}{a(\phi)} \quad \Rightarrow \quad U = \frac{Y - \mathrm{E}(Y)}{a(\phi)}.$$

Hence, the expected value of the score vector U is given by

$$\mathrm{E}(U) = \mathrm{E}\left(\frac{Y - \mathrm{E}(Y)}{a(\phi)}\right) = \frac{\mathrm{E}(Y)}{a(\phi)} - \frac{\mathrm{E}(Y)}{a(\phi)} = 0, \tag{2.2}$$

and the variance of the score vector U can be written as

$$\mathrm{Var}(U) = -\mathrm{E}\left(\frac{\partial U}{\partial \theta}\right) = \frac{b''(\theta)}{a(\phi)}.$$

Several of the most common probability distributions are members of the exponential family of distributions. This includes the normal, exponential, binomial, gamma, chi-squared and Poisson distribution. Our focus will be on the binomial distribution, which has the probability mass function given by

$$f(y; n, p) = \binom{n}{y} p^y (1-p)^{n-y}, \tag{2.3}$$

where $y = 0, 1, 2, \ldots, n$ is the number of successes in n independent trials, and p is the probability of success in each trial. If we rewrite (2.3) to the form given in (2.1), we get $a(\phi) = 1$, $b(\theta) = n \ln(\exp(\theta) + 1)$ and $c(y, \phi) = -\ln \binom{n}{y}$. Hence, we obtain

$$U = Y - np,$$

and

$$\mathrm{Var}(U) = np(1-p)$$

for the binomial distribution.

2.2 The logistic regression model

We define a random variable

$$Z = \begin{cases} 1 & \text{if the outcome is a success} \\ 0 & \text{if the outcome is a failure} \end{cases},$$

with $\mathrm{P}(Z = 1) = \pi$ and $\mathrm{P}(Z = 0) = 1 - \pi$. Hence, $Z \sim \mathrm{bin}(1, \pi)$. The logistic regression model is by Dobson & Barnett (2008) (p. 126) defined as

$$\mathrm{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{x}^T\boldsymbol{\beta}, \tag{2.4}$$

and is used when the outcome Z is a binary variable. In the model, $\boldsymbol{\beta}$ is a vector of parameters, and $\boldsymbol{x}$ is a vector of explanatory variables. These explanatory variables can be either categorical or continuous. We can find the expression for the probability of success by rewriting (2.4) to

$$\pi = \frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^T\boldsymbol{\beta})}. \tag{2.5}$$

We now define Y to be the number of successes in n independent trials, each with equal probability of success, π. This random variable is given by

$$Y = \sum_{i=1}^{n} Z_i,$$

and follows the binomial distribution with the probability mass function given by

$$f(y; n, \pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y},$$

where $y = 0, 1, 2, \ldots, n$.

Next, we consider N independent random variables, $Y_1, Y_2, ..., Y_N$, corresponding to the number of successes in N different subgroups. A subgroup is here defined as a group of observations with identical levels or values of the explanatory variables. The combination of the values of the explanatory variables is called the covariate pattern. If one or more of the explanatory variables are continuous, then there are often few observations for each covariate pattern, and the number of patterns may be equal to the number of observations. If all variables are categorical, there is a clearly limited number of possible combinations. This usually gives several observations for each covariate pattern if the total number of observations are

relatively large. Each subgroup j has its own probability of success, π_j. Due to the fact that the observations $Z_1, Z_2, ..., Z_N$ are independent, $Y_1, Y_2, ..., Y_N$ are also independent, and we obtain the joint probability density function

$$f(y_1, ..., y_N; n_1, ..., n_N, \pi_1, ..., \pi_N) = \prod_{j=1}^{N} \binom{n_j}{y_j} \pi_j^{y_j} (1 - \pi_j)^{n_j - y_j}.$$

The likelihood function $L(\boldsymbol{\theta}; \boldsymbol{y})$ is algebraically the same as the joint probability density function $f(\boldsymbol{y}; \boldsymbol{\theta})$, but there is a change in the notation. In the probability density function, we have a set of observations, $\boldsymbol{y}$, given the fixed values of the parameters $\boldsymbol{\theta}$, while in the likelihood function we have a set of parameter values, $\boldsymbol{\theta}$, given some observations $\boldsymbol{y}$. By taking the logarithm of the likelihood function for Y_1, Y_2, ..., Y_N, we obtain the log-likelihood function

$$l(\pi_1, ..., \pi_N; y_1, ..., y_N) = \sum_{j=1}^{N} \left[y_j \ln\left(\frac{\pi_j}{1 - \pi_j}\right) + n_j \ln(1 - \pi_j) + \ln\binom{n_j}{y_j} \right]. \quad (2.6)$$

By inserting the logit function, (2.4), and the probability of success, (2.5), into this log-likelihood function, we get

$$l(\pi_1, ..., \pi_N; y_1, ..., y_N) = \sum_{j=1}^{N} \left[y_j \boldsymbol{x}_j^T \boldsymbol{\beta} - n_j \ln(1 + \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})) + \ln\binom{n_j}{y_j} \right]. \quad (2.7)$$

2.3 Maximum likelihood estimation in logistic regression

To estimate the parameters in $\boldsymbol{\beta}$, we use the method of maximum likelihood. To maximize the likelihood, we need to obtain the derivative of the log-likelihood function (2.7) with respect to β_0, β_1, ..., β_{p-1} and β_p. This vector, the score vector $\boldsymbol{U}$, is given by

$$\boldsymbol{U} = \begin{bmatrix} U_0 \\ U_1 \\ \vdots \\ U_p \end{bmatrix} = \begin{bmatrix} \frac{\partial l}{\partial \beta_0} \\ \frac{\partial l}{\partial \beta_1} \\ \vdots \\ \frac{\partial l}{\partial \beta_p} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{N} \left(y_j - n_j \frac{\exp(\boldsymbol{x}_j^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})} \right) \\ \sum_{j=1}^{N} \left(x_{j,1} \left(y_j - n_j \frac{\exp(\boldsymbol{x}_j^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})} \right) \right) \\ \vdots \\ \sum_{j=1}^{N} \left(x_{j,p} \left(y_j - n_j \frac{\exp(\boldsymbol{x}_j^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})} \right) \right) \end{bmatrix}, \quad (2.8)$$

where $x_{j,k}$ is the kth element in $\boldsymbol{x}_j$. The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is the solution of $\boldsymbol{U}(\boldsymbol{\beta}) = \mathbf{0}$. This solution can not in general be given in a closed form, so we will use the Fisher scoring algorithm, which is a form of the Newton-Raphson method, to find a numerical solution. In order to use this method, we need to calculate the variance-covariance matrix of the score vector (2.8). By using (2.5), we can rewrite (2.8) to

$$\boldsymbol{U} = \begin{bmatrix} \sum_{j=1}^{N} (y_j - n_j\pi_j) \\ \sum_{j=1}^{N} x_{j,1}(y_j - n_j\pi_j) \\ \vdots \\ \sum_{j=1}^{N} x_{j,p}(y_j - n_j\pi_j) \end{bmatrix}. \tag{2.9}$$

The variance-covariance matrix of $\boldsymbol{U}$ is called the information matrix, denoted $\mathfrak{J}$, and its elements are given by $\mathfrak{J}_{jk} = \mathrm{E}(U_j U_k)$, where $j,\ k = 0,\ 1,\ \ldots,\ p$. This is derived by using

$$\begin{aligned} \mathrm{Var}(U_i) &= \mathrm{E}(U_i^2) - \mathrm{E}(U_i)^2, \\ \mathrm{Cov}(U_i, U_j) &= \mathrm{E}(U_i U_j) - \mathrm{E}(U_i)\mathrm{E}(U_j) \end{aligned}$$

and the fact that $\mathrm{E}(U_i) = 0$ for all i, as shown in (2.2). Inserting the score vector $\boldsymbol{U}$ from (2.9) into the expression for $\mathfrak{J}_{jk}$, we get

$$\mathfrak{J}_{jk} = \mathrm{E}(U_j U_k) = \mathrm{E}\left(\sum_{i=1}^{N} x_{i,j}(y_i - n_i\pi_i) \sum_{l=1}^{N} x_{l,k}(y_l - n_l\pi_l) \right),$$

where $x_{i,0} = x_{l,0} = 1$. We know that $Y_i \sim \mathrm{bin}(n_i, \pi_i)$ and that $Y_1,\ Y_2,\ \ldots,\ Y_N$ are independent. Hence,

$$\begin{aligned} \mathrm{E}(Y_i) &= n_i\pi_i \quad \text{and} \\ \mathrm{Var}(Y_i) &= n_i\pi_i(1 - \pi_i). \end{aligned} \tag{2.10}$$

Due to

$$\mathrm{E}((y_i - n_i\pi_i)(y_l - n_l\pi_l)) = \mathrm{E}((y_i - \mathrm{E}(Y_i))(y_l - \mathrm{E}(Y_l))) = \begin{cases} \mathrm{Var}(Y_i) & \text{when } i = l \\ 0 & \text{when } i \neq l \end{cases},$$

the information matrix can be written as

$$\mathfrak{J} = \begin{bmatrix} \sum_{i=1}^{N} \mathrm{Var}(Y_i) & \sum_{i=1}^{N} x_{i,1}\mathrm{Var}(Y_i) & \dots & \sum_{i=1}^{N} x_{i,p}\mathrm{Var}(Y_i) \\ \sum_{i=1}^{N} x_{i,1}\mathrm{Var}(Y_i) & \sum_{i=1}^{N} x_{i,1}^2\mathrm{Var}(Y_i) & \dots & \sum_{i=1}^{N} x_{i,1}x_{i,p}\mathrm{Var}(Y_i) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_{i,p}\mathrm{Var}(Y_i) & \sum_{i=1}^{N} x_{i,p}x_{i,1}\mathrm{Var}(Y_i) & \dots & \sum_{i=1}^{N} x_{i,p}^2\mathrm{Var}(Y_i) \end{bmatrix}, \quad (2.11)$$

where

$$\mathrm{Var}(Y_i) = n_i \frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\beta})}$$

from (2.5) and (2.10).

As mentioned earlier in this chapter, we will use the Fisher scoring algorithm to find the values of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, that solves $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}$. This algorithm is presented in Dobson & Barnett (2008) (p. 65). By slightly rewriting this, we get

$$\boldsymbol{b}^{(m+1)} = \boldsymbol{b}^{(m)} + [\mathfrak{J}^{(m)}]^{-1}\boldsymbol{U}^{(m)},$$

where $\boldsymbol{b}^{(m+1)}$ is the vector of the maximum likelihood estimates of the parameters in $\boldsymbol{\beta}$ at the $(m+1)$th iteration, $\mathfrak{J}$ is the information matrix given in (2.11) and $\boldsymbol{U}$ is the score vector from (2.9). Both the information matrix and the score vector are evaluated at $\boldsymbol{b}^{(m)}$. Given an initial guess, $\boldsymbol{b}^{(0)}$, of the values, the Fisher scoring algorithm obtains an improved estimate, $\boldsymbol{b}^{(1)}$. The algorithm uses this new estimate to find an even better estimate, $\boldsymbol{b}^{(2)}$. This procedure is repeated several times until the difference between $\boldsymbol{b}^{(m)}$ and $\boldsymbol{b}^{(m+1)}$ is sufficiently small. Then, $\boldsymbol{b}^{(m+1)}$ is used as the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$.

We have now given an introduction to the logistic regression model and how to estimate the parameters in this model using maximum likelihood estimation. In the next section, we will introduce the concept of odds ratio, and show how this is related to the logistic regression model. Then, in Section 2.5, we will see how different mean values for a normal distributed explanatory variable will lead to a logistic regression model. In the Sections 2.6-2.10, four statistical methods used to detect genotype-phenotype association are presented. These methods will later be evaluated and compared by using simulated datasets.

2.4 Odds ratio

Odds ratio (OR) is a measure of effect size which is widely utilized in the field of epidemiology. If we divide a population into subgroups based on some criterion of interest, then Agresti (1996) (p. 22) defines the odds of success within each subgroup i to be

$$\text{odds}_i = \frac{\pi_i}{1-\pi_i}. \tag{2.12}$$

The odds ratio is known as the ratio of odds from two subgroups. That is,

$$\text{OR} = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)},$$

where $i = 1$ corresponds to subgroup 1, and $i = 2$ corresponds to subgroup 2.

In a logistic regression setting, odds ratio may be used to interpret the estimated coefficients in the model. From (2.4) we can write the logistic regression model to the form

$$\frac{\pi}{1-\pi} = \exp(\boldsymbol{x}^T\boldsymbol{\beta}), \tag{2.13}$$

by using the exponential function on both sides. If π is the probability of success, then the left side of (2.13) is in (2.12) defined as the odds of success. We can now calculate the increase or decrease in the response when the level of one of the explanatory variables is increased or decreased by one unit. The model will have the form

$$\begin{aligned}\frac{\pi_i}{1-\pi_i} &= \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}) \\ &= \exp(\beta_0)\exp(\beta_1 x_{1i})\exp(\beta_2 x_{2i})\ldots\exp(\beta_p x_{pi}).\end{aligned}$$

A change in x_{ji} from x_{ji} to $x_{ji}+1$, gives the odds ratio

$$\text{OR} = \frac{\exp(\beta_0)\exp(\beta_1 x_{1i})\exp(\beta_2 x_{2i})\ldots\exp(\beta_j(x_{ji}+1))\ldots\exp(\beta_p x_{pi})}{\exp(\beta_0)\exp(\beta_1 x_{1i})\exp(\beta_2 x_{2i})\ldots\exp(\beta_j x_{ji})\ldots\exp(\beta_p x_{pi})} = \exp(\beta_j).$$

All the terms where the explanatory variable stays unchanged will vanish, and the only term left is the coefficient belonging to the variable that changes. Hence, the estimated coefficients returned from logistic regression are log-odds ratios. They can be interpreted as how the log-odds of success will change when the value of an explanatory variable changes with one unit. The sign of the log-odds ratio indicates whether the change is positive or negative.

2.5 Normal data and logistic regression

We may experience a situation where the distribution of the explanatory variables differ depending on the value of the response variable. We will here show how using logistic regression is a natural choice for such situations. First, consider the logistic regression model given by

$$\text{logit}(\pi(x)) = \beta_0 + \beta_1 x, \tag{2.14}$$

where x is the observation of the explanatory variable X. Suppose that the distribution of X is normal with mean μ_0 and variance σ^2 for all subjects where the response variable, Y, is equal to 0. For subjects where Y is equal to 1, the variance is the same, σ^2, but the mean is μ_1. That is, $X \sim N(\mu_0, \sigma^2)$ when $Y = 0$ and $X \sim N(\mu_1, \sigma^2)$ when $Y = 1$. If we define $\text{P}(Y = 1) = \lambda$ and $\text{P}(Y = 0) = 1 - \lambda$, then the probability distribution of X, f_X, is given by

$$f_X(x) = f_X(x \mid y = 0)(1 - \lambda) + f_X(x \mid y = 1)\lambda.$$

Further, by using Bayes' theorem, we can express the probability of the response variable to be equal to 1, given the value of the explanatory variable, as

$$\pi(x) = \text{P}(Y = 1 \mid X = x) = \frac{f_X(x; \mu_1, \sigma^2)\lambda}{f_X(x; \mu_0, \sigma^2)(1 - \lambda) + f_X(x; \mu_1, \sigma^2)\lambda}, \tag{2.15}$$

where $f_X(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\frac{1}{2\sigma^2}(x - \mu)^2)$. By inserting (2.15) into the logistic regression model given in (2.14), we get

$$\text{logit}(\pi(x)) = \log\left(\frac{\lambda}{1 - \lambda}\right) + \frac{1}{2\sigma^2}(\mu_0^2 - \mu_1^2) + \frac{\mu_1 - \mu_0}{\sigma^2}x,$$

which gives

$$\beta_0 = \log\left(\frac{\lambda}{1 - \lambda}\right) + \frac{1}{2\sigma^2}(\mu_0^2 - \mu_1^2)$$

and

$$\beta_1 = \frac{\mu_1 - \mu_0}{\sigma^2}.$$

Hence, the sign of β_1 is given by the sign of $\mu_1 - \mu_0$. This shows that when the population under study consists of one group of subjects where $Y = 0$, with a corresponding bell-shaped distribution of X, and one group where $Y = 1$, where X has a shifted bell-shaped distribution with similar variance, then the logistic regression

function is a good approximation for $\pi(x)$. If the distributions of the explanatory variable have highly different variance, Agresti (1996) (p. 108) suggests that the model should also include a quadratic term in order to be a good approximation for $\pi(x)$.

