Logistic Regression with Missing Haplotypes

by

Kelly Burkett

B.Sc., University of Guelph, 2000

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY
December 2002

APPROVAL

Name: Kelly Burkett

Degree: Master of Science

Title of project: Logistic Regression with Missing Haplotypes

Examining Committee: Dr. Richard Lockhart
Chair

Dr. Jinko Graham
Senior Supervisor
Simon Fraser University

Dr. Brad McNeney
Simon Fraser University

Dr. John Spinelli
External Examiner
BC Cancer Agency and Simon Fraser University

Date Approved: _____

Abstract

Complex diseases are thought to be caused by both environmental and genetic factors. Single nucleotide polymorphisms (SNPs) are currently being explored for use as genetic markers in association studies of complex diseases. SNPs are abundant in the genome, leading to much finer scans of candidate regions. Since multiple SNPs in a region are likely in linkage disequilibrium, it has been suggested that methods which use the information at several SNPs at a time, along the haplotype, will be better for finding disease-predisposing genes through association studies. For certain genotypes, it is not possible to determine the haplotype phase so missing data methods have been used to infer the haplotype or the haplotype frequencies in a sample. The inferred haplotypes or frequencies are then used in association analyses. A chi square test, either on the inferred haplotype counts or the estimated haplotype frequencies, is commonly used but does not account for environmental factors. Currently, many association analyses which adjust for environmental cofactors use imputed haplotypes but do not account for this uncertainty. In this project, the expectation maximization (EM) algorithm is used to handle the missing haplotype information in a logistic regression with haplotypes as the genetic covariates and other fully observed environmental covariates. The haplotypes are not imputed, instead the EM algorithm is used to obtain maximum likelihood estimates of the regression coefficients directly. The variance-covariance matrix is derived using Louis' formula for the observed information (1982) and the method is applied to a simulated cohort study dataset. This method, which allows for both genetic and environmental factors and does not assume that the haplotypes are fully observed, is shown to find the genetic signal better than an analysis which does not include environmental factors.

Acknowledgments

I would like to thank my supervisory committee and in particular my supervisor, Dr. Jinko Graham, for all her help and support throughout my 2 years at SFU. Also, thank you to my friends and fellow students in K9501, with a special thanks to Crystal and Michael for going first, Simon for of all his helpful comments and Jason L for the coffee breaks.

Contents

Approval Page ii
Abstract iii
Acknowledgments iv
List of Tables vi
List of Figures vii
1 Introduction 1
2 Background 5
2.1 Genetic background and association mapping 5
2.2 An EM algorithm to estimate haplotype frequencies 9
2.3 Improving association methods– an EM-based logistic regression 12
3 Logistic regression with missing haplotypes 13
3.1 The genetic data 13
3.2 EM algorithm by the method of weights 15
3.2.1 The Expectation step 16
3.2.2 The Maximization step 19
3.3 Parameterization of the covariate distribution 23
3.4 Standard errors for regression parameters 25
4 Analysis of a simulated data set 33
5 Summary and conclusions 42
Appendices
A The coalescent and the simulated data 47
Bibliography 49

List of Tables

A.1 Subset of the simulated data set. 48

List of Figures

4.1 Genetic signal for haplotype effect versus chromosome position; model with environmental covariates. 37
4.2 (a)Genetic signal for haplotype effect versus chromosome position; model without environmental covariates. (b) The analysis with the environmental covariates and without superimposed. 39
4.3 (a)Genetic signal for locus effect versus chromosome position. (b) Haplotype model and single-locus model results superimposed. 41

Chapter 1

Introduction

There is much interest in finding genetic susceptibility alleles for common "complex" traits, such as cancer, Type II diabetes, asthma and heart disease, which represent a significant health care burden. Complex traits are multi-factorial, with environmental and non-genetic factors such as weight and age, as well as genetic factors influencing disease susceptibility. To differentiate between genetic factors (*i.e.* genotypes or haplotypes for marker loci or candidate genes) on one hand, and environmental and other non-genetic factors on the other, we will refer to any non-genetic factor as an environmental factor throughout this project. With both environmental and genetic factors contributing to disease, the relationship between an observed phenotype and an underlying disease-predisposing genotype is not as clear as with simple Mendelian disorders. Often candidate regions for disease-predisposing genes are suggested by linkage studies or by biological considerations. These candidate regions can be relatively large because linkage studies do not have sufficient power to find the exact location of the disease-predisposing gene, and because many candidate genes potentially involved in a disease pathway may be clustered together over a large region in the genome. Therefore, linkage-disequilibrium (LD) mapping has become a popular tool for fine-mapping susceptibility loci in candidate regions.

In the presence of linkage disequilibrium, a disease mutation will be associated with the alleles at close surrounding markers. Therefore, differences in allele frequencies at genetic markers between affecteds and unaffecteds can be explained by the

marker being the actual disease-predisposing gene or being in linkage disequilibrium with the disease-predisposing gene. A χ^2 test is often used to test the frequency differences between the two groups. Traditionally, environmental factors have not been taken into account in LD mapping because gene mapping has been focussed on simple Mendelian traits rather than complex traits. However, with the focus now on mapping the more common complex traits, methods that can accommodate both genetic and environmental cofactors are required.

Single nucleotide polymorphisms (SNPs) are diallelic markers that are currently used in many genetic studies. They are abundant in the genome so it is thought that a fine map of SNPs close enough to detect linkage disequilibrium can be constructed. However, since they are diallelic, they have low heterozygosity, meaning that the informativity of a diallelic locus will be lower than for multi-allelic markers. The use of SNP haplotypes rather than single-loci as markers has been suggested to increase heterozygosity. Haplotypes are the particular combinations of alleles that are inherited from a parent. They are not observed for all single-locus genotypes, so to include haplotypes in the analysis, a missing data approach is required.

To compare the counts of haplotypes among affected and unaffected individuals, existing methods like the EM algorithm (Long *et al.* 1995; Hawley and Kidd 1995; Excoffier and Slatkin 1995) or Bayesian approaches (Stephens *et al.* 2001; Niu *et al.* 2002) are first used to estimate haplotypes using the data on single-locus genotypes. Commonly, the expected cell counts of haplotypes in affected and unaffected individuals are first found using the estimated frequencies and a χ^2 test is then applied to these expected counts to determine if they differ in the affected and unaffected groups (see for example Fallin *et al.* 2001). To include haplotype information and environmental factors in a more sophisticated analysis, a two-stage approach has also been adopted, which first involves imputing the missing haplotypes, then treating them as known in further analysis, such as in a logistic regression. In an analysis that uses the imputed values as the true values, the variance will not accurately reflect the added uncertainty associated with the haplotype estimation procedure. In addition, only the genetic data is used to impute haplotype values, even though data on disease outcomes and environmental covariates is available for each individual.

In this project, we describe a maximum likelihood approach that uses logistic regression to estimate and test associations of a complex disease with particular haplotypes, taking into account environmental risk factors and uncertain haplotypes. With this method the missing haplotype covariates are not imputed in a first stage of analysis. Instead, haplotype frequencies and regression coefficients are estimated jointly using the EM algorithm by method of weights (Ibrahim 1990). In contrast to the haplotype reconstruction method described above, all information known for an individual, not just the single-locus genotypes, is used to estimate the haplotype parameters. Rather than leave the genetic covariate distributions unspecified, we make the assumptions of Hardy-Weinberg equilibrium and independence of the genetic and environmental covariates. These assumptions are standard in genetic data analysis and, if valid, improve statistical efficiency and stability of estimation. However, the method should be applied with caution when there is reason to suspect that the candidate region contains genes which influence the environmental exposures. The method is expected to be robust to departures from Hardy-Weinberg equilibrium, as are its counterparts for haplotype reconstruction based on single-locus genotype data (Fallin and Schork 2000). We derive the variance of the estimated regression coefficients using Louis' formula (1982). With the logistic regression model, not only can environmental covariates be included, but more detailed modeling of gene-gene and gene-environment interactions can also be done.

Chapter 2 contains background information. Some genetic terminology is given, as well as a brief description of gene mapping by association methods. Other approaches for handling the missing haplotypes are described, with an emphasis placed on the maximum likelihood estimation of haplotype frequencies using the EM algorithm, since it is currently a popular approach for dealing with ambiguous haplotypes, and therefore a natural starting point for extension.

Chapter 3 describes the EM algorithm by the method of weights, as applied to a logistic model with logit link. The parameterization of the covariate distribution that we chose, which involves assuming Hardy-Weinberg equilibrium and independence of genetic and environmental covariates, is also described. Finally the standard errors of the regression coefficients are derived.