2.6 The Score test

We want to test whether a selection of the estimated parameters in the fitted model has a significant effect on the response variable or not. One of the methods which is frequently used for this purpose, is named the Score test.

To perform the Score test, we consider $l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y})$ which is a log-likelihood function like the one given in (2.7), but here the parameters are divided into two groups, such that $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]^T$. We want to test the hypothesis

$$H_0 : \boldsymbol{\theta}_2 = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\theta}_2 \neq \mathbf{0}.$$

In other words, we want to test if the parameters in $\boldsymbol{\theta}_2$ have a significant effect on the response. The parameters in $\boldsymbol{\theta}_1$ are not of interest in this test, but we need to find estimates for them as well in order to be able to compute a test statistic. These parameters which are not of interest, are called nuisance parameters. The information matrix, $\mathfrak{J}$, is according to Smyth (2003) given by the covariance matrix of the score vector. This can be partitioned as

$$\mathfrak{J} = \begin{bmatrix} \mathfrak{J}_{11} & \mathfrak{J}_{12} \\ \mathfrak{J}_{21} & \mathfrak{J}_{22} \end{bmatrix} \tag{2.16}$$

when the parameters are divided into the vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Here, $\mathfrak{J}_{11} = \mathrm{Var}\big(\frac{\partial l}{\partial \boldsymbol{\theta_1}}\big)$, $\mathfrak{J}_{12} = \mathfrak{J}_{21} = \mathrm{Cov}\big(\frac{\partial l}{\partial \boldsymbol{\theta_1}}, \frac{\partial l}{\partial \boldsymbol{\theta_2}}\big)$ and $\mathfrak{J}_2 = \mathrm{Var}\big(\frac{\partial l}{\partial \boldsymbol{\theta_2}}\big)$. We denote $\frac{\partial l}{\partial \boldsymbol{\theta}_1} = \boldsymbol{U}_1$ and $\frac{\partial l}{\partial \boldsymbol{\theta}_2} = \boldsymbol{U}_2$.

There are now two possible situations. The parameters in $\boldsymbol{\theta}_1$ may be known or unknown. If they are known, we follow Smyth (2003) and use the score test statistic given by

$$S = \boldsymbol{U}_2^T (\mathfrak{J}_{22})^{-1} \boldsymbol{U}_2,$$

where both $\boldsymbol{U}_2$ and $\mathfrak{J}_{22}$ are evaluated at $\boldsymbol{\theta}_2 = \mathbf{0}$.

If the parameters in $\boldsymbol{\theta}_1$ are unknown, we use the maximum likelikelihood estimate $\hat{\boldsymbol{\theta}}_1$, which is equivalent to solve $\boldsymbol{U}_1 = \frac{\partial l}{\partial \boldsymbol{\theta}_1} = 0$. The next step is to find the conditional distribution of $\boldsymbol{U}_2 \mid (\boldsymbol{U}_1 = 0)$. We know that

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_1 \\ \boldsymbol{U}_2 \end{bmatrix} \sim N(\mathbf{0}, \mathfrak{I}),$$

where the zero mean follows from (2.2), and $\mathfrak{I}$ is given in (2.16). Johnson & Wichern (2007) show that the conditional distribution of $(\boldsymbol{U}_2 \mid (\boldsymbol{U}_1 = \boldsymbol{u}_1)$ then is multivariate normal with mean equal to $\mathrm{E}(\boldsymbol{U}_2) + \mathfrak{I}_{21}\mathfrak{I}_{11}^{-1}(\boldsymbol{u}_1 - \mathrm{E}(\boldsymbol{U}_1))$, and covariance given by $\mathfrak{I}_{22} - \mathfrak{I}_{21}\mathfrak{I}_{11}^{-1}\mathfrak{I}_{12}$.

Since $\boldsymbol{u}_1 = \mathbf{0}$ and, as we know from (2.2), both $\mathrm{E}(\boldsymbol{U}_1)$ and $\mathrm{E}(\boldsymbol{U}_2)$ are equal to zero, the mean for the conditional distribution is also equal to zero. The covariance matrix for the conditional distribution is given by

$$\mathfrak{J}^* = \mathrm{Cov}(\boldsymbol{U}_2 \mid \boldsymbol{U}_1 = 0) = \mathfrak{I}_{22} - \mathfrak{I}_{21}\mathfrak{I}_{11}^{-1}\mathfrak{I}_{12}. \tag{2.17}$$

The score test statistic is by Smyth (2003) defined as

$$S = \boldsymbol{U}_2^T(\mathfrak{J}^*)^{-1}\boldsymbol{U}_2,$$

where both $\boldsymbol{U}_2$ and $\mathfrak{J}^*$ are evaluated at $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}_2 = \mathbf{0}$.

Both when the nuisance parameters are known and when they are unknown, the test statistic is χ^2-distributed with number of degrees of freedom equal to the dimension of the $\boldsymbol{\theta}_2$ vector. Hence, the null hypothesis will be rejected if $S > \chi^2_{\alpha,\dim(\boldsymbol{\theta}_2)}$ at a α-level of significance.

An advantage of the Score test is that the parameters in the model only have to be estimated once. If we do have many datasets, this will save computation time.

2.7 The Likelihood ratio test

The Likelihood ratio test (LRT) is a test based on comparing the values of the likelihood functions corresponding to two nested models. We call these models 'the model of interest' and 'the reference model'. By nested we mean that one of the models (the model of interest) contains a selection of the parameters from the other model (the reference model). If $L(\boldsymbol{b}; \boldsymbol{y})$ is the maximum likelihood function of the model of interest, and $L(\boldsymbol{b}_{\text{reference}}; \boldsymbol{y})$ is the maximum likelihood function of the reference model, then

$$\frac{L(\boldsymbol{b}_{\text{reference}}; \boldsymbol{y})}{L(\boldsymbol{b}; \boldsymbol{y})}$$

is the ratio of the likelihoods. How well the model of interest fits the data compared to the reference model is assessed by using the logarithm of this likelihood ratio, which gives

$$l(\boldsymbol{b}_{\text{reference}}; \boldsymbol{y}) - l(\boldsymbol{b}; \boldsymbol{y}).$$

Large values of this difference indicate that the model of interest has a poor fit compared to the reference model. The likelihood ratio test statistic, also known as the deviance statistic, is according to Dobson & Barnett (2008) (p. 80), defined as

$$D = 2[l(\boldsymbol{b}_{\text{reference}}; \boldsymbol{y}) - l(\boldsymbol{b}; \boldsymbol{y})]. \tag{2.18}$$

This statistic is used to test the hypothesis that all the parameters which are included in the reference model, but not in the model of interest, are equal to zero. The maximum likelihood estimates for the reference model are $\hat{\pi}_j = \frac{y_j}{n_j}$, and by inserting this into (2.6), we get the log-likelihood of the reference model,

$$l(\boldsymbol{b}_{\text{reference}}; \boldsymbol{y}) = \sum_{j=1}^{N} \left[y_j \ln \left(\frac{\frac{y_j}{n_j}}{1 - \frac{y_j}{n_j}} \right) + n_j \ln \left(1 - \frac{y_j}{n_j} \right) + \ln \binom{n_j}{y_j} \right]. \tag{2.19}$$

To obtain the log-likelihood of any model nested to the reference model, we need the estimated fitted values from this model. These estimates can be denoted $\hat{y}_i = n_i \hat{\pi}_i$, where the $\hat{\pi}_i$s are the estimates for the probabilities. The log-likelihood is then given by

$$l(\boldsymbol{b}; \boldsymbol{y}) = \sum_{j=1}^{N} \left[y_j \ln \left(\frac{\frac{\hat{y}_j}{n_j}}{1 - \frac{\hat{y}_j}{n_j}} \right) + n_j \ln \left(1 - \frac{\hat{y}_j}{n_j} \right) + \ln \binom{n_j}{y_j} \right]. \tag{2.20}$$

Hence, by inserting (2.19) and (2.20) into (2.18), we get

$$D = 2 \sum_{j=1}^{N} \left[y_j \ln \left(\frac{y_j}{\hat{y}_j} \right) + (n_j - y_j) \ln \left(\frac{n_j - y_j}{n_j - \hat{y}_j} \right) \right].$$

Dobson & Barnett (2008) state that if the hypothesis is true, then two times the difference between the log-likelihood for the reference model and for the model of interest is χ^2-distributed with $N - p$ degrees of freedom. Here, N is the number of parameters in the reference model, and p is the number of parameters in the model of interest. If we only have a few observations per covariate pattern, $D \sim \chi^2_{\alpha, N-p}$ is a poor approximation.

2.8 The Wald test

The Wald test is used to draw conclusions about the true value, $\boldsymbol{\beta}$, of a parameter based on its estimated value from the sample. Let $\boldsymbol{b}$, with p elements, be an maximum likelihood estimator of $\boldsymbol{\beta}$. The Wald test statistic is then in Dobson & Barnett (2008) (p. 77) defined as

$$W = (\boldsymbol{b} - \boldsymbol{\beta})^T \mathfrak{J}(\boldsymbol{b})(\boldsymbol{b} - \boldsymbol{\beta}),$$

where $\mathfrak{J}$ is the variance-covariance matrix from (2.11). Under the null hypothesis, the Wald statistic is χ^2-distributed with p degrees of freedom and significance level α, $W \sim \chi^2_{\alpha,p}$.

2.9 Graphically representation of the Score test, the Likelihood ratio test and the Wald test

We have now considered three likelihood based methods. These methods can all be used to test if leaving out one or several covariates will reduce how well the model fits the data, but they use different approaches. The null hypothesis for the three tests is that the smaller model is the true model. We will here let the coefficient of interest from the model be denoted θ. The score test statistic is calculated based on the slope of the log-likelihood function at the value of θ given in the null hypothesis, θ_0. This slope is used to estimate the change in the model fit if additional variables were included or removed from the model. The Likelihood ratio test compares the log-likelihood of a model where the value of θ is given in the null hypothesis with the log-likelihood of a model where θ is estimated. The estimated value of θ is denoted $\hat{\theta}$. The comparison is done by checking if the difference in the values of these two log-likelihood functions is statistical significant. The Wald test does not compare the difference between the log-likelihood functions of two models, but rather the difference between the estimated parameter corresponding to the fitted model and the parameter under the null hypothesis. If the estimated value is significant different from the value given in the null hypothesis, then the null hypothesis will be rejected.

Figure 2.1 shows graphically what the Score test, the Likelihood ratio test and the Wald test examine in order to draw conclusions about the null hypothesis H_0: $\theta = \theta_0$ when there are no nuisance parameters included. The Score test evaluates how quickly the log-likelihood is changing at $\theta = \theta_0$, the Likelihood ratio test compares $\ln(L(\hat{\theta}))$ with $\ln(L(\theta_0))$, and the Wald test compares $\hat{\theta}$ with θ_0.

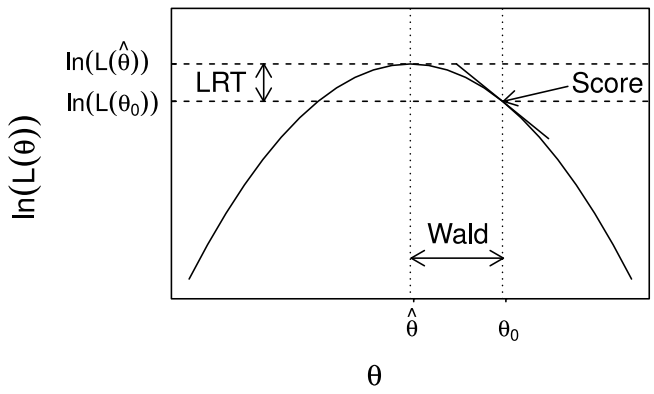

Figure 2.1: Relationship among the Score test statistic, the Likelihood ratio test statistic and the Wald test statistic. The figure is inspired by Figure D.16 in Fox (1997) (p. 570).

2.10 The Cochran-Armitage test for trend

The methods we have looked at so far have all been based on likelihood estimation. There are also other tests that may be used for the same purpose. We will here consider a non-model based method named the Cochran-Armitage test for trend (CATT). CATT is widely used in the statistical analysis in studies when the aim is to check if there is an association between a response variable with two levels, and an explanatory variable with k ordinal levels. The null hypothesis says that there is no trend. First we have to introduce what is called a contingency table, which is a table displaying the counts distribution of the variables we consider. Such a table, and its notation, is shown in Table 2.1.

	Genotype			
	aa	aA	AA	Sum
Case	x_0	x_1	x_2	n_1
Control	y_0	y_1	y_2	n_2
Sum	m_0	m_1	m_2	n

Table 2.1: Notation for the 2×3 contingency table.

Using the notation in Table 2.1, Langaas & Bakke (2013) obtain the CATT test statistic,

$$\text{CATT} = \frac{\sum_{i=0}^{2} s_i \left(n_2 x_i - n_1 y_i\right)}{\sqrt{n_1 n_2 \left(\sum_{i=0}^{2} s_i^2 m_i - \frac{1}{N}\left(\sum_{i=0}^{2} s_i m_i\right)^2\right)}},$$

where s_0, s_1 and s_2 are weights taking values depending on the choice of genetic model. By genetic model we mean the relationship between how the disease probabilities are modeled for the three genotypes. The most common genetic models are the three models called recessive, dominant and additive. In the recessive model, we assume that two copies of the high risk allele is necessary for developing the disease, while in the dominant model, only one copy of the high risk allele is necessary. For the additive genetic model, we assume that the allele combination 'aA' gives an increased risk of developing the disease compared to carrying the combination 'aa', but an decreased risk of developing the disease compared to carrying the combination 'AA'. According to Zheng et al. (2006), the trend test is optimal when $s_0 = 0$, $s_2 = 2$, and s_1 is equal to 0, 1 or 2 for the recessive, additive and dominant genetic model, respectively.

The CATT test statistic is asymptotically standard normal distributed under the null hypothesis, and the absolute value of it is invariant to linear transformations of the weights.

Zheng et al. (2012) (pp. 67-68) show that the squared CATT statistic is equal to the score test statistic for logistic regression when the genotype is the only covariate, and it is coded similarly as in CATT.

Chapter 3

Statistical models including genotype and environmental covariates

We will in this chapter consider several possible models constructed around the logistic regression model. The aim is to see what the score test statistic presented in Chapter 2 looks like for these models when considering appropriate statistical hypotheses. We will also derive expressions for parameter estimates for some models.

3.1 The simplest model

First, we look at the model where the intercept term is the only term. This model is not of interest by itself, but it provides a good structure for the upcoming models. We write this model as

$$\text{logit}(\pi_i) = \beta_0, \tag{3.1}$$

and the hypothesis we want to test is

$$H_0 : \beta_0 = 0 \quad \text{vs.} \quad H_1 : \beta_0 \neq 0.$$

The log-likelihood function for observation i is given by

$$l_i(\pi_i; y_i) = y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i) + \log \binom{n_i}{y_i},$$

and the score function then takes the form

$$U_i = \frac{\partial l_i}{\partial \pi_i} = \frac{Y_i}{\pi_i} + \frac{n_i - Y_i}{1 - \pi_i} = \frac{Y_i - n_i \pi_i}{\pi_i (1 - \pi_i)}.$$

By setting the score function equal to zero, we get $\hat{\beta}_0 = \ln\left(\frac{\bar{Y}}{1-\bar{Y}}\right)$, where $\bar{Y} = \frac{Y}{n}$ and n is the number of observations. Further, the information matrix is given by

$$\mathfrak{J} = \mathrm{Var}(U) = \frac{\mathrm{Var}(Y)}{\pi^2 (1 - \pi)^2} = \frac{n}{\pi(1 - \pi)}.$$

Under H_0, $\pi = \frac{1}{2}$. This can be shown by rewriting (3.1), using the link function $\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, to $\pi_i = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ and insert $\beta_0 = 0$.

The score test statistic for the model in (3.1) will then be

$$S = U^T \mathfrak{J}^{-1} U \mid_{H_0} = 4n\left(\bar{Y} - \frac{1}{2}\right)^2,$$

and the null hypothesis, H_0, is rejected if $S > \chi^2_{\alpha,1}$.

We now let Y be a random variable that follows the binomial distribution such that $Y \sim \mathrm{bin}(n, p)$, where n is the number of trials and p is the probability of success in each trial. Moreover, we let $\bar{Y} = \frac{Y}{n}$ which approximately follows the normal distribution, $\bar{Y} \sim N\left(p, \frac{p(1-p)}{n}\right)$. Hence, $\frac{\bar{Y}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathrm{N}(0,1)$. It is known that the sum of k squared normal distributed random variables follows the chi-squared distribution with k degrees of freedom. Thus, $\frac{(\bar{Y}-p)^2}{\frac{p(1-p)}{n}} \sim \chi^2_{\alpha,1}$. If we let $p = \frac{1}{2}$, we get $4n\left(\bar{Y} - \frac{1}{2}\right)^2$, which is identical to the score vector.

3.2 Model including a linear genotype covariate

The model we are going to consider in this section can be written as

$$\mathrm{logit}(\pi_i) = \beta_0 + \beta_G x_{Gi}, \tag{3.2}$$

where x_{Gi} can take the values 0, 1 and 2. The hypothesis of interest is now

$$H_0 : \beta_G = 0 \quad \text{vs.} \quad H_1 : \beta_G \neq 0,$$

and the log-likelihood function is given by

$$l_i(\pi_i; y_i) = y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i) + \log \binom{n_i}{y_i},$$

where

$$\pi_i = \frac{\exp(\beta_0 + \beta_G x_{Gi})}{1 + \exp(\beta_0 + \beta_G x_{Gi})}.$$

By using the log-likelihood function, we get the score vector $\boldsymbol{U} = [U_{\beta_0}, U_{\beta_G}]^T$, where

$$U_{\beta_0} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_0} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \pi_i} \frac{\partial \pi_i}{\partial \beta_0} = \sum_{i=1}^{n} (Y_i - n_i \pi_i)$$

and

$$U_{\beta_G} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_G} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \pi_i} \frac{\partial \pi_i}{\partial \beta_G} = \sum_{i=1}^{n} x_i (Y_i - n_i \pi_i).$$

From setting $U_{\beta_0} = 0$, we get $\hat{\pi} = \bar{Y}$, which gives $\hat{\beta}_0 = \ln\left(\frac{\bar{Y}}{1-\bar{Y}}\right)$. The estimate $\hat{\beta}_G$ can not be expressed in a closed form.

The information matrix for the model in (3.2) is given by

$$\mathfrak{J} = \begin{bmatrix} \sum\limits_{i=1}^{n} n_i \pi_i (1 - \pi_i) & \sum\limits_{i=1}^{n} x_i n_i \pi_i (1 - \pi_i) \\ \sum\limits_{i=1}^{n} x_i n_i \pi_i (1 - \pi_i) & \sum\limits_{i=1}^{n} x_i^2 n_i \pi_i (1 - \pi_i) \end{bmatrix}.$$

By using (2.17), we get

$$\mathfrak{J}^* = \sum_{i=1}^{n} x_i^2 n_i \pi_i (1 - \pi_i) - \frac{\left[\sum\limits_{i=1}^{n} x_i n_i \pi_i (1 - \pi_i)\right]^2}{\sum\limits_{i=1}^{n} n_i \pi_i (1 - \pi_i)}.$$

Inserting $\hat{\beta}_0$ and $\beta_G = 0$ gives

$$\mathfrak{J}^* \mid_{\hat{\beta}_0, \beta_G = 0} = \bar{Y}(1-\bar{Y}) \left[\sum_{i=1}^{n} n_i x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i n_i \right)^2}{\sum_{i=1}^{n} n_i} \right],$$

and the score test statistic is then expressed as

$$S = U^T (\mathfrak{J}^*)^{-1} U = \frac{\left[\sum_{i=1}^{n} \left(Y_i - n_i \bar{Y} \right) x_i \right]^2}{\bar{Y}(1-\bar{Y}) \left[\sum_{i=1}^{n} n_i x_i^2 - \left(\sum_{i=1}^{n} x_i n_i \right)^2 \left(\sum_{i=1}^{n} n_i \right)^{-1} \right]}.$$

The null hypothesis, H_0, is rejected if $S > \chi^2_{\alpha,1}$.

3.3 Model including a linear genotype covariate and a linear environmental covariate

In the previous section we considered a model with the genotype term as the only term in addition to the intercept term. In this section we will add an environmental covariate as well. We write the model as

$$\text{logit}(\pi_i) = \beta_0 + \beta_E x_{Ei} + \beta_G x_{Gi},$$

where x_{Ei} is the value of the environmental covariate for observation i, and x_{Gi} is equal to 0, 1 or 2 for the genotype 0, 1 and 2, respectively. The hypothesis we want to test is

$$H_0 : \beta_G = 0 \quad \text{vs.} \quad H_1 : \beta_G \neq 0.$$

The log-likelihood function for observation i takes the form

$$l_i(\pi_i; y_i) = y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i) + \log \binom{n_i}{y_i},$$

where

$$\pi_i = \frac{\exp(\beta_0 + \beta_E x_{Ei} + \beta_G x_{Gi})}{1 + \exp(\beta_0 + \beta_E x_{Ei} + \beta_G x_{Gi})}.$$

The elements in the score vector are given by

$$\boldsymbol{U} = \begin{bmatrix} U_{\beta_0} \\ U_{\beta_E} \\ U_{\beta_G} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} (Y_i - n_i\pi_i) \\ \sum_{i=1}^{n} x_{Ei}(Y_i - n_i\pi_i) \\ \sum_{i=1}^{n} x_{Gi}(Y_i - n_i\pi_i) \end{bmatrix}, \tag{3.3}$$

and the information matrix becomes

$$\mathfrak{J} = \begin{bmatrix} \sum_{i=1}^{n} n_i\pi_i(1-\pi_i) & \sum_{i=1}^{n} x_{Ei}n_i\pi_i(1-\pi_i) & \sum_{i=1}^{n} x_{Gi}n_i\pi_i(1-\pi_i) \\ \sum_{i=1}^{n} x_{Ei}n_i\pi_i(1-\pi_i) & \sum_{i=1}^{n} x_{Ei}^2n_i\pi_i(1-\pi_i) & \sum_{i=1}^{n} x_{Ei}x_{Gi}n_i\pi_i(1-\pi_i) \\ \sum_{i=1}^{n} x_{Gi}n_i\pi_i(1-\pi_i) & \sum_{i=1}^{n} x_{Ei}x_{Gi}n_i\pi_i(1-\pi_i) & \sum_{i=1}^{n} x_{Gi}^2n_i\pi_i(1-\pi_i) \end{bmatrix}.$$

Again, we use (2.17) and get

$$\mathfrak{J}^* = \mathfrak{J}_{22} - \mathfrak{J}_{21}\mathfrak{J}_{11}^{-1}\mathfrak{J}_{12}, \tag{3.4}$$

where

$$\mathfrak{J}_{22} = \sum_{i=1}^{n} x_{Gi}^2 n_i\pi_i(1-\pi_i),$$

$$\mathfrak{J}_{21} = \begin{bmatrix} \sum_{i=1}^{n} x_{Gi}n_i\pi_i(1-\pi_i) & \sum_{i=1}^{n} x_{Ei}x_{Gi}n_i\pi_i(1-\pi_i) \end{bmatrix},$$

$$\mathfrak{J}_{11} = \begin{bmatrix} \sum_{i=1}^{n} n_i\pi_i(1-\pi_i) & \sum_{i=1}^{n} x_{Ei}n_i\pi_i(1-\pi_i) \\ \sum_{i=1}^{n} x_{Ei}n_i\pi_i(1-\pi_i) & \sum_{i=1}^{n} x_{Ei}^2n_i\pi_i(1-\pi_i) \end{bmatrix}$$

and

$$\mathfrak{J}_{12} = \begin{bmatrix} \sum_{i=1}^{n} x_{Gi}n_i\pi_i(1-\pi_i) \\ \sum_{i=1}^{n} x_{Ei}x_{Gi}n_i\pi_i(1-\pi_i) \end{bmatrix}.$$

The score test statistic is then

$$S = U_{\beta_G}^T (\mathfrak{I}^*)^{-1} U_{\beta_G},$$

where U_{β_G} is given in (3.3), and $\mathfrak{I}^*$ is given in (3.4). The null hypothesis, H_0, is rejected if $S > \chi^2_{\alpha,1}$.

3.4 Model including an interaction term

In the present section we are going to consider a model similar to the one in the previous section, but we will also add an interaction term. We write the model as

$$\text{logit}(\pi_i) = \beta_0 + \beta_E x_{Ei} + \beta_G x_{Gi} + \beta_{EG} x_{Ei} x_{Gi},$$

where x_{Ei} is the value of the environmental covariate for observation i, and x_{Gi} is equal to 0, 1 or 2 for the genotype 0, 1 and 2, respectively. One possible hypothesis we want to test is

$$H_0 : \beta_{EG} = 0 \quad \text{vs.} \quad H_1 : \beta_{EG} \neq 0. \tag{3.5}$$

The log-likelihood function for observation i takes the form

$$l_i(\pi_i; y_i) = y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i) + \log \binom{n_i}{y_i},$$

where

$$\pi_i = \frac{\exp(\beta_0 + \beta_E x_{Ei} + \beta_G x_{Gi} + \beta_{EG} x_{Ei} x_{Gi})}{1 + \exp(\beta_0 + \beta_E x_{Ei} + \beta_G x_{Gi} + \beta_{EG} x_{Ei} x_{Gi})}.$$

The elements in the score vector are given by

$$\boldsymbol{U} = \begin{bmatrix} U_{\beta_0} \\ U_{\beta_E} \\ U_{\beta_G} \\ U_{\beta_{EG}} \end{bmatrix} = \begin{bmatrix} \sum_i (Y_i - n_i \pi_i) \\ \sum_i x_{Ei} (Y_i - n_i \pi_i) \\ \sum_i x_{Gi} (Y_i - n_i \pi_i) \\ \sum_i x_{Ei} x_{Gi} (Y_i - n_i \pi_i) \end{bmatrix}, \tag{3.6}$$

and the information matrix becomes

$$\mathfrak{J} = \begin{bmatrix} \sum_i \text{Var}(Y_i) & \sum_i x_{Ei}\text{Var}(Y_i) & \sum_i x_{Gi}\text{Var}(Y_i) & \sum_i x_{Ei}x_{Gi}\text{Var}(Y_i) \\ \sum_i x_{Ei}\text{Var}(Y_i) & \sum_i x_{Ei}^2\text{Var}(Y_i) & \sum_i x_{Ei}x_{Gi}\text{Var}(Y_i) & \sum_i x_{Ei}^2x_{Gi}\text{Var}(Y_i) \\ \sum_i x_{Gi}\text{Var}(Y_i) & \sum_i x_{Ei}x_{Gi}\text{Var}(Y_i) & \sum_i x_{Gi}^2\text{Var}(Y_i) & \sum_i x_{Ei}x_{Gi}^2\text{Var}(Y_i) \\ \sum_i x_{Ei}x_{Gi}\text{Var}(Y_i) & \sum_i x_{Ei}^2x_{Gi}\text{Var}(Y_i) & \sum_i x_{Ei}x_{Gi}^2\text{Var}(Y_i) & \sum_i x_{Ei}^2x_{Gi}^2\text{Var}(Y_i) \end{bmatrix}, \tag{3.7}$$

where $\text{Var}(Y_i) = n_i\pi_i(1-\pi_i)$.

Again, we use (2.17) and get

$$\mathfrak{J}^* = \mathfrak{J}_{22} - \mathfrak{J}_{21}\mathfrak{J}_{11}^{-1}\mathfrak{J}_{12}, \tag{3.8}$$

where

$$\mathfrak{J}_{22} = \sum_i x_{Ei}^2x_{Gi}^2n_i\pi_i(1-\pi_i),$$

$$\mathfrak{J}_{21} = \begin{bmatrix} \sum_i x_{Ei}x_{Gi}n_i\pi_i(1-\pi_i) & \sum_i x_{Ei}^2x_{Gi}n_i\pi_i(1-\pi_i) & \sum_i x_{Ei}x_{Gi}^2n_i\pi_i(1-\pi_i) \end{bmatrix},$$

$$\mathfrak{J}_{11} = \begin{bmatrix} \sum_i n_i\pi_i(1-\pi_i) & \sum_i x_{Ei}n_i\pi_i(1-\pi_i) & \sum_i x_{Gi}n_i\pi_i(1-\pi_i) \\ \sum_i x_{Ei}n_i\pi_i(1-\pi_i) & \sum_i x_{Ei}^2n_i\pi_i(1-\pi_i) & \sum_i x_{Ei}x_{Gi}n_i\pi_i(1-\pi_i) \\ \sum_i x_{Gi}n_i\pi_i(1-\pi_i) & \sum_i x_{Ei}x_{Gi}n_i\pi_i(1-\pi_i) & \sum_i x_{Gi}^2n_i\pi_i(1-\pi_i) \end{bmatrix}$$

and

$$\mathfrak{J}_{12} = \begin{bmatrix} \sum_i x_{Ei}x_{Gi}n_i\pi_i(1-\pi_i) \\ \sum_i x_{Ei}^2x_{Gi}n_i\pi_i(1-\pi_i) \\ \sum_i x_{Ei}x_{Gi}^2n_i\pi_i(1-\pi_i) \end{bmatrix}.$$

The score test statistic is then

$$S = U_{\beta_{EG}}^T(\mathfrak{J}^*)^{-1}U_{\beta_{EG}},$$

where $U_{\beta_{EG}}$ is given in (3.6), and $\mathfrak{J}^*$ is given in (3.8). The null hypothesis from (3.5) is rejected if $S > \chi_{\alpha,1}^2$.