In chapter 4, the method is applied to a simulated candidate-region scan of a cohort of diseased and unaffected individuals. The analysis including environmental covariates is compared to the analysis which does not adjust for environmental risk factors to demonstrate the importance of including environmental risk factors in association studies. The haplotype method is also compared to an association analysis of single markers to determine its ability to pick up genetic signal.

Finally, chapter 5 contains some concluding points about this approach and possible extensions of the method.

Chapter 2

Background

In order to understand the terminology and the current approaches used, this chapter provides a brief explanation of some genetic concepts and techniques that are used to map disease genes.

2.1 Genetic background and association mapping

At any given locus, each person receives one copy of a gene (called the allele) from their mother and one from their father. The two alleles are transmitted to the offspring independently of each other. However, when considering the transmission of two genes, the alleles at the two loci may not be independent of each other. If the two loci of interest are on different chromosomes, they are unlinked and the transmission of the genes is independent. If the two loci are on the same chromosome, they may be linked so that the genes tend to be transmitted together. During meiosis, homologous chromosomes may undergo the process of recombination. A recombination event or a crossover is said to occur if the DNA at one locus is of a different parental origin than the DNA at the second locus. That is, if on one of the chromosomes transmitted, one locus is of maternal origin and the other is of paternal origin, a crossover of DNA must have occurred somewhere between the two loci. The association in the population between the alleles at the two linked loci depends in part on the probability of a recombination event occurring between the two loci. The closer two loci are to

each other on a chromosome, the smaller the probability of a recombination occurring between them. Thus the values of the alleles at the two loci are not independent but are also not necessarily always of the same parental origin. The two loci may be so far apart on a chromosome that through recombination events the original association is expected to break down after transmission to the next generation; thus they are unlinked as well.

A haplotype is the particular set of alleles, at more than one locus, that are transmitted from one parent. Each person inherits two haplotypes, one from their mother and one from their father. One knows the multilocus genotype or the phase if the set of two haplotypes transmitted is known. In most cases, it is known which alleles the individual received at each locus (called just the single-locus genotypes), but not necessarily the parental haplotypes. For example, a person who has alleles 'A' and 'a' at one locus, and alleles 'B' and 'b' at a second locus (that is, the observed single-locus genotypes are Aa ; Bb), could have inherited haplotypes AB and ab (they are AB/ab), or haplotypes Ab and aB (they are Ab/aB). The number of possible haplotypes is the product of the number of alleles at each of the loci being considered.

An initial disease mutation occurs on a particular background haplotype. At first, the mutation allele is completely associated with the alleles at the surrounding loci. With time, recombination breaks down this association. However, loci that are tightly linked to the disease locus take many meioses for this association to degrade, since few recombination events will occur between the two loci. Thus, there is allelic association or linkage disequilibrium between the marker and disease mutation locus. For example, suppose that a mutation m occurs on a haplotype with alleles a_1 and b_1 at tightly linked loci A and B. The initial mutation-bearing haplotype is a_1mb_1 and only individuals with a_1 and b_1 at the adjacent loci will have the mutation allele. If A is assumed to be the closer of the two loci, after many generations, mutation-bearing haplotypes could be, for example, a_1mb_2 and a_1mb_1 in proportion to the frequencies of b_1 and b_2 in the population. The a_1 allele is still associated with the mutation allele but the b_1 is no longer associated with the mutation allele.

Gene mapping by association methods exploits the linkage disequilibrium between a disease locus and a marker that is tightly linked with the disease locus.

The methods used for non-family based samples examine marker alleles in affected individuals and unaffected individuals to determine if there are differences in allele frequencies between the two groups. A significant difference between an allele frequency in the two groups is taken as evidence that the allele either predisposes towards the disease or is in linkage disequilibrium with the disease-predisposing allele. A χ^2 test is often used to test whether the allele frequencies differ between the two groups.

Complex traits, as opposed to Mendelian diseases, do not follow typical segregation patterns with any single locus. There could be numerous reasons for the unknown underlying inheritance of complex traits. A complex trait is likely affected by both genetic and environmental factors. Given a certain genetic background, a person may be susceptible to disease but might not ever develop the disease (incomplete penetrance). Often with complex diseases the phenotype may appear the same in many individuals, but could have different underlying non-genetic causes (phenocopy). There may be multiple genes which cause the disease (locus heterogeneity), or multiple mutations within a disease gene which lead to the same disease (allelic heterogeneity). Finally with complex diseases, there are likely multiple genes that additively affect the disease outcome (polygenic inheritance) or interact to affect the outcome (epistasis).

Since complex traits, such as heart disease and asthma, are often common diseases, researchers are interested in finding their underlying genetic mechanisms. For these types of diseases, it is common to test for association of the disease with a genotype at candidate loci using data from a cohort or case-control study. However, traditional association methods for genetic analyses, like the χ^2 test given above, generally only examine the genetic effect of one locus and often do not correct for known environmental risk factors. These approaches have served well for simple Mendelian diseases, which are not influenced by environmental risk factors. However, for complex traits, methods that can incorporate environmental information are likely to be more successful.

Single nucleotide polymorphisms (SNPs) are currently being examined for use as genetic markers in association studies of complex diseases. They are abundant throughout the genome so it is possible to choose closely linked SNPs for association

studies of a candidate region. Due to their diallelic nature, they may also be easier to type using automated techniques. However, a drawback to using SNPs for association studies is the decreased informativity associated with markers having only two alleles. Association between a diallelic marker and a disease locus may be difficult to find unless the marker is closely linked to the disease locus. Therefore, the use of haplotypes rather than single loci in association studies can increase heterozygosity (information content) at a marker locus and better capture linkage disequilibrium in a region. Zöllner and Von Haeseler (2000) and Akey *et al.* (2001) showed that the power of χ^2 tests to detect genetic association is improved using haplotypes. However, using simulated data, Kaplan and Morris (2001) found that even with the haplotype phase known, there was rarely an advantage to using multiple locus haplotypes to detect association with a single disease-predisposing allele. More recently, Morris and Kaplan (2002) have studied the power of a likelihood ratio test to detect association in a case-control design in the presence of multiple susceptibility alleles, using both single-locus genotypes and both known and missing haplotype information. Their results show that the haplotype analysis is more powerful when there are multiple susceptibility alleles.

It is not always possible to determine haplotype phase from the single-locus genotypes. If a person is heterozygous at $h > 0$ loci, then the number of multilocus genotypes that are consistent with the observed genotype is 2^{h-1}. For example, if $h = 1$, and there are three loci A, B and C, with the third one heterozygous, the multilocus genotype is $a_1b_1c_1/a_1b_1c_2$. If $h = 2$ and loci B and C are heterozygous, the 2 possible multilocus genotypes are $a_1b_1c_1/a_1b_2c_2$ and $a_1b_1c_2/a_1b_2c_1$. Phase can only be determined using technologically demanding and cost prohibitive laboratory techniques (Judson and Stephens 2001) or by collecting genotype data on family members. For example, if parental genotypes are known, the multilocus genotype may be inferred by determining which alleles were inherited from each parent. In many cases however, genotyping of more members of a family would have to be done to infer the haplotypes with certainty. If the disease being studied is a late onset disease, it may not be possible to get the genotypes of the parents or other family members. If genotyping is not done on any family members, it is unlikely that the

haplotype phase can be inferred for a whole sample, which is a drawback to the use of haplotypes in association studies.

Since it is often infeasible to accurately determine the multilocus genotypes, several methods have been proposed for estimating the haplotype frequencies or imputing the unknown haplotypes. Clark (1990) imputes the unknown multilocus genotype with unambiguous haplotypes in the sample or with combinations of haplotypes already seen, so a haplotype not yet seen can be explained with recombination events. The expectation maximization (EM) algorithm (Dempster *et al.* 1977) has been used to estimate haplotype frequencies (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long *et al.* 1995). The implementation described by Long *et al.* is given below in more detail. Recently, Bayesian approaches have been suggested. Stephens *et al.* (2001) describe a method (SSD) which imputes haplotype phase by modeling the prior distribution of haplotypes using population genetic theory. The SSD method has been implemented in the program PHASE. Niu *et al.* (2002) also describe a Bayesian approach to haplotype estimation that only assumes Hardy-Weinberg equilibrium.