Another hypothesis that may be of interest to test is

$$H_0 : \beta_G = \beta_{EG} = 0 \quad \text{vs.} \quad H_1 : \text{at least one of } \beta_G \text{ or } \beta_{EG} \text{ is not equal to zero.} \tag{3.9}$$

In this situation, the $\boldsymbol{\theta}_2$ vector from Section 2.6 will be

$$\boldsymbol{\theta}_2 = \begin{bmatrix} \beta_G \\ \beta_{EG} \end{bmatrix}.$$

The information matrix is as given in (3.7), but now

$$\mathfrak{J}_{22} = \begin{bmatrix} \sum_i x_{Gi}^2 n_i \pi_i (1-\pi_i) & \sum_i x_{Ei} x_{Gi}^2 n_i \pi_i (1-\pi_i) \\ \sum_i x_{Ei} x_{Gi}^2 n_i \pi_i (1-\pi_i) & \sum_i x_{Ei}^2 x_{Gi}^2 n_i \pi_i (1-\pi_i) \end{bmatrix},$$

$$\mathfrak{J}_{21} = \begin{bmatrix} \sum_i x_{Gi} n_i \pi_i (1-\pi_i) & \sum_i x_{Ei} x_{Gi} n_i \pi_i (1-\pi_i) \\ \sum_i x_{Ei} x_{Gi} n_i \pi_i (1-\pi_i) & \sum_i x_{Ei}^2 x_{Gi} n_i \pi_i (1-\pi_i) \end{bmatrix},$$

$$\mathfrak{J}_{11} = \begin{bmatrix} \sum_i n_i \pi_i (1-\pi_i) & \sum_i x_{Ei} n_i \pi_i (1-\pi_i) \\ \sum_i x_{Ei} n_i \pi_i (1-\pi_i) & \sum_i x_{Ei}^2 n_i \pi_i (1-\pi_i) \end{bmatrix},$$

and

$$\mathfrak{J}_{12} = \begin{bmatrix} \sum_i x_{Gi} n_i \pi_i (1-\pi_i) & \sum_i x_{Ei} x_{Gi} n_i \pi_i (1-\pi_i) \\ \sum_i x_{Ei} x_{Gi} n_i \pi_i (1-\pi_i) & \sum_i x_{Ei}^2 x_{Gi} n_i \pi_i (1-\pi_i) \end{bmatrix}.$$

Hence, the score test statistic becomes

$$S = \boldsymbol{U}_{\boldsymbol{\theta}_2}^T (\mathfrak{J}^*)^{-1} \boldsymbol{U}_{\boldsymbol{\theta}_2},$$

where $\boldsymbol{U}_{\boldsymbol{\theta}_2}$ is given by

$$\boldsymbol{U}_{\boldsymbol{\theta}_2} = \begin{bmatrix} U_{\beta_G} \\ U_{\beta_{EG}} \end{bmatrix},$$

and both $\mathfrak{J}^*$ and $\boldsymbol{U}_{\boldsymbol{\theta}_2}$ are evaluated at $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}_1}$ and $\boldsymbol{\theta}_2 = \mathbf{0}$. The null hypothesis in (3.9) is rejected if $S > \chi^2_{\alpha,2}$. Note that $\boldsymbol{\theta}_2$ here has two elements. This causes the score test statistic to be χ^2-distributed with two degrees of freedom, not one as for the previous situations.

3.5 The genotype variable and the environmental variable as factors

In this section, both the genotype variable and the environmental variable are considered as factors. We will not follow the structure from the previous sections, and the focus will no longer be on hypothesis testing, but rather on estimating the parameters in the model. The estimates will be given in a closed form when this is possible.

We assume that the genotype variable has three levels, and that the environmental variable has k levels, and consider a $2 \times 3 \times k$ table like the one shown in Table 3.1.

	Genotype			
	0	1	2	Sum
Case	x_{0j}	x_{1j}	x_{2j}	n_{1j}
Control	y_{0j}	y_{1j}	y_{2j}	n_{2j}
Sum	m_{0j}	m_{1j}	m_{2j}	N_j

Table 3.1: Notation for the $2 \times 3 \times k$ contingency table.

Here, j is the level of the environmental variable, and can take values in the interval $0 \leq j \leq k-1$. The genotype is either 0, 1 or 2, and the response variable has two levels; case (diseased) and control (non-diseased).

First, we take a look at the saturated model. A saturated model is a model where the number of parameters is equal to the number of covariate patterns. For logistic regression, the saturated model may be written as

$$\ln\left(\frac{f_{ij}}{1-f_{ij}}\right) = \psi_{ij}, \tag{3.10}$$

where f_{ij} is the probability of developing the disease given the levels of i and j. The index i refers to the genotype, while the index j refers to the level of the environmental variable.

By rewriting (3.10), we get

$$f_{ij} = \frac{\exp(\psi_{ij})}{1+\exp(\psi_{ij})}$$

and

$$1 - f_{ij} = \frac{1}{1+\exp(\psi_{ij})}.$$

The likelihood function is then given by

$$\begin{aligned} \mathrm{L} &= \prod_{j=0}^{k-1} \left[\left(\frac{\exp(\psi_{0j})}{1+\exp(\psi_{0j})} \right)^{x_{0j}} \left(\frac{\exp(\psi_{1j})}{1+\exp(\psi_{1j})} \right)^{x_{1j}} \left(\frac{\exp(\psi_{2j})}{1+\exp(\psi_{2j})} \right)^{x_{2j}} \right. \\ &\quad \left. \left(\frac{1}{1+\exp(\psi_{0j})} \right)^{y_{0j}} \left(\frac{1}{1+\exp(\psi_{1j})} \right)^{y_{1j}} \left(\frac{1}{1+\exp(\psi_{2j})} \right)^{y_{2j}} \right]. \end{aligned}$$

We get the maximum likelihood estimates for the parameters included by first taking the logarithm of the likelihood function, and then differentiate with respect to each of the parameters. Next, we set each expression equal to zero and solve for the parameter of interest. Using the random variables corresponding to the cells in Table 3.1, the estimates are given by

$$\begin{aligned} \exp(\hat{\psi}_{0j}) &= \frac{X_{0j}}{Y_{0j}}, \\ \exp(\hat{\psi}_{1j}) &= \frac{X_{1j}}{Y_{1j}} \quad \text{and} \\ \exp(\hat{\psi}_{2j}) &= \frac{X_{2j}}{Y_{2j}}. \end{aligned}$$

We now consider both the genotype variable and the environmental variable as factors, and rewrite the model from (3.10) to

$$\ln\left(\frac{f_{ij}}{1-f_{ij}}\right) = \alpha + \beta_i + \gamma_j + \delta_{ij}, \tag{3.11}$$

where α is the intercept, β_i is the main effect of the genotype, γ_j is the main effect of the environment variable, and the term δ_{ij} is the interaction between genotype and environment. We use β_0, γ_0 and δ_{00} as reference categories. Hence, $\beta_0 = \gamma_0 = \delta_{00} = \delta_{0j} = \delta_{i0} = 0$. From (3.11) we get

$$f_{ij} = \frac{\exp(\alpha + \beta_i + \gamma_j + \delta_{ij})}{1+\exp(\alpha + \beta_i + \gamma_j + \delta_{ij})}$$

and

$$1 - f_{ij} = \frac{1}{1+\exp(\alpha + \beta_i + \gamma_j + \delta_{ij})}.$$

The likelihood function can be expressed as

$$\begin{aligned} \mathrm{L} \;=\; & \prod_{j=0}^{k-1} \Bigg[\left(\frac{\exp(\alpha+\gamma_j)}{1+\exp(\alpha+\gamma_j)} \right)^{x_{0j}} \left(\frac{\exp(\alpha+\beta_1+\gamma_j+\delta_{1j})}{1+\exp(\alpha+\beta_1+\gamma_j+\delta_{1j})} \right)^{x_{1j}} \\ & \left(\frac{\exp(\alpha+\beta_2+\gamma_j)}{1+\exp(\alpha+\beta_2+\gamma_j+\delta_{2j})} \right)^{x_{2j}} \left(\frac{1}{1+\exp(\alpha+\gamma_j)} \right)^{y_{0j}} \\ & \left(\frac{1}{1+\exp(\alpha+\beta_1+\gamma_j+\delta_{1j})} \right)^{y_{1j}} \left(\frac{1}{1+\exp(\alpha+\beta_2+\gamma_j+\delta_{2j})} \right)^{y_{2j}} \Bigg]. \end{aligned}$$

By taking the logarithm of the likelihood function, differentiate for each parameter, and then set each expression equal to zero, we get the maximum likelihood estimates in a closed form. Using the random variables corresponding to the cells in Table 3.1, the estimates are given by

$$\exp(\hat{\alpha}) = \frac{X_{00}}{Y_{00}}, \tag{3.12}$$

$$\exp(\hat{\beta}_1) = \frac{X_{10}}{Y_{10}} \frac{Y_{00}}{X_{00}}, \tag{3.13}$$

$$\exp(\hat{\beta}_2) = \frac{X_{20}}{Y_{20}} \frac{Y_{00}}{X_{00}}, \tag{3.14}$$

$$\exp(\hat{\gamma}_j) = \frac{X_{0j}}{Y_{0j}} \frac{Y_{00}}{X_{00}}, \tag{3.15}$$

$$\exp(\hat{\delta}_{1j}) = \frac{X_{1j}}{Y_{1j}} \frac{Y_{0j}}{X_{0j}} \frac{Y_{10}}{X_{10}} \frac{X_{00}}{Y_{00}} \quad \text{and} \tag{3.16}$$

$$\exp(\hat{\delta}_{2j}) = \frac{X_{2j}}{Y_{2j}} \frac{Y_{0j}}{X_{0j}} \frac{Y_{20}}{X_{20}} \frac{X_{00}}{Y_{00}}. \tag{3.17}$$

Here, $\frac{X_{00}}{Y_{00}}$ is the relationship between the cases and the controls for the individuals carrying genotype 0 and an environmental variable at level 0. This fraction is used as a reference level for the other estimates.

The estimates given in (3.12)-(3.17) apply for the recessive and the dominant genetic models only. For the additive genetic model, the expressions can not be written in a closed form, so we need to use numerical methods. Also, if the genotype and the environmental variables are not factors, but continuous variables, we need to solve for the estimates numerically.

The model in (3.11) has, in addition to an intercept and a genotype term, both an environmental term and an interaction term. We can also find estimates for the

parameters in the models where the interaction term, or both the environmental and the interaction term, is omitted. These two possible models can be written as

$$\text{logit}(\pi_i) = \alpha + \beta_i$$

and

$$\text{logit}(\pi_i) = \alpha + \beta_i + \gamma_j.$$

The procedure for calculating the estimates for these models is similar to the one outlined for the model in (3.11).

Chapter 4

Statistical test theory

Before we continue with a power study in Chapter 5, we need to introduce some basic statistical test theory. In this chapter we will give a brief introduction to hypothesis testing, p-values, significance level and error in hypothesis testing. We will also present some helpful tools and techniques we can use to evaluate and compare several statistical methods. The content in Sections 4.1 and 4.2 is based on Chapter 8 in Castella & Berger (2002).

4.1 Hypothesis testing

A statistical hypothesis can be seen as an assumption about one or several population parameters. In order to determine whether the hypothesis is true or false, we perform what is called an hypothesis test. First we need to define the hypothesis. The null hypothesis, denoted by H_0, is usually the assumption that the observations are a result of pure chance. The alternative hypothesis, here denoted by H_1, usually says that the observations are influenced by some non-random cause.

The best way to test such an hypothesis is to examine the entire population. Since that is normally impossible, or at least impractical, we rather examine a random sample from the population. If the data from this sample is not consistent with the null hypothesis, the null hypothesis is rejected. To determine whether we do have consistency between the random sample and the null hypothesis, we calculate a test statistic, $S(X)$, and a corresponding p-value, $p(x)$. Assume that large values of $S(X)$ give evidence of H_1. The p-value is then defined as

$$p(x) = \mathrm{P}(S(X) \geq s(x)).$$

Given that the null hypothesis is true, the p-value gives information about the

probability of observing something at least as extreme as our observation. Small p-values lead to rejection of the null hypothesis. The chosen level of significance, α, determines whether the null hypothesis should be rejected or not. A p-value less than the significance level leads to rejection of the null hypothesis. In other words, the probability of rejecting the null hypothesis is less than or equal to the chosen level of significance.

When we test a statistical hypothesis, there are two types of possible errors that may occur. By errors we here mean drawing incorrect conclusions. First, we have what is called Type I error, which is defined as rejecting a true null hypothesis. Second, we have what is called Type II error, which is defined as not rejecting a false null hypothesis. A Type I error is also called a false positive error, while a Type II error sometimes is called a false negative error. Type I and Type II errors are related to each other. If the probability of a Type I error decreases, then the probability of a Type II error increases, and vice versa.

4.2 Statistical power and test size

The power of a statistical test, which we denote γ, is defined as the probability that the test rejects the null hypothesis when the null hypothesis is false. Hence, the statistical power can be expressed as 1-P(Type II error). When the probability that a Type II error occurs decreases, the power increases.

The power of a statistical method depends on several factors. One of these is the criterion level of statistical significance used in the test. An increasing of this criterion level increases the power of the method, because by doing this, the risk of a Type II error is reduced. In other words, the chance of rejecting the null hypothesis when the null hypothesis is false increases when the criterion level of statistical significance increases.

Another factor that affects the power of a method, is the sample size. It is more difficult to detect effects in smaller samples. Hence, increasing the sample size will probably increase the statistical power of a method. The magnitude of the effect or association of interest in the population under study does also have an impact on the statistical power. Larger effects are easier to detect. Hence, the statistical power increases when the size of the effect increases.

The probability of Type I error is called the statistical size of a test, or sometimes the significance level of a test.

4.3 Estimating power by simulation

The power of a test, γ, can be calculated at a parameter vector θ. In our setting,

$$\theta = (\beta_0, \beta_E, \beta_G, \text{MAF}, \mu_E, \sigma_E).$$

Here, β_0, β_E and β_G are used in the calculation of the probability of disease, while MAF, μ_E and σ_E are used in the distribution calculations for the genotype and the environmental factor. The power when using significance level α is given by

$$\gamma(\theta) = \text{P}(p(X) \leq \alpha), \tag{4.1}$$

where $p(X)$ is the p-value at θ for an observation X. However, the distribution of the p-value for a parameter vector θ is in many cases unknown. We may thus instead estimate the power by simulation. We generate a dataset x at a parameter vector θ. For this dataset we calculate a p-value associated with the test we want to perform. This p-value may be greater or smaller than a chosen significance level α. We repeat this procedure m times, and a statistical test is said to be exact if $\gamma(\theta) = \alpha$ for all α, while it is called valid if $\gamma(\theta) \leq \alpha$.

Let W be the number of p-values less than or equal to α, then $\hat{\gamma} = \frac{W}{m}$. Since $W \sim \text{bin}(m, \gamma)$, the expected value of $\hat{\gamma}$ is given by

$$\text{E}(\hat{\gamma}) = \gamma,$$

and the variance of $\hat{\gamma}$ is given by

$$\text{Var}(\hat{\gamma}) = \frac{\gamma(1-\gamma)}{m}.$$

Furthermore, γ follows approximately the normal distribution. The limits of a confidence interval for γ is thus given by

$$\gamma \pm z_\alpha \text{SE}(\gamma),$$

where α is the significance level, and z_α is the α percentile of the normal distribution. The standard error (SE) is the square root of the variance of γ.

Chapter 5

Power study

In Chapter 2, we introduced statistical methods that can be used to detect association between genotype, environmental factors and disease. These methods have different approaches to determine whether the overall hypothesis of interest should be rejected or not. The majority of the methods presented and outlined in this thesis require fitting one or several models using logistic regression, but we have also included a method which is independent of any model fitting. In the present chapter, we will look at the performance of all these methods when testing for association between genotype and disease. It will here be useful to know the true underlying model when drawing conclusions, so for this purpose we simulate datasets for a given genetic marker.

5.1 Procedure for data simulation

We start by outlining a method to simulate data in a cohort study design. This method is inspired by the procedure given in Zheng et al. (2012) (pp. 85-86).

Let X_G denote the genotype variable, X_E denote the environmental variable, and the random variable Y be defined as

$$Y = \begin{cases} 0 & \text{if non-diseased} \\ 1 & \text{if diseased} \end{cases}.$$

Then, by using the logistic regression model, the probability of developing the disease, given the genotype and the value of the environmental variable, is given by

$$f = \mathrm{P}(Y = 1 \mid X_E = x_E, X_G = x_G) = \frac{\exp(\beta_0 + \beta_E x_E + \beta_G x_G)}{1 + \exp(\beta_0 + \beta_E x_E + \beta_G x_G)}. \tag{5.1}$$

Here, we assume X_E to be normally distributed with mean μ and variance σ^2. The covariate x_G is coded 0, $\frac{1}{2}$ and 1 for genotype 0, 1 and 2, respectively, and the environmental effect and the genotype effect are assumed to be independent. The intercept parameter, β_0, is a predetermined effect, and the parameters β_E and β_G are determined by the odds ratio for the environmental and the genotype effect, respectively. From Section 2.4, we know that $\beta_j = \ln(\mathrm{OR}_j)$. Hence, $\beta_E = \ln(\mathrm{OR}_E)$ and $\beta_G = \ln(\mathrm{OR}_G)$, where OR_E is the odds ratio for the environmental factor, and OR_G is the odds ratio for the genotype.

Next, we define the probability of carrying genotype i as g_i, where $i = 0, 1, 2$. For a given MAF, these probabilities are given by

$$\begin{aligned} g_0 &= (1 - \mathrm{MAF})^2, \\ g_1 &= 2\mathrm{MAF}(1 - \mathrm{MAF}) \quad \text{and} \\ g_2 &= \mathrm{MAF}^2, \end{aligned}$$

when assuming the Hardy-Weinberg equilibrium described in Section 1.4.

For each individual, we simulate a genotype, x_G, and an environmental covariate, x_E. The genotype is simulated by drawing from $\mathrm{P}(X_G = x_G) = g_i$, $i = 0, 1, 2$, while the environmental covariate is simulated by drawing from the normal distribution with mean μ and variance σ^2. Now, when the genotype and the environmental covariate are known, we draw the disease status from a binomial distribution where the probability of disease is given by the expression in (5.1).

By following the procedure outlined above, we obtain data generated in a cohort setting. That is, the disease status is determined when the values of the exposure variables are known. For our testing and comparing of the methods from Chapter 2, we would like to use data from a case-control setting. That is, we would like to pick a number of diseased individuals, and a number of non-diseased individuals, and then check what the values of the explanatory variables are. To make a case-control dataset by using the cohort procedure, we follow the given steps, and continue to simulate until we have a predetermined number, n_0, non-diseased individuals, and another predetermined number, n_1, diseased individuals.

For each combination of the odds ratio values for the environmental factor and the odds ratio values for the genotype, we make m datasets, and for each dataset we perform the CATT, the Score test, the LRT and the Wald test. Next, for each test we count the total number of p-values less than or equal to a chosen value of α, and add up the total for all the m datasets. This gives us one total value for each test. Then, these values are divided by the total number of datasets, m, to obtain

the simulated power, as defined in (4.1), of each of the tests. The results are put into a table where we easily can see how the power changes when the odds ratio for the environmental factor and the genotype change.

5.2 Data simulation and hypothesis testing

In this section we consider the model given by

$$\text{logit}(\pi) = \beta_0 + \beta_E x_E + \beta_G x_G,$$

and test the null hypothesis that says $\beta_G = 0$. This model, and the corresponding hypothesis, is the same as the one considered in Section 3.3. Before we follow the procedure outlined above and simulate datasets for our study, we need to set the values for the input parameters. We let the environmental random variable, X_E, follow a standardized normal distribution. Hence, for each individual we draw the value of x_E from $\text{N}(0, 1)$. The intercept parameter, β_0, is set to -3.5, and the MAF value is set to 0.3. The choice of these values are to be discussed later. To determine the values of the parameters β_E and β_G, we use odds ratio values in the range from 1 to 20 for the environmental factor, and from 1.0 to 1.8 for the genotype.

Furthermore, we use $m = 10\ 000$, $n_0 = n_1 = 1\ 000$ and $\alpha = 0.05$. The statistical power obtained from the simulated datasets for each of the four methods are shown in Tables 5.1 (CATT), 5.2 (Score test), 5.3 (LRT) and 5.4 (Wald test).

When using $m = 10\ 000$, obtaining a test size of 0.05 gives a standard error of 2.18e-3. This is calculated using the expression obtained in Section 4.3. Similarly, a power of 80% will have a standard error of 4.0e-3.

	OR_E					
OR_G	1.0	1.5	2.0	5.0	10	20
1.0	0.0497	0.0496	0.0553	0.0479	0.0491	0.0494
1.2	0.2616	0.2605	0.2633	0.1874	0.1355	0.1055
1.4	0.6939	0.6937	0.6780	0.5164	0.3530	0.2367
1.6	0.9356	0.9319	0.9209	0.7959	0.5932	0.4165
1.8	0.9910	0.9899	0.9886	0.9322	0.7887	0.5963

Table 5.1: Statistical power for the CATT when $m = 10\ 000$, $n_0 = n_1 = 1\ 000$, MAF = 0.3 and $\alpha = 0.05$.

	OR_E					
OR_G	1.0	1.5	2.0	5.0	10	20
1.0	0.0496	0.0499	0.0536	0.0479	0.0505	0.0502
1.2	0.2628	0.2547	0.2482	0.1787	0.1599	0.1331
1.4	0.6938	0.6758	0.6463	0.5095	0.4191	0.3477
1.6	0.9361	0.9249	0.9048	0.7876	0.6860	0.6015
1.8	0.9912	0.9879	0.9829	0.9325	0.8755	0.7927

Table 5.2: Statistical power for the Score test when $m = 10\ 000$, $n_0 = n_1 = 1\ 000$, MAF $= 0.3$ and $\alpha = 0.05$.

	OR_E					
OR_G	1.0	1.5	2.0	5.0	10	20
1.0	0.0497	0.0500	0.0539	0.0480	0.0505	0.0503
1.2	0.2630	0.2549	0.2484	0.1789	0.1603	0.1338
1.4	0.6941	0.6766	0.6465	0.5103	0.4199	0.3492
1.6	0.9362	0.9251	0.9048	0.7880	0.6872	0.6025
1.8	0.9912	0.9879	0.9830	0.9325	0.8757	0.7934

Table 5.3: Statistical power for the LRT when $m = 10\ 000$, $n_0 = n_1 = 1\ 000$, MAF $= 0.3$ and $\alpha = 0.05$.

	OR_E					
OR_G	1.0	1.5	2.0	5.0	10	20
1.0	0.0494	0.0496	0.0534	0.0474	0.0503	0.0498
1.2	0.2625	0.2544	0.2473	0.1784	0.1591	0.1327
1.4	0.6932	0.6753	0.6461	0.5086	0.4176	0.3469
1.6	0.9359	0.9249	0.9046	0.7863	0.6854	0.5999
1.8	0.9912	0.9878	0.9829	0.9320	0.8755	0.7917

Table 5.4: Statistical power for the Wald test when $m = 10\ 000$, $n_0 = n_1 = 1\ 000$, MAF $= 0.3$ and $\alpha = 0.05$.

5.3 Comparison of the performance of the Score test, the LRT and the Wald test

From Tables 5.2-5.4, we observe that the Score test, the LRT and the Wald test give very similar results. The LRT seems to be the method obtaining the highest statistical power, but by adding 95% confidence intervals, we observe that the difference in the performance of these three methods is not statistically significant. For both the LRT and the Wald test, two models need to be fitted for each dataset, while the Score test requires only one model fitting. This gives the Score test reduced computational time compared to the LRT and the Wald test. Based on the fact that the three methods seem to draw the same conclusion about the null

hypothesis, and that the Score test has a computational advantage, we consider the Score test as the preferred one among these three. In the following analysis, we will therefore neither include the LRT nor the Wald test, but focus on the Score test in addition to the CATT.

5.4 Performance of the CATT and the Score test

Table 5.1 shows how the power for the CATT varies for several combinations of the effect size for the environmental factor and for the genotype. The null hypothesis says that the genotype has no effect on the risk of develop the disease, which is equivalent to $\text{OR}_G = 1$. The level of significance is set to 0.05. Thus, the test is said to hold its level if the test size is 0.05. As we can see from the first row of Table 5.1, all values are close to 0.05. In order to determine whether the values are close enough for the test to be considered as a test which holds its level, we calculate confidence intervals for each value of the statistical test size. The lower and upper bounds of these confidence intervals for the CATT are shown in Table 5.5.

	OR_E					
	1.0	1.5	2.0	5.0	10	20
Lower bound	0.0454	0.0453	0.0508	0.0437	0.0449	0.0452
Upper bound	0.0540	0.0539	0.0598	0.0521	0.0533	0.0536

Table 5.5: Lower and upper bounds of 95% confidence intervals for the test size of the CATT for different values of the environmental effect.

As we can see from Table 5.5, 0.05 is included in all the given 95% confidence intervals, except for the one where the environmental odds ratio is equal to 2.0. Here, 0.05 is just outside the interval. Despite this, we conclude that the CATT holds its level for the environmental odds ratio values and the other parameter values we have used in this simulation.

If we again consider Table 5.1, we observe that as the effect of the environmental factor increases, the effect of the genotype becomes harder to detect. When the odds ratio for the environmental factor is 20, which is relatively high in this setting, the probability that the CATT rejects the null hypothesis when it is false is about 60% when the odds ratio for the genotype is 1.8. For environmental odds ratios less than or equal to 5, the power of the CATT is greater than 80% when the genotype odds ratio is at least 1.6.

Similarly as for the CATT, we calculate confidence intervals for the test size of the Score test. These intervals are given in Table 5.6.

From Table 5.6 we observe that 0.05 is included in all the given 95% confidence intervals of the test size for the Score test. Hence, the Score test holds its level for

	OR_E					
	1.0	1.5	2.0	5.0	10	20
Lower bound	0.0453	0.0456	0.0492	0.0437	0.0462	0.0459
Upper bound	0.0539	0.0542	0.0580	0.0521	0.0548	0.0545

Table 5.6: Lower and upper bounds of 95% confidence intervals for the test size of the Score test for different values of the environmental effect.

the environmental odds ratio values and the other parameter values we have used in this simulation.

As the squared CATT statistic is equal to the Score test statistic for logistic regression when genotype is the only covariate, the first column in the Score table (Table 5.2), is close to identical to the first column in the CATT table (Table 5.1). The reason why these two columns are not exactly identical is due to the parameter β_E. For the CATT, this parameter is not included at all, which is equivalent to set $\beta_E = 0$, while for the Score test, we have to estimate β_E. This estimate, $\hat{\beta}_E$, will be close to zero, but not necessarily equal to zero, which leads to a slightly different test statistic for the CATT and the Score test.

Similarly as for the CATT, the effect of the genotype is harder to detect for the Score test when the effect from the environmental factor increases. For the odds ratio equal to 1.4 for the genotype, the power decreases to the half from about 70% to about 35% when the odds ratio for the environmental factor increases from 1.0 to 20.

If we consider both the performance of the CATT and of the Score test, we can see that the power follows the same trend for the two methods. It increases as the effect of the genotype increases, and decreases as the effect of the environmental effect increases. This is not a very surprising result, but what is of greater interest is what we observe when we compare the performance of the two methods. Here, the CATT has greater statistical power than the Score test when the odds ratio effect from the environment is ≤ 5. For the rest of the environmental odds ratios included in this simulation, the Score test rejects the null hypothesis more often than the CATT does, which means it detects the effect of the genotype more frequently. We have not been able to find any articles or other available literature where this finding is presented. To determine whether the results from the two methods are significantly different or not, a paired hypothesis test strategy is needed. We may apply a test called McNemar's test for this purpose, but we have not included any such tests in our study. This would require counting of the number of discordant pairs of conclusions for the CATT and the Score test for a given combination of the genotype and environmental effect sizes. Also, we have not performed any analysis about whether the estimates $\hat{\beta}_G$ are closer to the true value when using a logistic regression model without environmental influence.

When calculating the CATT test statistic, the information about the environmen-

tal influence is ignored. The Score test, on the other hand, takes this additional information into account when the test statistic is computed. One might expect that using more of the available information will lead to greater statistical power, but this seems not to be true for all combinations of the environmental- and genotype effects in our analysis. When an additional variable is taken into account, the uncertainty becomes greater, and an increased uncertainty leads to decreased statistical power.

A graphical representation of how the statistical power varies for different effects sizes for the genotype and the environmental factor are shown in Figures 5.1-5.4.

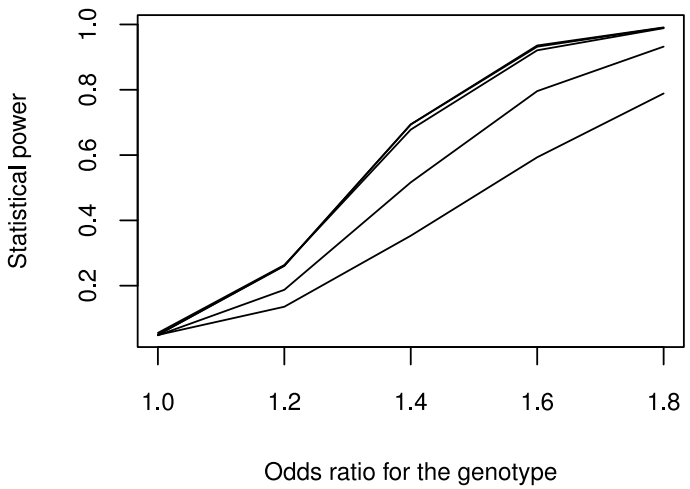

Figure 5.1: Statistical power obtained for CATT for several genotype odds ratio values. The lines, from top to bottom, represent odds ratio values from 1.0 to 20 for the environmental effect.

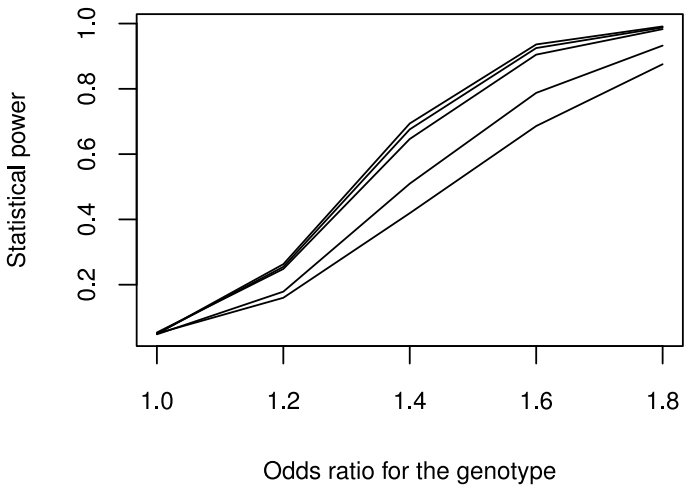

Figure 5.2: Statistical power obtained for Score test for several genotype odds ratio values. The lines, from top to bottom, represent odds ratio values from 1.0 to 20 for the environmental effect.

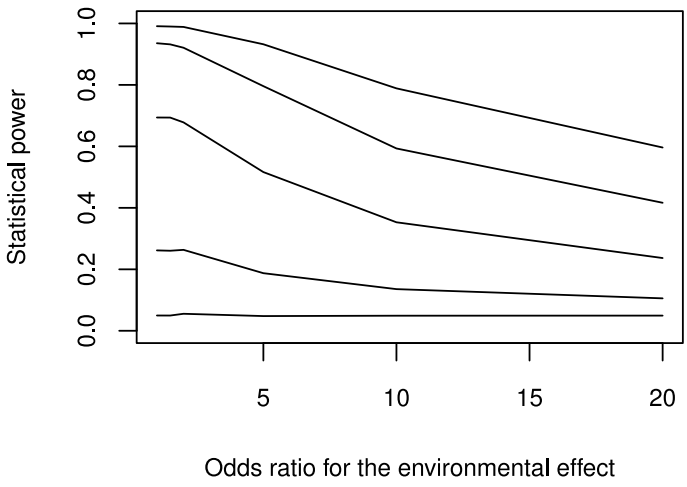

Figure 5.3: Statistical power obtained for CATT for several environmental odds ratio values. The lines, from top to bottom, represent odds ratio values from 1.8 to 1.0 for the genotype effect.