2.2 An EM algorithm to estimate haplotype frequencies

The implementation of the EM algorithm for estimating haplotype frequencies described here is from Long *et al.* (1995) and is used to estimate haplotype frequencies from data randomly sampled from a population. The method finds the maximum likelihood estimates of the frequencies of each haplotype by assuming Hardy-Weinberg equilibrium of the haplotypes. Hardy-Weinberg equilibrium (HWE) is achieved through the random union of alleles, or in this case haplotypes, in the population. If p and q are the frequencies of the two possible alleles A and a, then after one generation of random-mating in a large population, the genotype frequencies are p^2, q^2 and $2pq$ (for AA, aa and Aa respectively). Equilibrium is achieved since the allele frequencies calculated from these genotype frequencies are again p and q.

The log-likelihood over the n independent individuals is given by

$$\ln L = \sum_{i=1}^{n} \ln p_i,$$

where p_i is the probability of the i^{th} person's observed genotype and is

$$p_i = \sum \Pr\{h_k/h_l\}$$

where the sum of the probabilities is over all possible multilocus genotypes consistent with the observed single-locus genotypes and h_k/h_l is the multilocus genotype made up of haplotypes l and k. If γ_j is the probability of haplotype j, then given HWE,

$$\Pr\{h_l/h_k\} = \begin{cases} \gamma_l^2 & \text{if } k = l \\ 2\gamma_l\gamma_k & \text{otherwise} \end{cases}.$$

If N_{l_i} is the number of h_l haplotypes within individual i, then the EM algorithm at the t^{th} iteration can be written as:

1. E step- Impute the expected numbers of each haplotype.

$$n_{l_i} = E[N_{l_i}|\text{genotypes of individual } i] = \frac{2\gamma_l^{(t)}\gamma_k^{(t)}}{p_i^{(t)}},$$

where multilocus genotype h_l/h_k is consistent with i's observed genotype. Then

$$n_l = E[N_l|\text{data}] = \sum_{i=1}^{n} n_{l_i}.$$

2. M step- Find maximum likelihood estimates of haplotype frequencies using the imputed counts.

$$\gamma_l^{(t+1)} = \frac{n_l}{2n}$$

Note that if the multilocus genotype is known and, for example, homozygous, the E step gives $n_{l_i} = 2\gamma_l^2/\gamma_l^2 = 2$. In the multinomial likelihood, this 2 corresponds to the individual contributing two h_l haplotypes to the total count. If the phase is known and the two haplotypes combined are h_l and h_k, the expected count of haplotype h_l is

$n_{l_i} = 2\gamma_l\gamma_k/(2\gamma_l\gamma_k) = 1$. Individual i contributes one h_l haplotype to the total count of h_l haplotypes.

The algorithm can be generalized for many diallelic loci. As the number of loci or alleles increase, the number of possible haplotypes also increases. The number of frequencies to be estimated can get quite large, depending on the number of loci and alleles. Since more unknown multilocus genotypes means more frequencies that need to be estimated, the estimates will suffer if the heterozygosity is too large, limiting the size of SNP haplotypes that can be considered. Constraining haplotype frequencies to 0 when they appear to be close to 0 can increase the maximum number of loci, however the small size of haplotypes that can be estimated remains a limitation of this approach. The Bayesian approaches of Stephens *et al.* (2001) and Niu *et al.* (2002) can handle many more diallelic loci than the EM methods because they effectively limit the number of haplotypes to be considered.

Using data simulated under varying conditions and assumption violations, Fallin and Schork (2000) studied the accuracy of the EM estimated frequencies compared to the sample frequencies and population frequencies. Because the haplotype phase was simulated, both the generating population haplotype frequencies and sample haplotype frequencies were known. They looked at the effects of sample size, number of loci, heterozygosity, and the presence of rare haplotypes, and found the EM estimates to be reasonably close to the sample frequencies under most circumstances (no more than 5% difference if the sample size was larger than 100). They found that most of the error in estimation was related to the error due to sampling versus error in the EM estimation procedure. They did not address the subsequent error that would be incurred if the EM frequencies were then used to impute haplotypes for further statistical analysis.

2.3 Improving association methods– an EM-based logistic regression

In the present analysis, we are interested in finding an association of a complex disease with a particular haplotype using a statistical model that allows for other, possibly continuous, environmental risk factors to be included. A simple χ^2 test for association of haplotype frequencies with disease is not adequate since it will not allow detailed covariate information about the patient to be included in the analysis.

Logistic regression is a common method for analyzing cohort or case-control data. However, to include haplotype covariates, a missing data method is required. One could take a two-stage approach by using any of the haplotype estimation procedures given above to impute the values in a first stage, and then treat them as known in a second stage of logistic regression analysis. Although the methods given have been shown to be relatively robust to violation of assumptions such as HWE, and the estimates are relatively close to the sample estimates (Fallin and Schork 2000), the effects of using the imputed values in statistical analyses which treat them as known have not been studied. Any analysis that uses the imputed values as the true values risks underestimating the variance, since it will not include the uncertainty associated with the estimation of haplotypes. Finally in reconstructing haplotypes, we would also like to use available information on environmental covariates and disease status.

In the following chapter, an EM-based logistic regression for binary response data from a cohort study is described. With this method, the EM algorithm is not first used to impute the haplotype covariates. Instead, this maximum likelihood approach jointly estimates the haplotype and environmental risk parameters and the haplotype frequencies on the basis of data on affection status, non-genetic covariates and single-locus genotypes. The approach is a natural extension of the maximum likelihood estimation of haplotype frequencies, by use of the EM algorithm, on the basis of data on single-locus genotypes presented in section 2.2.

Chapter 3

Logistic regression with missing haplotypes

Assume that we have single-locus genotypic information on 2 loci for n randomly sampled individuals. In addition, we have complete information on environmental covariates for each of the n individuals. We would like to perform a logistic regression, using haplotypes and environmental factors as covariates. However, the haplotypic information will not be known for all individuals. For this reason, the EM algorithm will be used to estimate the regression parameters. To simplify the description of the approach, we start by considering examples where the risk model involves only genetic covariates and then introduce environmental covariates in section 3.3

3.1 The genetic data

Indicator variables are used to describe which of the multilocus genotypes an individual has. For example, for two-loci with two alleles there are 10 multilocus genotypes. Let a_1 and a_2 be the alleles at locus a, and b_1 and b_2 be the alleles at the second locus b. The variables in a saturated genetic model with an intercept term for the baseline

group of a_2b_2/a_2b_2 homozygotes would be:

$$\begin{aligned}
x_{i0} &= 1 \\
x_{i1} &= \text{I(individual } i \text{ is } a_1b_1/a_1b_1) \\
x_{i2} &= \text{I(individual } i \text{ is } a_1b_1/a_1b_2) \\
x_{i3} &= \text{I(individual } i \text{ is } a_1b_1/a_2b_1) \\
x_{i4} &= \text{I(individual } i \text{ is } a_1b_1/a_2b_2) \\
x_{i5} &= \text{I(individual } i \text{ is } a_1b_2/a_1b_2) \\
x_{i6} &= \text{I(individual } i \text{ is } a_1b_2/a_2b_1) \\
x_{i7} &= \text{I(individual } i \text{ is } a_1b_2/a_2b_2) \\
x_{i8} &= \text{I(individual } i \text{ is } a_2b_1/a_2b_1) \\
x_{i9} &= \text{I(individual } i \text{ is } a_2b_1/a_2b_2)
\end{aligned}$$

A second genetic model may involve the number of copies of a particular haplotype. Allowing an intercept term for the baseline group a_2b_2/a_2b_2 and considering non-a_2b_2 haplotypes, the variables are:

$$x_{ih} = \begin{cases} 0 & \text{individual } i \text{ has 0 copies of haplotype } h \\ 1 & \text{individual } i \text{ has 1 copy of haplotype } h \\ 2 & \text{individual } i \text{ has 2 copies of haplotype } h. \end{cases},$$

for $h = 1, 2, 3$. For example, if each haplotype consists of 2 diallelic loci, a and b, then this haplotype-dose model could be coded as:

$$\begin{aligned}
x_{i0} &= 1 \\
x_{i1} &= \#\text{ copies of haplotype } a_1b_1 \\
x_{i2} &= \#\text{ copies of haplotype } a_1b_2 \\
x_{i3} &= \#\text{ copies of haplotype } a_2b_1
\end{aligned}$$

Let $\mathcal{S}$ be the set of all possible genetic covariate vectors for any individual. An element of $\mathcal{S}$ for the saturated example given above would be $(1, 0, 1, 0, 0, 0, 0, 0, 0, 0)$

for an a_1b_1/a_1b_2 heterozygote. Let $\mathbf{x}^{(j)}$ denote the j^{th} covariate vector in $\mathcal{S}$. Since an individual can only be of one type of multilocus genotype, the number of elements in the set $\mathcal{S}$ is $r =$(# multilocus genotypes). Let the probability that an individual is of covariate type j be γ_j and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_{r-1})$. Thus, the counts of individuals having each of the possible covariate assignments are multinomial$(n, \boldsymbol{\gamma})$.