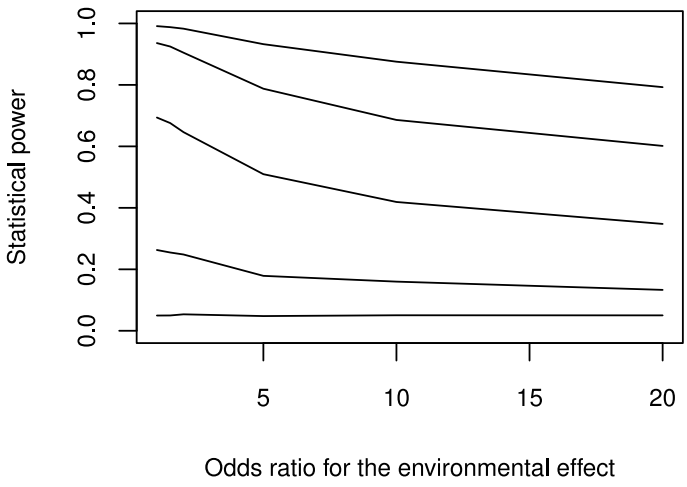

Figure 5.4: Statistical power obtained for Score test for several environmental odds ratio values. The lines, from top to bottom, represent odds ratio values from 1.8 to 1.0 for the genotype effect.

5.5 Effect of parameter values

There are several parameters that have an impact on the power of a statistical test. We are here going to discuss these, and observe how the power of the methods changes when we use different values for some of these parameters in our simulations.

5.5.1 Intercept parameter

For the simulations in Section 5.2, we used an intercept value equal to -3.5. This value was set such that the probability of disease is in the interval [0.03, 0.20] for all combinations of the chosen odds ratio values for both the environmental factor and the genotype. How to obtain an expression for the probability of disease is shown in Chapter 6. A change in the MAF will barely influence the probability of disease, thus -3.5 seems to be an appropriate choice for the intercept parameter value independent of the MAF.

5.5.2 Minor allele frequency

The MAF in the population under study was in our simulation set to 0.3. This value was inspired by the MAF value in the data used in Gabrielsen (2013) (paper III). We assumed the Hardy-Weinberg equilibrium, and used the MAF value to determine the genotype frequencies. Figure 5.5 shows how the probabilities of carrying the possible genotypes change when the MAF value changes.

From Figure 5.5, we see that an increased MAF leads to increased probability of genotype 2. Also, we can observe that a MAF equal to 0.5 gives the most balanced probabilities for the three genotypes, and we experience a symmetry in the plot about this value.

In order to see how the statistical power for the CATT and the Score test changes when the MAF changes, we use fixed values for the effects, and MAF values in the range from 0.05 to 0.60. The results are shown in Table 5.7. Here the odds ratio for the environmental factor is set to 2, and the odds ratio for the genotype is set to 1.5.

MAF	0.05	0.10	0.15	0.20	0.30	0.40	0.50	0.60
CATT	0.2928	0.4947	0.6364	0.7190	0.8349	0.8741	0.8827	0.8591
Score test	0.2797	0.4678	0.6112	0.6919	0.8115	0.8510	0.8600	0.8364

Table 5.7: Power of the CATT and the Score test for different minor allele frequencies. The odds ratio for the environmental factor is equal to 2, and the odds ratio for the genotype is equal to 1.5. The level of significance used is 0.05.

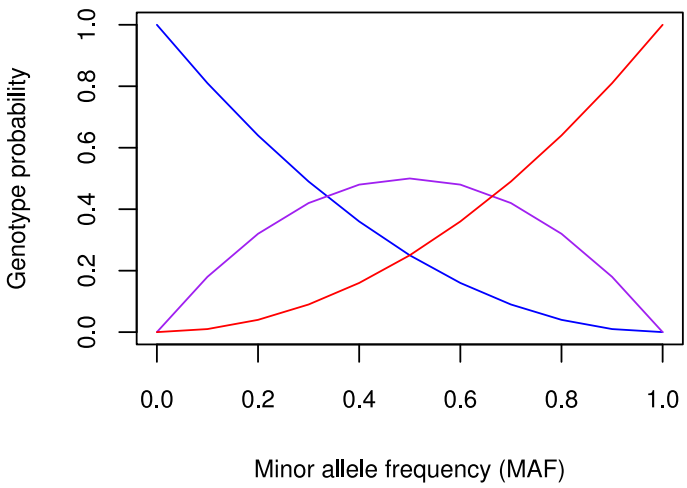

Figure 5.5: The probability of carrying each of the genotypes 0 (blue line), 1 (purple line) and 2 (red line) for different values of the minor allele frequency.

As we can see from Table 5.7, the power increases as the MAF increases. The MAF of 0.3 is the smallest out of the tested values which gives power greater than 80% for both the CATT and the Score test. For other values of the environmental effect, we may experience slightly different powers, but the trend is the same. Thus, increased MAF implies increased power.

If we estimate the genotype parameter, $\hat{\beta}_G$, and its standard error for several values of MAF, we observe that the standard error decreases as MAF increases. The greater the standard error is, the harder it is to use the estimate to draw a conclusion. This explains why the power increases when the MAF increases.

5.5.3 Distribution of the environmental variable

As mentioned in Chapter 1, it can be several environmental factors that influence the risk of developing a disease. These factors may vary depending on the disease. Instead of making a model where several such factors are included, we have chosen to combine them into one common, standard normal distributed variable. Hence, in the general population, X_E is normal distributed with mean 0 and variance equal to 1. Because the environment has an impact on the risk of developing disease, the

distribution of X_E in each of the case and control group is not identical.

5.5.4 Size of the effects

The effect size of the covariates included in the model has, of course, a great impact on the statistical power of a method. We have in our study included an effect from the environment, and an effect from the genotype. Later, we will also test hypotheses where an interaction between the genotype and the environment is considered as well. The effect of the environmental factor in the model may vary depending on the exposures included in the factor and which disease we are studying. Smoking habits, exposure to air or water pollution, diet, physical activity and age are all well known examples of environmental factors that might influence the risk of developing certain diseases. The influence can be in a positive or a negative direction, and a factor which has a positive impact on the risk of developing one disease, can for another disease provide a negative impact.

Obviously, when determining the β parameter values in the model, the distribution of the corresponding variable is of great importance. Remember, if we denote a covariate x_j and its parameter β_j, a change from x_j to $x_j + 1$, gives an odds ratio equal to $\exp(\beta_j)$. Note that an odds ratio equal to 1.0 indicates that there is no effect.

In order to choose effect sizes for our simulation which reflects what is observed in other studies, we have been searching in the available literature for such values. Guan et al. (2012) include an overview of the genotype odds ratio values obtained in studies where an association between type 2 diabetes susceptibility and common variants was detected. Here we find odds ratio values in the range from 1.05 to 1.40. The sample sizes in these studies are in the order of 50 000. Our choice of effect sizes is motivated from the studies referred to in this article, but there is a tradeoff because we use a smaller sample size. Thus, we have included genotype effects given by odds ratio values in the range from 1.0 to 1.8. Since $\beta = \ln(\text{OR})$, this implies that β_G will take values from 0 to about 0.6. The genotype covariate is coded $0, \frac{1}{2}, 1$, so carrying genotype 2 compared to genotype 0 implies an increase given by $\ln(\text{OR}_G)$ for the $\text{logit}(\pi(x))$, where OR_G is the given genotype odds ratio.

When it comes to the effect size for the environmental influence, it turned out that most articles within this field of study do not specify which effect sizes they observe. Our general impression is that an environmental odds ratio equal to 5 when considering a standard normal distributed variable, is rather high. For most common diseases, we would expect a smaller effect, but we can not exclude the possibility of observing an environmental influence with a corresponding odds ratio values much greater than 5. We have therefore in our simulation used odds ratio values in the range from 1.0 to 20, which gives β_E values from 0 to 3.0. This can be interpreted as for an individual with environmental covariate equal to 1, the $\text{logit}(\pi(x))$ function will be $\ln(\text{OR}_E)$ greater than for an individual with

environmental covariate equal to 0, when OR_E is the given environmental odds ratio.

5.5.5 Level of significance

In GWA studies, numerous hypotheses are tested simultaneously. Each null hypothesis which is tested has a corresponding level of significance. This level is often set to 0.05, which means that the probability of making a Type I error is less than 5%. It is of interest to not just control the error rate for each test, but also the overall error rate. A possible way to do this is by using what is called the familywise error rate (FWER). The FWER is defined as the probability of at least one Type I error in total (Ge et al., 2003). If we use the significance level of 0.05 for each individual hypothesis test, the FWER will be much greater than 0.05. Hence, if we want an overall level of significance, α^*, of 0.05, we have to decrease the significance level for each hypothesis test. A common, but very conservative, method that may be used to control the FWER, is named the Bonferroni correction. To perform a such correction, we divide the desired overall significance level, α^*, by the number of tests, m. Thus, the new p-value is given by $\frac{\alpha^*}{m}$. Hence, if we use $\alpha^* = 0.05$ and $m = 10\ 000$, then we have to use $\alpha =$ 5e-6 as significance level for each test in order to control the FWER. In GWA studies, the p-values usually have to be about 1e-7 or 1e-8 to be considered as significant.

In the simulation performed in Section 5.2, we used $\alpha = 0.05$. We will also use $\alpha =$ 5e-6 in some of the following simulations. In the cases where the significance level is decreased, the FWER will be controlled in a greater extent.

5.5.6 Sample size

The number of participants in the case group and in the control group is also of importance for the statistical power of a test. Table 5.8 shows how the power of the CATT and the Score test changes when the number of individuals in the two groups changes. Here, the effects of the genotype and the environmental factor are fixed. In Figure 5.6, the changes are shown graphically.

	1 000	2 000	3 000	4 000	5 000	10 000
CATT	0.0499	0.3286	0.6912	0.9020	0.9755	1.0000
Score test	0.0417	0.2788	0.6271	0.8621	0.9601	1.0000

Table 5.8: Statistical power for CATT and Score test for different sample sizes when the level of significance is set to 5e-6. The odds ratio for the environmental factor is equal to 2, the odds ratio for the genotype is equal to 1.5, and the MAF is equal to 0.3.

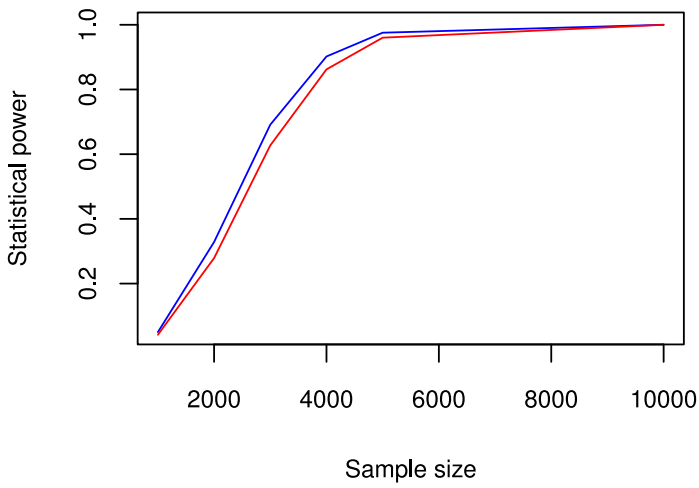

Figure 5.6: Statistical power for CATT (blue line) and Score test (red line) for different sample sizes when the level of significance is set to 5e-6. The odds ratio for the environmental factor is equal to 2, the odds ratio for the genotype is equal to 1.5, and the MAF is equal to 0.3.

Figure 5.6 shows that we need about 4 000 individuals in each of the case and control group to obtain statistical power of at least 80% when the level of significance is set to 5e-6, and the MAF is equal to 0.3. This yields for both the CATT and the Score test. For the effect sizes used here, CATT has slightly greater power than the Score test for any number of participants in the study, but this may change if using different effect sizes.

5.6 Simulation inspired by the TOP study

The Thematically Organized Psychosis (TOP) study is an ongoing research study launched in Oslo, Norway, in the year of 2003 (TOP, 2013). It is a joint collaboration between Oslo University Hospital and the University of Oslo. The participants are mainly individuals living in the Oslo area, but in the later years, also individuals from other parts of the country have been invited to participate. The aim of the TOP study is to identify the causes for a selection of mental illnesses with main focus on schizophrenia and bipolar disorder. The database now includes more than

1 100 diseased individuals which constitute the case group, and about 500 healthy individuals in a control group. The TOP study is a complex study where the participants are taken through clinical examinations, neuropsychological tests, magnetic resonance imaging (MRI) and genetic analyses to collect the necessary information.

In the simulation we performed in Section 5.2, we used an equal number of cases and controls. For the TOP study, we have about twice as many cases than controls. In order to see how this influences the power of the CATT and the Score test, we will perform another simulation where $n_0 = 500$ and $n_1 = 1\ 000$.

From Halle (2012) (Figure 5.1, p. 43) we observe that for chromosome 22, the density of the MAF distribution decreases slightly when the MAF increases, but it seems like there is no clear choice of which MAF it would be appropriate to use. We will therefore use a selection of MAF values for our simulation. This selection contains the MAF values 0.05, 0.1, 0.3 and 0.5. When it comes to the level of significance, we choose to use α equal to 5e-6. From Tables 5.1 and 5.2 we know that it is hard to detect the genotype effect when it is small, so we perform this simulation for the odds ratio values 2, 3 and 5 for the genotype. For the environmental effect, we use odds ratio equal to 2 and 10. The power values obtained are given in Tables 5.9 (CATT) and 5.10 (Score test).

From Tables 5.9 and 5.10, we observe the same trends as in previous simulations for both methods; high MAF values, small environmental effects and high genotype effects give the greatest power values. Also, when the environmental odds ratio is equal to 2, the CATT performs best, but when it is equal to 10, the Score test gives the best results.

		OR_E	
MAF	OR_G	2	10
0.05	2	0.0026	0.0000
	3	0.0626	0.0020
	5	0.5997	0.0390
0.1	2	0.0229	0.0009
	3	0.4102	0.0267
	5	0.9829	0.2790
0.3	2	0.2900	0.0219
	3	0.9743	0.3118
	5	1.0000	0.9187
0.5	2	0.4122	0.0384
	3	0.9866	0.4373
	5	1.0000	0.9561

Table 5.9: Statistical power for the CATT when $n_0 = 500$, $n_1 = 1\ 000$ and $\alpha = 5\text{e-}6$.

		OR_E	
MAF	OR_G	2	10
0.05	2	0.0022	0.0003
	3	0.0559	0.0069
	5	0.5699	0.1212
0.1	2	0.0192	0.0028
	3	0.3778	0.0672
	5	0.9784	0.5791
0.3	2	0.2572	0.0590
	3	0.9638	0.5690
	5	1.0000	0.9942
0.5	2	0.3751	0.0877
	3	0.9817	0.6982
	5	1.0000	0.9980

Table 5.10: Statistical power for the Score test when $n_0 = 500$, $n_1 = 1\ 000$ and $\alpha = 5\text{e-}6$.

5.7 Data simulation and hypothesis testing when an interaction effect is present

In Section 5.2 we assumed there were no interaction between the genotype and the environmental factor. We have to assume the possibility of situations where some kind of correlation between the effect of the genotype and the effect of the environmental factor is present. As an example, we may experience that a given SNP will affect the nicotine addiction for certain individuals (Gabrielsen, 2013, paper III). Hence, the genotype an individual is carrying will affect the size of the effect from the environmental variable, provided that this includes nicotine habits. In such situations, including an interaction effect in the logistic regression model, may give a more accurate result. We are here going to perform a power study where an interaction term is included in the model. As mentioned in Section 3.4, there are two hypotheses that may be of interest to test for this model. These two hypotheses are given in Table 5.11.

No	Hypothesis	
1	H_0:	$\beta_{EG} = 0$
	H_1:	$\beta_{EG} \neq 0$
2	H_0:	$\beta_G = \beta_{EG} = 0$
	H_1:	at least one of β_G or β_{EG} is not equal to zero

Table 5.11: Hypotheses for the model where an interaction term is included.

The Score test and the LRT can be used to test both hypothesis 1 and 2 in Table

5.11, while the Wald test only can be used when testing hypothesis 1, because it can only test one parameter at the time.

We follow the procedure outlined in Section 5.1 to simulate data, but now we also include an interaction term $\beta_{EG}x_Ex_G$. The intercept parameter, β_0, is set to -3.5 as in the previous simulations. This value may not be optimal when the model we are fitting includes an interaction term, but in order to be able to compare the results with the previous simulations, we use $\beta_0 = -3.5$ here as well. Furthermore, we use odds ratio values equal to 1.4 and 1.8 for the interaction parameter β_{EG}. For the environmental parameter, β_E, we now use odds ratio values in the range from 1 to 15, while for the genotype parameter, β_G, we still use odds ratio values in the range from 1.0 to 1.8. As for the simulation in Section 5.2, we use $n_0 = n_1 = 1\ 000$ and $m = 10\ 000$. Also, MAF = 0.3 and $\alpha = 0.05$. The results from the hypothesis tests are again displayed in terms of the statistical power obtained for each method. These results are shown in Tables 5.12 (Score test), 5.14 (LRT) and 5.14 (Wald test) for hypothesis 1, and in Tables 5.15 (Score test) and 5.16 (LRT) for hypothesis 2.

		OR_E			
OR_{EG}	OR_G	1	2	5	15
1.4	1.0	0.6686	0.5492	0.2806	0.1319
	1.2	0.6869	0.5464	0.2792	0.1434
	1.5	0.6942	0.5404	0.2805	0.1362
	1.8	0.6855	0.5541	0.2863	0.1332
1.8	1.0	0.9882	0.9442	0.6348	0.2970
	1.2	0.9882	0.9432	0.6418	0.2971
	1.5	0.9891	0.9418	0.6559	0.2985
	1.8	0.9897	0.9424	0.6473	0.3115

Table 5.12: Statistical power for the Score test when hypothesis 1 from Table 5.11 is tested. Here, $m = 10\ 000$, $n_0 = n_1 = 1\ 000$, MAF = 0.3 and $\alpha = 0.05$.

		OR_E			
OR_{EG}	OR_G	1	2	5	15
1.4	1.0	0.6713	0.5565	0.2892	0.1411
	1.2	0.6885	0.5541	0.2907	0.1512
	1.5	0.6956	0.5473	0.2890	0.1448
	1.8	0.6877	0.5604	0.2938	0.1402
1.8	1.0	0.9882	0.9466	0.6466	0.3104
	1.2	0.9882	0.9447	0.6511	0.3112
	1.5	0.9892	0.9438	0.6650	0.3129
	1.8	0.9903	0.9444	0.6542	0.3241

Table 5.13: Statistical power for the LRT when hypothesis 1 from Table 5.11 is tested. Here, $m = 10\ 000$, $n_0 = n_1 = 1\ 000$, MAF = 0.3 and $\alpha = 0.05$.

		OR_E			
OR_{EG}	OR_G	1	2	5	15
1.4	1.0	0.6668	0.5481	0.2795	0.1315
	1.2	0.6861	0.5449	0.2782	0.1424
	1.5	0.6924	0.5394	0.2799	0.1347
	1.8	0.6843	0.5531	0.2852	0.1321
1.8	1.0	0.9881	0.9440	0.6333	0.2951
	1.2	0.9881	0.9430	0.6402	0.2950
	1.5	0.9890	0.9416	0.6543	0.2977
	1.8	0.9897	0.9420	0.6466	0.3094

Table 5.14: Statistical power for the Wald test when hypothesis 1 from Table 5.11 is tested. Here, $m = 10\ 000$, $n_0 = n_1 = 1\ 000$, MAF $= 0.3$ and $\alpha = 0.05$.

When hypothesis 1 in Table 5.11 is tested, all the three methods give very similar results. As mentioned previously, to determine whether the difference is significant or not, a paired hypothesis strategy is required. When the main environmental effect and the interaction effect are fixed, we notice that the statistical power only experiences minor changes when the main genotype effect increases. If the environmental effect or the interaction effect increases, the power will undergo greater changes. Hence, the environmental effect and the interaction effect influence the performance of the methods in a greater extent. To obtain statistical power greater than 80%, the environmental impact has to be non-existent or small (odds ratio ≤ 2). When its odds ratio is equal to 15, all the three relevant methods give poor results, independent of the effect from the genotype and the interaction. By comparing the three methods, we observe that the LRT is the one that provides greatest statistical power for all combinations of the effect sizes. The Wald test turns out to be the method with the overall poorest performance.

As we can see from Table 5.15 and Table 5.16, testing of hypothesis 2 from Table 5.11 gives great statistical power for both the Score test and the LRT. Here we even obtain power values greater than 80% for some effect combinations where the environmental odds ratio is equal to 15. The probability of rejecting the false null hypothesis is overall very high, which indicates that the Score test and the LRT perform well when testing for main genotype effect and interaction effect at the same time.

		OR_E			
OR_{EG}	OR_G	1	2	5	15
1.4	1.0	0.5664	0.5048	0.3443	0.1999
	1.2	0.7315	0.7558	0.6325	0.4567
	1.5	0.9574	0.9687	0.9305	0.7987
	1.8	0.9980	0.9984	0.9940	0.9526
1.8	1.0	0.9749	0.9452	0.8043	0.5319
	1.2	0.9919	0.9885	0.9408	0.7779
	1.5	0.9996	0.9998	0.9938	0.9529
	1.8	1.0000	1.0000	0.9999	0.9918

Table 5.15: Statistical power for the Score test when hypothesis 2 from Table 5.11 is tested. Here, $m = 10\ 000$, $n_0 = n_1 = 1\ 000$, MAF $= 0.3$ and $\alpha = 0.05$.

		OR_E			
OR_{EG}	OR_G	1	2	5	15
1.4	1.0	0.5693	0.5148	0.3576	0.2114
	1.2	0.7336	0.7614	0.6408	0.4690
	1.5	0.9580	0.9699	0.9331	0.8038
	1.8	0.9980	0.9984	0.9941	0.9551
1.8	1.0	0.9760	0.9484	0.8147	0.5461
	1.2	0.9925	0.9890	0.9430	0.7859
	1.5	0.9996	0.9998	0.9941	0.9550
	1.8	1.0000	1.0000	0.9999	0.9921

Table 5.16: Statistical power for the LRT when hypothesis 2 from Table 5.11 is tested. Here, $m = 10\ 000$, $n_0 = n_1 = 1\ 000$, MAF $= 0.3$ and $\alpha = 0.05$.

Since the results from the test of hypothesis 2 gave power values close, or equal, to 1 for several combinations of the effect sizes, we would like to perform another simulation with α equal to 5e-6. By using this significance level, we can control the FWER as well as the level of each hypothesis test. The statistical power obtained from this simulation is given in Tables 5.17 (Score test) and 5.18 (LRT). To get power greater than 80% here, both the odds ratio for the genotype and for the interaction have to be at least 1.8, and the environmental effect has to be relatively small (odds ratio ≤ 5).

		OR_E			
OR_{EG}	OR_G	1	2	5	15
1.4	1.0	0.0074	0.0034	0.0008	0.0001
	1.2	0.0246	0.0245	0.0100	0.0023
	1.5	0.2011	0.2384	0.1275	0.0376
	1.8	0.6154	0.6652	0.4414	0.1724
1.8	1.0	0.2634	0.1405	0.0342	0.0026
	1.2	0.4081	0.3438	0.1402	0.0262
	1.5	0.7402	0.7523	0.4904	0.1625
	1.8	0.9453	0.9539	0.8208	0.4329

Table 5.17: Statistical power for the Score test when hypothesis 2 from Table 5.11 is tested. Here, $m = 10\ 000$, $n_0 = n_1 = 1\ 000$, MAF $= 0.3$ and $\alpha =$ 5e-6.

		OR_E			
OR_{EG}	OR_G	1	2	5	15
1.4	1.0	0.0087	0.0045	0.0013	0.0004
	1.2	0.0270	0.0304	0.0125	0.0037
	1.5	0.2111	0.2581	0.1430	0.0458
	1.8	0.6257	0.6820	0.4644	0.1895
1.8	1.0	0.2852	0.1704	0.0468	0.0045
	1.2	0.4312	0.3860	0.1679	0.0339
	1.5	0.7583	0.7787	0.5284	0.1890
	1.8	0.9498	0.9591	0.8426	0.4693

Table 5.18: Statistical power for the LRT when hypothesis 2 from Table 5.11 is tested. Here, $m = 10\ 000$, $n_0 = n_1 = 1\ 000$, MAF $= 0.3$ and $\alpha =$ 5e-6.

The two plots in Figure 5.7 show the performance of the Score test for different values of the environmental effect when hypothesis 1 (left plot) and 2 (right plot) from Table 5.11 are tested. In the left plot, the red and the blue lines are clearly separated, while in the right plot they are overlapping.

Figure 5.8 includes two plots showing the performance of the Score test for different values of the genotype effect when hypothesis 1 (left plot) and 2 (right plot) from Table 5.11 are tested. Here we observe that while the power is about constant in

the left plot, we experience a major increase for some values of the environmental effect in the right plot.

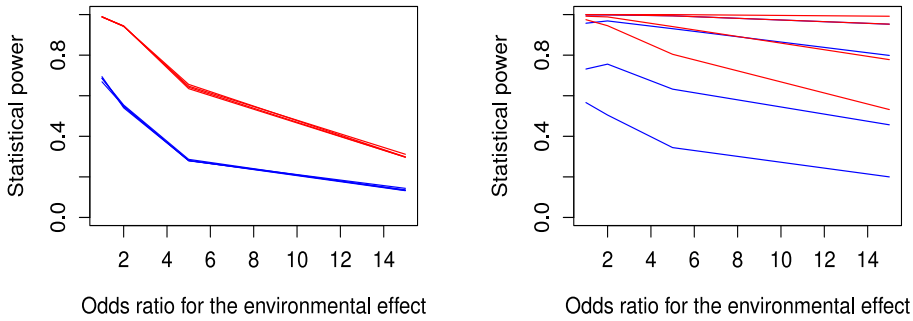

Figure 5.7: Statistical power obtained for different values of the environmental effect for the Score test when testing hypothesis 1 (left plot) and hypothesis 2 (right plot) from Table 5.11. The lines, from top to bottom within each color, represent genotype odds ratio values from 1.8 to 1.0. For the blue lines, the interaction odds ratio is equal to 1.4, while for the red lines, the interaction odds ratio is equal to 1.8.

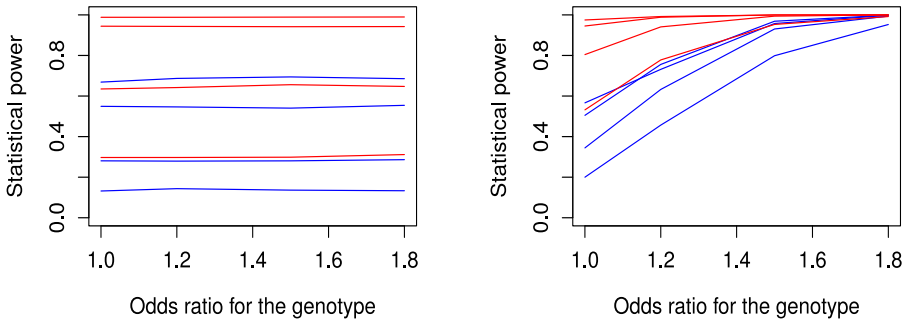

Figure 5.8: Statistical power obtained for different values of the genotype effect for the Score test when testing hypothesis 1 (left plot) and hypothesis 2 (right plot) from Table 5.11. The lines, from top to bottom within each color, represent environmental odds ratio values from 1.0 to 15. For the blue lines, the interaction odds ratio is equal to 1.4, while for the red lines, the interaction odds ratio is equal to 1.8.

Chapter 6

Simulation of case-control data

In Section 5.2 we constructed datasets by first simulating cohort data. From this, we used n_0 of the non-diseased individuals, and n_1 of the diseased individuals for our power study. It is possible to simulate case-control data directly, but when several covariates are considered, this requires a more complex procedure.

Zheng et al. (2012) (pp. 85-86) present a method to simulate case-control data. This method may work for some special cases, but it is not a general approach to obtain case-control data. Using this procedure requires available data with corresponding parameter values. We wanted to make a general setup, so we had to find a different way to perform the data simulation. The procedure outlined in Section 5.1 provides what we want, but we would like to present a more elegant method where case-control data is simulated directly. First, we consider a situation where the genotype is the only covariate included. Note that the notation in this chapter deviates from the notation used in previous chapters.

6.1 Method to simulate case-control data when the genotype is the only covariate

Let g_0, g_1 and g_2 denote the probability of carrying genotype 0, 1 and 2, respectively. If the MAF is known, and we assume the Hardy-Weinberg equilibrium, these probabilities can be calculated as shown in Section 5.1. Further, let β_0 and β_G be predetermined effect sizes for the intercept parameter and the genotype parameter, respectively. If X_G denotes the genotype variable, and the random variable Y is defined as

$$Y = \begin{cases} 0 \text{ if control} \\ 1 \text{ if case} \end{cases},$$

then the conditional probability of X_G given $Y = y$ can be expressed as

$$\mathrm{P}(X_G = i \mid Y = y) = \frac{\mathrm{P}(Y = y \mid X_G = i)\mathrm{P}(X_G = i)}{\sum_i \mathrm{P}(Y = y \mid X_G = i)\mathrm{P}(X_G = i)}, \tag{6.1}$$

where $i = 0, \frac{1}{2}, 1$, represents the genotypes 0, 1 and 2, respectively. Next, we denote the probability of developing the disease, given the genotype, as

$$f_i = \mathrm{P}(Y = 1 \mid X_G = i) = \frac{\exp(\beta_0 + \beta_G i)}{1 + \exp(\beta_0 + \beta_G i)}. \tag{6.2}$$

The overall probability of developing the disease is denoted k, which furthermore is given by

$$k = \mathrm{P}(Y = 1) = f_0 g_0 + f_1 g_1 + f_2 g_2. \tag{6.3}$$

By inserting (6.2) and (6.3) into (6.1), we get

$$\mathrm{P}(X_G = i \mid Y = 1) = \frac{f_i g_i}{k} \tag{6.4}$$

for the cases, and

$$\mathrm{P}(X_G = i \mid Y = 0) = \frac{(1 - f_i) g_i}{1 - k} \tag{6.5}$$

for the controls. To simulate a dataset, draw the genotype for each individual by using (6.4) for the cases and (6.5) for the controls.

6.2 Method to simulate case-control data including a genotype covariate and an environmental covariate

When a continuous, environmental effect also is included in the model, the simulation procedure becomes slightly more complicated. We will here present an idea of how it can be performed. We assume the genotype and the environmental effect to be independent. As input parameters we need to know the genotype frequencies

and the distribution of the environmental effect. Also, the effect sizes β_0, β_E and β_G need to be predetermined.