For real genetic data we cannot observe these covariate values since the haplotype phase is not known. The number that are not observed depends on the number of heterozygote loci. For example, if the multilocus genotype is not ambiguous, then for the saturated genetic model $\mathbf{x} = (1, x_1, x_2, x_3, 0, x_5, 0, x_7, x_8, x_9)$, where the sum of x_1 through x_9 is a maximum of 1. If the multilocus genotype is ambiguous, then $\mathbf{x} = (1, 0, 0, 0, x_4, 0, x_6 = 1 - x_4, 0, 0, 0)$. For example, if the multilocus genotype is a_1b_2/a_2b_2, then $\mathbf{x} = (1, 0, 0, 0, 0, 0, 0, 1, 0, 0)$. For the haplotype-dose model, an individual who is heterozygous at both loci will either be $\mathbf{x} = (1, 1, 0, 0)$ or $\mathbf{x} = (1, 0, 1, 1)$. Let $\mathcal{S}_i$ denote the subset of $\mathcal{S}$ containing only the covariate vectors that are compatible with the observed covariates for subject i. For the saturated genetic model, the heterozygote will have $\mathcal{S}_i = \{(1, 0, 0, 0, 1, 0, 0, 0, 0, 0), (1, 0, 0, 0, 0, 0, 1, 0, 0, 0)\}$ and for the haplotype-dose model the heterozygote will have $\mathcal{S}_i = \{(1, 1, 0, 0), (1, 0, 1, 1)\}$. We may think of these possible covariate vectors as belonging to "pseudo-individuals" whose data are compatible with subject i. In other words, each individual of unknown multilocus genotype has been broken up into pseudo-individuals representing each of the possible haplotype configurations for that individual.

3.2 EM algorithm by the method of weights

The EM algorithm is described using a logistic model but it is applicable for all generalized linear models. Let $\mathbf{x}_{obs,i}$ be the observed covariate values for the i^{th} individual, let $\mathbf{x}_i$ be the complete-data covariate vector for individual i and let $\mathbf{x}^{(j)}$ be a possible value of $\mathbf{x}_i$. If the haplotype information is not ambiguous, then $\mathbf{x}_{obs,i} = \mathbf{x}_i$. Let $\boldsymbol{\theta}^{(\mathbf{t})} = (\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)})$ be the current parameter estimates, where $\boldsymbol{\beta}$ is the regression parameter and $\boldsymbol{\gamma}$ is the genetic covariate parameter.

3.2.1 The Expectation step

The Expectation step involves computing $Q(\ \boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\ )$, the conditional expected log-likelihood of the complete-data $(\mathbf{x}, \mathbf{y})$ given the observed data $(\mathbf{x}_{obs}, \mathbf{y})$ and the current parameter estimates $\boldsymbol{\theta}^{(t)}$:

$$\begin{aligned} Q(\ \boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\ ) &= E\left[l_{y,\mathbf{x}}(\ \boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}) \mid \mathbf{x}_{obs}, \mathbf{y}, \boldsymbol{\theta}^{(t)}\right] \\ &= \sum_{i=1}^{n} E\left[l_{y,\mathbf{x}}(\ \boldsymbol{\theta} \mid \mathbf{x}_i, y_i) \mid \mathbf{x}_{obs,i}, y_i, \boldsymbol{\theta}^{(t)}\right] \\ &= \sum_{i=1}^{n} Q_i(\ \boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\ ) \end{aligned}$$

The E step is not as easy as imputing the missing data values and maximizing the log-likelihood with the imputed values, a method possible for certain EM implementations. Genetic examples of when this is applicable are the "gene counting" algorithm (Ceppelini *et al.* 1955), which is used to estimate genotype frequencies when only the phenotype is known, and the EM algorithm to estimate haplotype frequencies from genotype data that was described in section 2.2. In both of these cases, the distribution of the (complete) counts is multinomial so the complete-data log-likelihood takes on the form

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{k} T_i(\mathbf{X}) \log(\theta_i),$$

where the sum is over the k different classes and $T_i(\mathbf{X})$ is a linear function of the counts in each class. Therefore

$$Q(\ \boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\ ) = \sum_{i}^{k} E(T_i(\mathbf{X}) | \mathbf{X}_{obs}, \boldsymbol{\theta}^{(t)}) \log(\theta_i),$$

and since T_i is a linear function of the counts, the E step reduces to finding the expected complete counts given what is observed. The M step maximizes the complete-data log-likelihood with the expected counts replacing the unknown counts.

The E step for a logistic regression with missing covariates, however, is not as straightforward. The EM strategy used for generalized linear models with missing

covariates has been named the method of weights (Ibrahim 1990), since the E step is a problem of computing a weight of one or less and scaling each pseudo-individual's contribution to the likelihood by that value. The method of weights described here uses the notation from the review by Horton and Laird (1999) and is derived as follows:

$$
\begin{aligned}
Q(\ \boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\ ) &= \sum_{i=1}^{n} E[l_{y,\mathbf{x}}(\ \boldsymbol{\theta} \mid \mathbf{x}_i, y_i) \mid \mathbf{x}_{obs,i}, y_i, \boldsymbol{\theta}^{(t)}] \\
&= \sum_{i=1}^{n} \left[\sum_{j=1}^{|\mathcal{S}|} l_{y,\mathbf{x}}(\ \boldsymbol{\theta} \mid \mathbf{x}^{(j)}, y_i) \mathrm{Pr}\{\mathbf{x}_\mathrm{i} = \mathbf{x}^{(\mathrm{j})} | \mathbf{x}_{\mathrm{obs,i}}, \mathrm{y_i}, \boldsymbol{\theta}^{(\mathrm{t})}\} \right] \\
&= \sum_{i=1}^{n} \left[\sum_{j=1}^{|\mathcal{S}|} l_{y,\mathbf{x}}(\ \boldsymbol{\theta} \mid \mathbf{x}^{(j)}, y_i) w_{ij}(\boldsymbol{\theta}^{(t)}) \right] \\
&= \sum_{i=1}^{n} \left[\sum_{j=1}^{|\mathcal{S}|} \left(l_{y|\mathbf{x}}(\ \boldsymbol{\theta} \mid \mathbf{x}^{(j)}, y_i) + l_{\mathbf{x}}(\ \boldsymbol{\theta} \mid \mathbf{x}^{(j)}) \right) w_{ij}(\boldsymbol{\theta}^{(t)}) \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{|\mathcal{S}|} w_{ij}(\boldsymbol{\theta}^{(t)}) l_{y|\mathbf{x}}(\ \boldsymbol{\theta} \mid \mathbf{x}^{(j)}, y_i) + \sum_{i=1}^{n} \sum_{j=1}^{|\mathcal{S}|} w_{ij}(\boldsymbol{\theta}^{(t)}) l_{\mathbf{x}}(\ \boldsymbol{\theta} \mid \mathbf{x}^{(j)})
\end{aligned}
$$

where $l_{y|\mathbf{x}}$ is the log-likelihood for the regression model and $l_{\mathbf{x}}$ refers to the log-likelihood for the parameters of the covariate model.

The weights are denoted w_{ij} and are equal to $\mathrm{Pr}\{\mathbf{x}_\mathrm{i} = \mathbf{x}^{(\mathrm{j})} | \mathbf{x}_{\mathrm{obs,i}}, \mathrm{y_i}, \boldsymbol{\theta}^{(\mathrm{t})}\}$. If an individual is completely observed, the weight is 1 for the genotype actually observed since there are no other possible compatible multilocus genotypes given the observed covariate information. If the individual has missing haplotype information, most of the weights will be 0 since $\mathbf{x}_{obs,i}$ limits the possible values $\mathbf{x}^{(j)}$ that $\mathbf{x}_i$ can take on. For the case of two diallelic loci, there are a maximum of two $\mathbf{x}^{(j)}$ which will be compatible with $\mathbf{x}_{obs,i}$. Therefore, all the summations over the set $\mathcal{S}$ can be rewritten as summations over the set $\mathcal{S}_i$ since many of the covariate vectors are not compatible with the observed data and therefore have a conditional probability of 0.