Let X_G denote the genotype variable with $\mathrm{P}(X_G = i) = g_i$ for $i = 0$, 1 and 2. Moreover, let X_E denote the environmental variable with density function f, and let the random variable Y be defined such that

$$Y = \begin{cases} 0 \text{ if control} \\ 1 \text{ if case} \end{cases}.$$

Also, we denote

$$h(x, i) = \mathrm{P}(Y = 1 \mid X_G = i, X_E = x) = \frac{\exp(\beta_0 + \beta_E x + \beta_G i)}{1 + \exp(\beta_0 + \beta_E x + \beta_G i)}.$$

To be able to simulate a value for the environmental covariate and for the genotype, we need expressions for the conditional density of X_E given $Y = y$, which we denote f_Y, for $y = 0$, 1, and $\mathrm{P}(X_G = i \mid Y = y)$ for $i = 0$, 1, 2, and $y = 0$, 1. Since we have both a discrete random variable and a continuous random variable, we can obtain the mixed joint densities given by

$$h(x, i) g_i f(x)$$

when $y = 1$, and

$$(1 - h(x, i)) g_i f(x)$$

when $y = 0$. We can now find an expression for

$$\mathrm{P}(X_G = i \mid Y = y) = \frac{\mathrm{P}(X_G = i, Y = y)}{\mathrm{P}(Y = y)} \tag{6.6}$$

by summing over and integrating out the variables we do not want to include. For the denominator, we get

$$k = \mathrm{P}(Y = 1) = \int_{-\infty}^{\infty} f(x) \left(\sum_{i=0}^{2} g_i h(x, i) \right) dx$$

when $y = 1$, and

$$\mathrm{P}(Y = 0) = 1 - k$$

when $y = 0$. We can then rewrite (6.6) to

$$\mathrm{P}(X_G = i \mid Y = 1) = \frac{1}{k} \int_{-\infty}^{\infty} g_i h(x, i) f(x) dx, \quad i = 0, 1, 2, \tag{6.7}$$

when $y = 1$, and

$$\mathrm{P}(X_G = i \mid Y = 0) = \frac{1}{1-k} \int_{-\infty}^{\infty} g_i (1 - h(x, i)) f(x) dx, \quad i = 0, 1, 2, \tag{6.8}$$

when $y = 0$. Now, (6.7) and (6.8) can be used to simulate genotype for the cases and the controls, respectively. For the environmental covariate, we get

$$f_1(x) = \frac{1}{k} \sum_{i=0}^{2} g_i h(x, i) f(x) \tag{6.9}$$

when $y = 1$, and

$$f_0(x) = \frac{1}{1-k} \sum_{i=0}^{2} g_i (1 - h(x, i)) f(x) \tag{6.10}$$

when $y = 0$.

Generally, k can not be given in a closed form, so we need to use numerical integration. When an expression for k is found, we can obtain the three possible probabilities from (6.7) and the three possible probabilities from (6.8), and furthermore simulate genotypes by using these. To simulate environmental covariates, we solve (6.9) and (6.10), and draw values from these probability distributions. Rejection sampling is one possible technique that may be used for this purpose.

6.3 Comments

The method we used to simulate data for our power study, provides data in a case-control setting, but is has some drawbacks. For small MAF values, we experience that a large number of simulations is needed in order to obtain the number of cases we want. Hence, there is a relationship between the MAF value and the number of simulations needed, which again may affect the computational time. The advantage of this method is that the expressions for the covariates included may be given in a closed form. This makes the simulation procedure less complex.

As mentioned, the simulation method outlined by Zheng et al. (2012) (pp. 85-86) can not be applied in general. It is only valid for situations where we know the

relationship between β, μ and σ for the environmental effect, like we do in Section 2.5. If this relationship is not known, the given procedure is not applicable.

The simulation methods presented in Sections 6.1 and 6.2 are more general methods for case-control data simulation than the one provided by Zheng et al. (2012) (pp. 85-86). Hence, they have a greater area of application, but also some drawbacks. Since we assume independency between the genotype effect and the environmental effect, the method in Section 6.2 can not be used if we want to include an interaction effect between the two variables. For such situations, the cohort procedure has to be followed. Also, some uncertainty may be expected when performing numerical integration and rejection sampling.

Chapter 7

Discussion and conclusion

We have in this thesis presented, tested and compared statistical methods that can be used to detect association between a genetic marker and a common disease. We have seen how assuming an environmental effect, and later also an interaction effect, influence the performance of the methods. The effects of the different parameter values have already been discussed in Chapter 5. In the present chapter, we will take a closer look at the hypotheses tested during this thesis, and discuss possible approaches when the aim is to detect genotype-phenotype association.

7.1 Different approaches

We have during this thesis focused on three different statistical hypotheses. These hypotheses are summed up and enumerated in Table 7.1.

No	Hypothesis	
0	H_0:	$\beta_G = 0$
	H_1:	$\beta_G \neq 0$
1	H_0:	$\beta_{EG} = 0$
	H_1:	$\beta_{EG} \neq 0$
2	H_0:	$\beta_G = \beta_{EG} = 0$
	H_1:	at least one of β_G or β_{EG} is not equal to zero

Table 7.1: List of the statistical hypotheses considered in this thesis.

When analyzing biological datasets, we often do not know whether an interaction effect is present or not. The question of interest is often as simple as; "Is there

any association between the genotype and the disease?". There are several possible approaches to find an answer to this question. The classical statistical method is to first test hypothesis 1 from Table 7.1. If this hypothesis is rejected, the answer to the question is yes, and no more testing needs to be done. If hypothesis 1 is not rejected, then we can continue with testing hypothesis 0 from Table 7.1. If the result of this test leads to rejection of the null hypothesis, then the conclusion is that the genotype does have an influence on the probability of developing the given disease. If the null hypothesis is not rejected, then we do not have evidence to say that there is any association between the genotype and the disease.

A competing approach is to only test the composite hypothesis 2 from Table 7.1. This will probably require a smaller number of tests to be performed, which is preferable. When comparing these two approaches, the overall power is what is of interest. From the results in Chapter 5, we know that testing hypothesis 2 provides greater power than testing hypothesis 1. For hypothesis 1, we also have to consider that an additional testing of hypothesis 0 is necessary if the null hypothesis is not rejected. To be able to compare the two approaches outlined here, it all comes down to which level of significance do we have to use for hypothesis 1 and 0 to obtain an overall level of significance equal to the level α chosen for hypothesis 2. When such values are set, we can compare the statistical power for the two approaches, and based on this conclude which procedure is preferable. A possible choice of significance level for the first procedure, is to use $\frac{\alpha}{2}$ for hypothesis 1 and 0, and α for hypothesis 2. This is a rather conservative correction. To obtain less strict significance levels for the hypotheses 1 and 0, the correlation between the two tests has to be measured and taken into account.

As in many other contexts, previous knowledge and experience can be useful when deciding which hypothesis (or hypotheses) one should focus on testing. If we want to test for an interaction effect which is not present, using the first approach will, in worst case, lead to twice as many hypothesis tests than the second approach. If we have some previous knowledge, and do not expect an interaction effect to be present at all, testing only hypothesis 0 should be considered. Also, if we follow the classical statistical procedure, and do not reject hypothesis 1, we have two options when testing hypothesis 0. We can either use the same model, or refit and get a simpler model. This may also influence which procedure is preferable.

From this we can conclude that it is difficult to give an overall and general recommendation of which hypothesis approach one should use when testing for associations in biological datasets. More analyses on this topic is an interesting idea for further work.

7.2 Conclusion

We have in this thesis presented the statistical methods Score test, Likelihood ratio test, Wald test and Cochran-Armitage test for trend. By using simulated datasets,

we have compared the performance of these methods when the goal is to detect association between a genetic marker and a common disease, with environmental effects present as well. Based on our observations, logistic regression models are appropriate for detecting genotype-phenotype association. We recommend to use either Score test, LRT or Wald test when the environmental effect is large (odds ratio greater than 5 if the variable is standard normal distributed). Among these, the Score test has a computational advantage which makes it preferable. If the environmental effect has an odds ratio less than or equal to 5, the CATT provides greater power values than the other methods. Hence, CATT is recommended for such situations. This is an interesting finding which is useful for future studies. The recommendations given here, apply both when the number of cases and controls in the study is balanced, and when it is unbalanced.

We have also provided insight into the process of generating data for cohort and case-control studies. The method outlined in Section 6.2 gives case-control data, but if it is of interest to include an interaction effect, this method needs some adjustments.

As discussed in the previous section, when an interaction effect between the genotype and the environmental influence may be present, we can either use the classical statistical- or the alternative approach to detect genotype-phenotype association. More analyses and calculations are needed in order to determine which one is preferable in certain situations. Also, there may be other hypotheses that one should consider testing under such conditions.

Bibliography

A. Agresti (1996). *An Introduction to Categorical Data Analysis*. John Wiley & Sons.

G. Castella & R. L. Berger (2002). *Statistical inference*. Brooks/Cole.

A. K. Daly & C. P. Day (2001). 'Candidate Gene Case-Control Association Studies: Advantages and Potential Pitfalls'. *British Journal of Clinical Pharmacology* **52(5)**:489–499.

A. J. Dobson & A. G. Barnett (2008). *An introduction to Generalized Models*. Taylor & Francis Group.

J. Fox (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Sage Publications.

M. Gabrielsen (2013). *Genetic Risk Factors for Lung Cancer: Relationships to Smoking Habits and Nicotine Addiction*. Ph.D. thesis, Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology.

Y. Ge, et al. (2003). 'Resampling-based Multiple Testing for Microarray Data Analysis'. *Sociedad de Estadistica e Investigacion Operativa Test* **12**:1–77.

W. Guan, et al. (2012). 'Identifying Plausible Genetic Models Based on Association and Linkage Results: Application to Type 2 Diabetes'. *Genetic Epidemiology* **36**:820–828.

K. K. Halle (2012). 'Statistical Methods for Multiple Testing in Genome-Wide Association Studies'. Master's thesis, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Human Genome Project Information (2013). 'SNP Fact Sheet'. `http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml`.

R. A. Johnson & D. W. Wichern (2007). *Applied Multivariate Statistical Analysis*. Pearson Education.

M. Langaas & Ø. Bakke (2013). 'Increasing power with the unconditional maximization enumeration test in small samples − a detailed study of the MAX3 test statistic'. Tech. rep., Department of Mathematical Sciences, Norwegian University of Science and Technology. Statistics Preprint, 1/2013.

P. McCullagh & J. Nelder (1989). *Generalized Linear Models, 2nd edition.* Chapman & Hall.

National Human Genome Research Institute (2013). 'Genome-Wide Association Studies'. `http://www.genome.gov/20019523`.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

K. J. Rothman, et al. (2008). *Modern Epidemiology.* Lippincott Williams & Wilkins.

N. J. Schork, et al. (2009). 'Common vs. rare allele hypotheses for complex diseases'. *Current Opinion in Genetics & Development* **19**:212–219.

G. K. Smyth (2003). 'Pearson's Goodness of Fit Statistic as a Score Test Statistic'. *Science and Statistics: A Festschrift for Terry Speed* **40**:115–126.

TOP (2013). 'TOP'. `http://www.med.uio.no/klinmed/forskning/sentre/kgj-psykoseforskning/`.

X. Wang, et al. (2010). 'The Meaning of Interaction'. *Hum Hered* **70**:269–277.

G. Zheng, et al. (2006). 'Robust Genomic Control for Association Studies'. *The American Journal of Human Genetics* **78(2)**:350–356.

G. Zheng, et al. (2012). *Analysis of Genetic Association Studies.* Springer Science+Business Media.

A. Ziegler & I. R. König (2010). *A Statistical Approach to Genetic Epidemiology.* WILEY-VCH Verlag GmbH & Co. KGaA.

Appendix A

R code

A.1 Data simulation

R code for simulating datasets consisting of `n0` controls and `n1` cases. The input variables are, in addition to `n0` and `n1`, odds ratio value for the environmental (`OR_E`) and the genotype (`OR_G`) effect, minor allele frequency (`MAF`), and mean (`muE`) and standard deviation (`sigmaE`) for the environmental variable. The function returns a matrix with the columns `y`, `xE` and `xG`.

```
simulate <- function(n0,n1,OR_E,OR_G,MAF,muE,sigmaE){
  g0<-(1-MAF)^2
  g1<-2*MAF*(1-MAF)
  g2<-MAF^2

  beta0<-(-3.5)
  betaE<-log(OR_E)
  betaG<-log(OR_G)

  i<-1
  countContr<-0
  countCase<-0
  y<-NULL
  xE<-NULL
  xG<-NULL
  while(countContr<n0 || countCase<n1){
    xGTemp<-sample(c(0,0.5,1),1,prob=c(g0,g1,g2))
    xEtemp<-rnorm(1,muE,sigmaE)
    yTemp<-rbinom(1,1,
        prob=exp(beta0+betaE*xEtemp+betaG*xGTemp)/
```

```
        (1+exp(beta0+betaE*xEtemp+betaG*xGTemp)))

    if(yTemp==0 && countContr<n0){
      y[i]<-yTemp
      xG[i]<-xGTemp
      xE[i]<-xEtemp
      countContr<-countContr+1
      i<-i+1
    }
    if(yTemp==1 && countCase<n1){
      y[i]<-yTemp
      xG[i]<-xGTemp
      xE[i]<-xEtemp
      countCase<-countCase+1
      i<-i+1
    }
  }
  return(cbind(y,xE,xG))
}
```

A.2 Hypothesis testing without assuming an interaction effect

Cochran-Armitage test for trend

Performing the CATT by using the function `CATT()` from the package *Rassoc*. The obtained p-values are stored in the matrix `pvaluesCATT`.

```
contmat<-matrix(c(sum(y==0&xG==0),sum(y==0&xG==0.5),sum(y==0&xG==1),
  sum(y==1&xG==0),sum(y==1&xG==0.5),sum(y==1&xG==1)),ncol=3,byrow=T)
pvaluesCATT[i]<-pchisq(CATT(contmat)^2,1,lower.tail=F)
```

Model fitting

Fitting the models needed for the Score test, the LRT and the Wald test.

```
fit<-glm(y~xE,family=binomial)
fit2<-glm(y~xE+xG,family=binomial)
```

Score test

Performing the Score test by using the function `glm.scoretest()` from the package *statmod*. The obtained p-values are stored in the matrix `pvaluesScore`.

```
pvaluesScore[i]<-pchisq(glm.scoretest(fit,xG)^2,1,lower.tail=F)
```

Likelihood ratio test

Performing the LRT. The obtained p-values are stored in the matrix `pvaluesLRT`.

```
pvaluesLRT[i]<-pchisq(anova(fit2)$Deviance[3],1,lower.tail=F)
```

Wald test

Performing the Wald test. The obtained p-values are stored in the matrix `pvaluesWald`.

```
pvaluesWald[i]<-summary(fit2)$coef[3,4]
```

A.3 Hypothesis testing when assuming an interaction effect

Model fitting

Fitting the models needed for the Score test, the LRT and the Wald test.

```
fit<-glm(y~xE,family=binomial)
fit2<-glm(y~xE+xG,family=binomial)
fit3<-glm(y~xE*xG,family=binomial)
```

Score test

Performing the Score test by using the function `glm.scoretest()` from the package *statmod* for hypothesis 1 in Table 5.11. The obtained p-values are stored in the matrix `pvaluesScore1`.

```
pvaluesScore1[i]<-pchisq(glm.scoretest(fit2,xE*xG)^2,1,lower.tail=F)
```

Performing the Score test for hypothesis 2 in Table 5.11. The obtained p-values are stored in the matrix `pvaluesScore2`. It is possible to implement the Score test without fitting the model `fit3`, but for the convenience, the following method is used.

```
pvaluesScore2[i]<-anova(fit,fit3,test="Rao")$"Pr(>Chi)"[2]
```

Likelihood ratio test

Performing the LRT for hypothesis 1 in Table 5.11. The obtained p-values are stored in the matrix `pvaluesLRT1`.

```
pvaluesLRT1[i]<-pchisq(anova(fit3)$Deviance[4],1,lower.tail=F)
```

Performing the LRT for hypothesis 2 in Table 5.11. The obtained p-values are stored in the matrix `pvaluesLRT2`.

```
pvaluesLRT2[i]<-pchisq(anova(fit,fit3)$Deviance[2],2,lower.tail=F)
```

Wald test

Performing the Wald test for hypothesis 1 in Table 5.11. The obtained p-values are stored in the matrix `pvaluesWald1`.

```
pvaluesWald1[i]<-summary(fit3)$coef[4,4]
```