The weights are calculated using Bayes rule:

$$
\begin{aligned}
w_{ij}^{(t)} &= \Pr(\mathbf{x}_{\mathrm{i}} = \mathbf{x}^{(\mathrm{j})} |\ \mathbf{x}_{\mathrm{obs,i}}, \mathrm{y}_{\mathrm{i}}, \boldsymbol{\theta}^{(\mathrm{t})}) \\
&= \frac{\Pr(\mathbf{x}_{\mathrm{i}} = \mathbf{x}^{(\mathrm{j})},\ \mathbf{x}_{\mathrm{obs,i}},\ \mathrm{y}_{\mathrm{i}} |\ \boldsymbol{\theta}^{(\mathrm{t})})}{\Pr(\mathrm{y}_{\mathrm{i}},\ \mathbf{x}_{\mathrm{obs,i}} \mid \boldsymbol{\theta}^{(\mathrm{t})})} \\
&= \begin{cases} 0 & \text{if } \mathbf{x}^{(j)} \text{ is not compatible with } \mathbf{x}_{obs,i} \\ \frac{\Pr(\mathbf{x}_{\mathrm{i}}=\mathbf{x}^{(\mathrm{j})},\ \mathrm{y}_{\mathrm{i}} |\ \boldsymbol{\theta}^{(\mathrm{t})})}{\Pr(\mathrm{y}_{\mathrm{i}},\ \mathbf{x}_{\mathrm{obs,i}} \mid \boldsymbol{\theta}^{(\mathrm{t})})} & \text{if } \mathbf{x}^{(j)} \text{ is compatible with } \mathbf{x}_{obs,i}, \end{cases} \\
&= \begin{cases} 0 & \text{if } \mathbf{x}^{(j)} \text{ is not compatible with } \mathbf{x}_{obs,i} \\ \frac{\Pr(\mathrm{y}_{\mathrm{i}}|\mathbf{x}_{\mathrm{i}}=\mathbf{x}^{(\mathrm{j})},\boldsymbol{\theta}^{(\mathrm{t})})\Pr(\mathbf{x}_{\mathrm{i}}=\mathbf{x}^{(\mathrm{j})} |\ \boldsymbol{\theta}^{(\mathrm{t})})}{\sum \Pr(\mathrm{y}_{\mathrm{i}},\mathbf{x}_{\mathrm{i}}=\mathbf{x}^{(\mathrm{k})} \mid \boldsymbol{\theta}^{(\mathrm{t})})} & \text{if } \mathbf{x}^{(j)} \text{ is compatible with } \mathbf{x}_{obs,i}, \end{cases} \\
&= \begin{cases} 0 & \text{if } \mathbf{x}^{(j)} \text{ is not compatible with } \mathbf{x}_{obs,i} \\ \frac{\Pr(\mathrm{y}_{\mathrm{i}}|\mathbf{x}^{(\mathrm{j})},\boldsymbol{\theta}^{(\mathrm{t})})\Pr(\mathbf{x}^{(\mathrm{j})} |\ \boldsymbol{\theta}^{(\mathrm{t})})}{\sum \Pr(\mathrm{y}_{\mathrm{i}}|\mathbf{x}^{(\mathrm{k})},\boldsymbol{\theta}^{(\mathrm{t})})\Pr(\mathbf{x}^{(\mathrm{k})}|\boldsymbol{\theta}^{(\mathrm{t})})} & \text{if } \mathbf{x}^{(j)} \text{ is compatible with } \mathbf{x}_{obs,i}, \end{cases}
\end{aligned}
$$

where the summation in the denominator on the last line is over all $\mathbf{x}^{(k)}$ that are compatible with $\mathbf{x}_{obs,i}$ (*i.e.* over the elements in $\mathcal{S}_i$).

The term $\Pr(\mathbf{x}^{(\mathrm{j})}|\boldsymbol{\theta}^{(\mathrm{t})})$ simplifies to $\Pr(\mathbf{x}^{(\mathrm{j})}|\boldsymbol{\gamma}^{(\mathrm{t})})$ since the distribution of the covariates depends only on the parameters $\boldsymbol{\gamma}^{(t)}$. For example, in a saturated genetic model where an individual's genotype has a multinomial distribution, the probability of the covariate vector taking on value $\mathbf{x}^{(j)}$ at the t^{th} iteration is $\boldsymbol{\gamma}_j^{(t)}$, the probability of that genotype in the population. If the haplotype phase is known, only one covariate vector will be compatible with the observed covariates so the weight will be one. Thus, only the γ_j corresponding to ambiguous multilocus genotypes are required when implementing the algorithm. Making assumptions on the genetic model can simplify the $\boldsymbol{\gamma}$ parameters; this is discussed in section 3.3. In particular, the multilocus genotype frequencies will be written in terms of the haplotype frequencies.

The term $\Pr(\mathrm{y}_{\mathrm{i}}|\mathbf{x}^{(\mathrm{j})}, \boldsymbol{\theta}^{(\mathrm{t})})$ simplifies to $\Pr(\mathrm{y}_{\mathrm{i}}|\mathbf{x}^{(\mathbf{j})}, \beta^{(\mathbf{t})})$, where $\boldsymbol{\beta}$ parameterizes the generalized linear model of disease risk. In our context, this probability is either the probability that $y_i = 1$ or $y_i = 0$ given the covariates and current parameter values. If we assume a logistic model, these values are

$$\Pr(\mathrm{y}_{\mathrm{i}} = 1|\mathbf{x}, \boldsymbol{\beta}^{(\mathrm{t})}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}^{(\mathrm{t})})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta}^{(\mathrm{t})})}$$

and

$$\Pr(\mathrm{y}_{\mathrm{i}} = 0|\mathbf{x}, \boldsymbol{\beta}^{(\mathrm{t})}) = \frac{1}{1 + \exp(\mathbf{x}'\boldsymbol{\beta}^{(\mathrm{t})})}.$$

Two examples are now given to illustrate the calculation of the weights for a saturated genetic model. Individual 1 has single-locus genotypes $a_1a_1;\ b_1b_2$. In this case, the haplotype phase is known: the multilocus genotype is a_1b_1/a_1b_2. Thus, the set $\mathcal{S}_1$ consists of only the element, say the k^{th} covariate type, that matches the multilocus genotype, and

$$w_{1j} = \begin{cases} 0 & j \neq k \\ \frac{\Pr(\mathrm{y_i}|\mathbf{x}^{(\mathrm{j})},\boldsymbol{\theta}^{(\mathrm{t})})\ \gamma_\mathrm{j}}{\Pr(\mathrm{y_i}|\mathbf{x}^{(\mathrm{j})},\boldsymbol{\theta}^{(\mathrm{t})})\ \gamma_\mathrm{j}} = 1 & j = k \end{cases}.$$

Individual 2 has single-locus genotypes $a_1a_2;\ b_1b_2$ so the phase is not known and could be either a_1b_1/a_2b_2 or a_2b_1/a_1b_2. The set $\mathcal{S}_2$ contains two elements, say the k^{th} and l^{th} covariate types and

$$w_{2j} = \begin{cases} 0 & j \neq k, l \\ \frac{\Pr(\mathrm{y_i}|\mathbf{x}^{(\mathrm{j})},\boldsymbol{\theta}^{(\mathrm{t})})\ \gamma_\mathrm{j}}{\Pr(\mathrm{y_i}|\mathbf{x}^{(\mathrm{k})},\boldsymbol{\theta}^{(\mathrm{t})})\ \gamma_\mathrm{k}+\Pr(\mathrm{y_i}|\mathbf{x}^{(\mathrm{l})},\boldsymbol{\theta}^{(\mathrm{t})})\ \gamma_\mathrm{l}} & j = k, l \end{cases}$$

To summarize, the E step involves computing Q, the conditional expected value of the complete-data log-likelihood given the observed data. This function is calculated as the log-likelihood for each pseudo-individual multiplied by a weight corresponding to the possible covariate vector and the observed responses. For individuals with known haplotypes, the weight will be 1 for the multilocus genotype that is made up of the haplotypes, and 0 for all others. For unknown multilocus genotypes, the weights are determined using Bayes rule and are the conditional probabilities of the possible covariate vectors given the observed data.

3.2.2 The Maximization step

The conditional expected values of the complete-data log-likelihood given the observed data and parameter estimates are maximized in the M step to find the new parameter estimates. In the previous section the weights for the likelihood were denoted $w_{ij}^{(t)}$ where i referred to the individual and j referred to one of the haplotype configurations consistent with the individual's multilocus genotype. For simplicity, assume now that we have augmented the data so that for all individuals of unknown multilocus genotype in the sample, a pseudo-individual has been added for each of the possible

multilocus genotypes consistent with the individual's single-locus genotypes. Thus, we can replace the double subscript on the w_{ij}'s with one subscript that goes from 1 to $n + (\#\text{individuals added}) = M$. The weights for each of these pseudo-individuals will now be denoted by a_i to differentiate them from the previous w_{ij}.

The expected complete-data log-likelihood $Q(\ \boldsymbol{\theta} \mid \boldsymbol{\theta}^t\ )$ consists of two portions, one for $l_{y|\mathbf{x}}$ and one for $l_{\mathbf{x}}$ with summands for each pseudo-individual multiplied by the weight a_i. Notice that $l_{y|\mathbf{x}}$ is the log-likelihood for the regression parameter $\boldsymbol{\beta}$, and does not involve the parameters of the distribution of covariates $\boldsymbol{\gamma}$, and vice versa for $l_{\mathbf{x}}$. The weights are a function of $\boldsymbol{\theta}^{(t)}$ and are constant in the M-step. Therefore, $Q(\ \boldsymbol{\theta} \mid \boldsymbol{\theta}^t\ )$ breaks into two portions, just like the complete-data log-likelihood, so the maximization to update the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ estimates can be done separately.

At each iteration of the EM algorithm, estimation of the multilocus genotype or haplotype distribution parameters will depend on the genetic model used, but in general it is similar to the maximization of a multinomial likelihood. The parameters are estimated by summing up the expected numbers belonging in each of the r genotype or haplotype classes and dividing by the total numbers in all classes. If no assumptions are made about the covariate distribution for the genetic model, the maximization can be seen by examining the relevant portion of $Q(\ \boldsymbol{\theta} \mid \boldsymbol{\theta}^t\ ), \sum_{i=1}^{M} a_i^{(t)} l_{\mathbf{x}}(\boldsymbol{\theta}|\mathbf{x}_i)$, more thoroughly. Each pseudo-individual having genetic covariate type k will contribute a portion $a_i^{(t)} \times 1 \times \log \gamma_k$ to the weighted log-likelihood, where γ_k is the probability that an individual is in genotype class k. For example, if the pseudo-individual is homozygous for haplotype 1, the contribution to the log-likelihood is $a_i^{(t)} \log \gamma_1$, where γ_1 is the probability of having two copies of haplotype 1. Note that $\sum_{i=1}^{M} a_i^{(t)} = n$ and since $\gamma_r = 1 - \sum_{j=1}^{r-1} \gamma_j$

$$\frac{\partial Q(\ \boldsymbol{\theta} \mid \boldsymbol{\theta}^t\ )}{\partial \gamma_k} = \frac{A_k^{(t)}}{\gamma_k} - \frac{n - \sum_{k=1}^{r-1} A_k^{(t)}}{\gamma_r},$$

where $A_k^{(t)} = \sum\ \{i \text{ with covariate type } k\}\ a_i^{(t)}$ is the expected count in covariate class k and $(n - \sum_{k=1}^{r-1} A_k^{(t)})$ is the number of pseudo-individuals in the last haplotype class r, both at the t^{th} iteration. Note that the derivative is like the derivative of a multinomial likelihood. Hence, setting the derivative equal to 0 and solving for γ_k, the parameters

are updated at the t^{th} EM iteration by

$$\gamma_k^{(t+1)} = \frac{A_k^{(t)}}{n}.$$

That is, the new estimate for γ_k is the expected count in covariate class k divided by the total count in all covariate classes.

To update the p regression parameters, the weighted logistic log-likelihood is maximized. The weighted log-likelihood is given by

$$\sum_{i=1}^{M} a_i \left\{ y_i \log(\frac{p_i}{1-p_i}) + \log(1-p_i) \right\},$$

for the logistic model, where p_i is the probability that the i^{th} person has the disease. Putting this into canonical form for the logistic model gives:

$$\sum_{i=1}^{M} a_i \left\{ y_i \delta_i - b(\delta_i) \right\},$$

where

$$\delta_i = \log(\frac{p_i}{1-p_i}) \quad \text{and} \quad b(\delta_i) = \log(1+e^{\delta_i}).$$

With a logit link function, we have:

$$\eta_i = \mathbf{x}_i\boldsymbol{\beta} = \delta_i \quad \text{and} \quad p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}.$$

Now, maximize $\sum_{i=1}^{M} a_i l_{y|\mathbf{x}}(\boldsymbol{\beta}|\mathbf{x}_i, y_i)$ with respect to $\boldsymbol{\beta}$ to find the new parameter estimates.

$$\frac{\partial(\sum_{i=1}^{M} a_i l_{y|\mathbf{x}}(\boldsymbol{\beta}|\mathbf{x}_i, y_i))}{\partial \beta_j} = \sum_{i=1}^{M} a_i \frac{\partial l}{\partial \delta_i}\frac{\partial \delta_i}{\partial p_i}\frac{\partial p_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

and

$$\begin{aligned}
\frac{\partial l}{\partial \delta_i} &= y_i - b'(\delta_i) = y_i - p_i, \\
\frac{\partial \delta_i}{\partial p_i} &= \frac{1}{p_i} + \frac{1}{1-p_i} = \frac{1}{p_i(1-p_i)}, \\
\frac{\partial p_i}{\partial \eta_i} &= \frac{e^{\eta_i}}{1+e^{\eta_i}} - \frac{e^{\eta_i}e^{\eta_i}}{(1+e^{\eta_i})^2} = p_i(1-p_i), \\
\frac{\partial \eta_i}{\partial \beta_j} &= x_{ij}.
\end{aligned}$$

Putting it all together,

$$\frac{\partial(\sum_{i=1}^{M} a_i l_{y|\mathbf{x}}(\boldsymbol{\beta}|\mathbf{x}_i, y_i))}{\partial \beta_j} = \sum_{i=1}^{M} a_i(y_i - p_i)x_{ij} \tag{3.1}$$

and the score function is

$$S(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial(\sum_{i=1}^{M} a_i l_{y|\mathbf{x}}(\boldsymbol{\beta}|\mathbf{x}_i, y_i))}{\partial \beta_1} \\ \frac{\partial(\sum_{i=1}^{M} a_i l_{y|\mathbf{x}}(\boldsymbol{\beta}|\mathbf{x}_i, y_i))}{\partial \beta_2} \\ \vdots \\ \frac{\partial(\sum_{i=1}^{M} a_i l_{y|\mathbf{x}}(\boldsymbol{\beta}|\mathbf{x}_i, y_i))}{\partial \beta_p} \end{bmatrix}$$

The new estimate $\boldsymbol{\beta}^{(t)}$ can be found by using a Newton-Raphson procedure or code for Fisher scoring. For the Newton-Raphson algorithm, the Hessian, $I(\boldsymbol{\beta}) = \frac{\partial^2(\sum_{i=1}^{M} a_i l_{y|\mathbf{x}}(\boldsymbol{\beta}|\mathbf{x}_i, y_i))}{\partial \beta_j \partial \beta_k}$ is needed.

$$\begin{aligned} I(\boldsymbol{\beta}) = \frac{\partial^2(\sum_{i=1}^{M} a_i l_{y|\mathbf{x}}(\boldsymbol{\beta}|\mathbf{x}_i, y_i))}{\partial \beta_j \partial \beta_k} &= \sum_{i=1}^{M} -a_i x_{ij} \frac{\partial p_i}{\partial \beta_k} \\ &= \sum_{i=1}^{M} -a_i x_{ij} \frac{\partial p_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \\ &= \sum_{i=1}^{M} -a_i x_{ij} x_{ik} p_i(1 - p_i) \end{aligned}$$

Therefore, the jk^{th} element of the Hessian is

$$I_{jk} = -\sum_{i=1}^{M} a_i x_{ij} x_{ik} p_i(1 - p_i)$$

and $I(\boldsymbol{\beta}) = -X'WVX$ where

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{Mp} \end{bmatrix},$$

$$W = \begin{bmatrix} a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_M \end{bmatrix}.$$

$$V = \begin{bmatrix} p_1(1-p_1) & 0 & 0 & \cdots & 0 \\ 0 & p_2(1-p_2) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & p_M(1-p_M) \end{bmatrix}.$$

These are the components needed to find the MLEs of the regression parameters using Newton-Raphson. The algorithm can easily be programmed. Briefly, the $t+1^{th}$ estimates of the regression parameters are found by solving

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - [I(\boldsymbol{\beta})^{-1} S(\boldsymbol{\beta})]|_{\beta=\beta^t}.$$

Fisher scoring involves taking the expected value of the negative of the Hessian matrix when covariates are known. For a logistic regression model with a canonical link, Newton-Raphson optimization and Fisher scoring are equivalent. Statistical software packages, such as Splus, have functions that find MLEs of generalized linear models (GLMs) using Fisher Scoring. These functions have an option to include a set of weights for each individual. Therefore, it is easy to implement the method of weights by writing a simple function to calculate the weights, then finding the new parameter estimates by using the package's pre-existing GLM procedure with the calculated weights.

3.3 Parameterization of the covariate distribution

It is possible to use population genetic theory to model the distribution of the covariates. This will reduce the number of parameters required for the covariate distribution. The first assumption that can be made is that there is independence between the genetic and environmental covariates. This means that the covariate vector for an

individual has two components $\mathbf{x}_g$ and $\mathbf{x}_e$ whose distributions depend only on the parameters $\boldsymbol{\gamma}_g$ and $\boldsymbol{\gamma}_e$, respectively. Therefore, $l_{\mathbf{x}}$ is a sum of genetic and environmental components: $l_{\mathbf{x}}(\boldsymbol{\gamma}) = l_{\mathbf{x_g}}(\boldsymbol{\gamma}_g) + l_{\mathbf{x_e}}(\boldsymbol{\gamma}_e)$.

Assuming that only the haplotype information is missing results in further simplification of the E step. Let $\mathbf{x}_{obs,i} = (\mathbf{x}_g, \mathbf{x}_e)_{obs,i}$, and assume as above. Then, the weights are calculated as

$$\begin{aligned} w_{ij}^{(t)} &= \Pr(\mathbf{x}_{\mathrm{i}} = \mathbf{x}^{(\mathrm{j})} |\ \mathbf{x}_{\mathrm{obs,i}}, \mathrm{y_i}, \boldsymbol{\theta}^{(\mathrm{t})}) \\ &= \frac{\Pr(\mathrm{y_i}|\mathbf{x}^{(\mathrm{j})}, \boldsymbol{\theta}^{(\mathrm{t})}) \Pr(\mathbf{x}^{(\mathrm{j})}|\ \boldsymbol{\theta}^{(\mathrm{t})})}{\sum \Pr(\mathrm{y_i}|\mathbf{x}^{(\mathrm{k})}, \boldsymbol{\theta}^{(\mathrm{t})}) \Pr(\mathbf{x}^{(\mathrm{k})}|\boldsymbol{\theta}^{(\mathrm{t})})} \quad \text{(assuming compatibility of the covariate vectors)} \\ &= \frac{\Pr(\mathrm{y_i}|\mathbf{x}^{(\mathrm{j})}, \boldsymbol{\theta}^{(\mathrm{t})}) \Pr(\mathbf{x}_{\mathrm{g}}^{(\mathrm{j})}|\ \boldsymbol{\theta}^{(\mathrm{t})}) \Pr(\mathbf{x}_{\mathrm{e}}^{(\mathrm{j})}|\ \boldsymbol{\theta}^{(\mathrm{t})})}{\sum \Pr(\mathrm{y_i}|\mathbf{x}^{(\mathrm{k})}, \boldsymbol{\theta}^{(\mathrm{t})}) \Pr(\mathbf{x}_{\mathrm{g}}^{(\mathrm{k})}|\boldsymbol{\theta}^{(\mathrm{t})}) \Pr(\mathbf{x}_{\mathrm{e}}^{(\mathrm{k})}|\ \boldsymbol{\theta}^{(\mathrm{t})})}. \end{aligned}$$

Note that the summation in the denominator is only over compatible covariate types. Since the environmental covariate is not missing, only covariate categories with the same environmental covariate value are acceptable. Thus, the probability of the environmental covariate in the numerator and denominator will cancel and so estimation of γ_e is not required. Both the independence of the genetic and environmental covariates, and the assumption of completely observed environmental covariates, mean that the algorithm can be implemented with continuous environmental covariates. Without these assumptions, γ_e would have to be estimated. Such estimation would be impractical with continuous environmental covariates unless their distributions could be specified in terms of a relatively small number of parameters

Another assumption that can be made is of Hardy-Weinberg equilibrium (HWE). This means that the probability of a multilocus genotype will be the probability of the haplotype squared if the individual is homozygous and 2 times the probabilities of each of the haplotypes if the person is heterozygous. For example, if the multilocus genotype is $abc/a'b'c'$, then

$$\Pr\{\mathrm{abc/a'b'c'}\} = \begin{cases} p_{abc}^2 & \text{if } abc = a'b'c' \\ 2p_{abc}p_{a'b'c'} & \text{otherwise} \end{cases}$$

where p_{abc} is the population frequency of the *abc* haplotype. With this assumption, the covariate model parameters are the probabilities of each of the haplotypes, so the number of parameters is $r - 1$, where r is the number of haplotypes. If Hardy-Weinberg equilibrium is not assumed, the number of parameters is equal to the number of possible multilocus genotypes minus one, or $r(r + 1)/2 - 1$. This assumption therefore results in a large reduction in the number of parameters. Even for two-locus multilocus genotypes there are 4 haplotypes and 10 multilocus genotypes. A pseudo-individual homozygous for haplotype k will contribute $2 \times 1 \times \log \gamma_k$ to the expected complete-data log-likelihood for $\boldsymbol{\gamma}$ since their weight is 1, and a heterozygous pseudo-individual with haplotypes k and l will contribute $a_i^{(t)} \log \gamma_k \gamma_l = a_i^{(t)} \log \gamma_k + a_i^{(t)} \log \gamma_l$ to the expected complete-data log-likelihood for $\boldsymbol{\gamma}$. In the maximization step, the probability distribution of the genetic covariates is again multinomial-like, with n now equal to the total number of haplotypes in the sample (two times the sample size) and $A_k^{(t)} = \sum_{i=1}^{M} a_i^{(t)} \times$(# copies of haplotype k in individual i). The new estimates for the haplotype frequencies are then estimated similar to the standard EM algorithm described in section 2.2 for estimating haplotype frequencies from single-locus genotypes: sum up the number of expected copies of a haplotype and divide by 2 times the number of haplotypes.

3.4 Standard errors for regression parameters

Maximum likelihood estimation of the regression parameters and covariate parameters was outlined in the previous section describing the weighted EM algorithm. The standard errors output from programs to fit a generalized linear model with weights are not correct in this situation since they do not take into account the uncertainty of the covariates. For this reason, the variance covariance matrix for the parameter estimators must be derived separately.

Louis (1982) showed that

$$\begin{aligned} \mathrm{I}(\theta) &= E_\theta[\ \mathrm{I}_c(\theta) \mid \mathbf{x}_{obs}, \mathbf{y}\ ] \\ &\quad -(E_\theta[\ \mathrm{S}_c(\theta)\mathrm{S}_c(\theta)^T \mid \mathbf{x}_{obs}, \mathbf{y}\ ] - E_\theta[\ \mathrm{S}_c(\theta) \mid \mathbf{x}_{obs}, \mathbf{y}\ ]E_\theta[\ \mathrm{S}_c(\theta)^T \mid \mathbf{x}_{obs}, \mathbf{y}]) \\ &= E_\theta[\ \mathrm{I}_c(\theta) \mid \mathbf{x}_{obs}, \mathbf{y}\ ] - \mathrm{cov}[\ \mathrm{S}_c(\theta) \mid \mathbf{x}_{obs}, \mathbf{y}\ ] \end{aligned}$$

where $\mathrm{I}(\theta)$ is the negative Hessian of the observed data log-likelihood, and $\mathrm{I}_c(\theta)$ and $\mathrm{S}_c(\theta)$ are the negative Hessian and score of the complete-data log-likelihood function, respectively.

We assume independence between the genetic and environmental covariates, fully observed environmental covariates and HWE, as in section 3.3. Given the assumptions, the observed data likelihood is a sum of two terms, one involving the $(\boldsymbol{\beta}, \boldsymbol{\gamma}_g)$ and the other involving only $\boldsymbol{\gamma}_e$ since

$$\begin{aligned} \log \Pr(\ \mathbf{X} = \mathbf{x}_{obs}, \mathbf{y}|\boldsymbol{\theta}) &= \log(\sum_{\{\mathbf{x}\ compatible\ with\ \mathbf{x}_{obs}\}} \Pr(\ \mathbf{x}, \mathbf{y}|\boldsymbol{\theta})) \\ &= \log(\sum \Pr(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\Pr(\mathbf{x}|\boldsymbol{\theta})) \\ &= \log(\sum \Pr(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta})\Pr(\mathbf{x}_g|\boldsymbol{\gamma}_g)\Pr(\mathbf{x}_e|\boldsymbol{\gamma}_e)) \\ &= \log(\Pr(\mathbf{x}_e|\boldsymbol{\gamma}_e)\sum \Pr(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta})\Pr(\mathbf{x}_g|\boldsymbol{\gamma}_g)) \\ &= \log \Pr(\mathbf{x}_e|\boldsymbol{\gamma}_e) + \log(\sum \Pr(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta})\Pr(\mathbf{x}_g|\boldsymbol{\gamma}_g)). \end{aligned}$$

Therefore, the second derivative of $\log L(\boldsymbol{\theta})$ can be written as the block diagonal matrix

$$\begin{bmatrix} \frac{\partial^2 \log(\sum \Pr(\mathbf{y}|\mathbf{x},\boldsymbol{\beta})\Pr(\mathbf{x}_g|\boldsymbol{\gamma}_g))}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{*T}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial^2 \log \Pr(\mathbf{x}_e|\boldsymbol{\gamma}_e)}{\partial \boldsymbol{\gamma}_e \partial \boldsymbol{\gamma}_e^T} \end{bmatrix}$$

where $\boldsymbol{\theta}^* = (\boldsymbol{\beta}, \boldsymbol{\gamma}_g)$. Since we are only interested in estimating the variance of the regression parameters, we do not need to compute $\frac{\partial^2 \log \Pr(\mathbf{x}_e|\boldsymbol{\gamma}_e)}{\partial \boldsymbol{\gamma}_e \partial \boldsymbol{\gamma}_e^T}$; the standard errors for $\boldsymbol{\beta}$ can be obtained by inverting the upper left-hand sub-matrix, $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}_\mathrm{g}) = \frac{\partial^2 \log(\sum \Pr(\mathbf{y}|\mathbf{x},\boldsymbol{\beta})\Pr(\mathbf{x}_\mathrm{g}|\boldsymbol{\gamma}_\mathrm{g}))}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{*\mathrm{T}}}$.

Let $l(\boldsymbol{\theta}^*) = l(\boldsymbol{\theta}^*|\mathbf{x}, \mathbf{y})$ and $l_i(\boldsymbol{\theta}^*) = l(\boldsymbol{\theta}^*|\mathbf{x}_i, y_i)$. Then applying Louis' formula,

the first term for $\mathrm{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}_\mathrm{g})$ can be found as follows

$$\begin{aligned}
\mathrm{E}[\ I_c(\boldsymbol{\beta}, \gamma_g)|\mathbf{x}_{obs}, \mathbf{y}\ ] &= \mathrm{E}[\ -\sum_{i=1}^{n} \frac{\partial^2 l_i(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{*T}} |\mathbf{x}_{obs}, \mathbf{y}\ ] \\
&= -\sum_{i=1}^{n} \mathrm{E}[\ \frac{\partial^2 l_i(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{*T}} |\mathbf{x}_{obs}, \mathbf{y}\ ] \\
&= -\sum_{i=1}^{n} \sum_{j=1}^{|\mathcal{S}_i|} \frac{\partial^2 l_i(\boldsymbol{\theta}^*|\mathbf{x}^{(j)})}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{*T}} \Pr(\mathbf{x}_i = \mathbf{x}^{(j)}|\mathbf{x}_{obs,i}, y_i) \\
&= -\sum_{i=1}^{n} \sum_{j=1}^{|\mathcal{S}_i|} w_{ij} \frac{\partial^2 l_i(\boldsymbol{\theta}^*|\mathbf{x}^{(j)})}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{*T}}.
\end{aligned}$$

Since $l(\boldsymbol{\beta}, \gamma_g) = l_{Y|X}(\boldsymbol{\beta}) + l_X(\boldsymbol{\gamma}_g)$, $\mathrm{E}\{I_c(\boldsymbol{\beta}, \gamma_g)|\mathbf{x}_{obs}, \mathbf{y}\}$ is the block diagonal matrix

$$-\begin{bmatrix} \mathrm{E}[\ \frac{\partial^2 l_{Y|X}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} |\mathbf{x}_{obs}, \mathbf{y}\ ] & \mathbf{0} \\ \mathbf{0} & \mathrm{E}[\ \frac{\partial^2 l_X(\boldsymbol{\gamma}_g)}{\partial \boldsymbol{\gamma}_g \partial \boldsymbol{\gamma}_g^T} |\mathbf{x}_{obs}, \mathbf{y}\ ] \end{bmatrix}.$$

Note that $-E[\ \frac{\partial^2 l_{Y|X}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} |\ \mathbf{x}_{obs}, \mathbf{y}\ ] = X^T W V X$ from the calculations for obtaining regression estimates using Newton-Raphson in section 3.2.2.

To compute $\mathrm{E}[\ \frac{\partial^2 l_X(\boldsymbol{\gamma}_g)}{\partial \boldsymbol{\gamma}_g \partial \boldsymbol{\gamma}_g^T} |\ \mathbf{x}_{obs}, \mathbf{y}\ ]$, the assumption of Hardy-Weinberg equilibrium is used. Supposing complete covariate data on subject i, the contribution from i to the log-likelihood of $\boldsymbol{\gamma}_g$ is $\log \gamma_{g,i}$, where $\gamma_{g,i}$ is the probability of i's two haplotypes. For example, if i has haplotypes l and k and there are r possible haplotypes in the population, the assumption of HWE gives

$$\gamma_{g,i} = \begin{cases} \gamma_l^2 & \text{if } l = k \\ 2\gamma_l \gamma_k & \text{if } l \neq k \end{cases},$$

where $\gamma_r = 1 - \sum_{j=1}^{r-1} \gamma_j$. Thus,

$$\log \gamma_{g,i} = \begin{cases} 2 \log \gamma_l & \text{if } l = k \\ \log 2 + \log \gamma_l + \log \gamma_k & \text{if } l \neq k \end{cases}.$$

Or, if $\mathbf{x}_{g,i}$ gives the row vector of length r of haplotype counts for i (that is, an element in $\mathbf{x}_{g,i}$ is 0,1 or 2 depending on the number of copies of that haplotype that i has)

then the contribution to the log-likelihood $l_{\mathbf{X}}(\boldsymbol{\gamma}_g)$ can be written as the product of the two vectors

$$\mathbf{x}_{g,i} \begin{bmatrix} \log \gamma_1 \\ \log \gamma_2 \\ \vdots \\ \log(1 - \sum_{i=1}^{(r-1)} \gamma_i) \end{bmatrix}_{r \times 1},$$

Thus,

$$\frac{\partial l_i(\boldsymbol{\gamma}_g)}{\partial \gamma_k} = \mathbf{x}_{g,i} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1/\gamma_k \\ 0 \\ \vdots \\ 0 \\ -1/(1 - \sum_{i=1}^{(r-1)} \gamma_i) \end{bmatrix}_{r \times 1}, \tag{3.2}$$

$$\frac{\partial^2 l_i(\boldsymbol{\gamma}_g)}{\partial \gamma_k \partial \gamma_k} = \mathbf{x}_{g,i} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -1/\gamma_k^2 \\ 0 \\ \vdots \\ 0 \\ -1/(1 - \sum_{i=1}^{(r-1)} \gamma_i)^2 \end{bmatrix}_{r \times 1},$$

and

$$\frac{\partial^2 l_i(\boldsymbol{\gamma}_g)}{\partial\gamma_k\partial\gamma_l} = \mathbf{x}_{g,i}\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ -1/(1-\sum_{i=1}^{(r-1)}\gamma_i)^2 \end{bmatrix}_{r\times 1}.$$

Finally $\mathrm{E}[\ \frac{\partial^2 l_X(\boldsymbol{\gamma}_g)}{\partial\boldsymbol{\gamma}_g\partial\boldsymbol{\gamma}_g^T}|\ \mathbf{x}_{obs},\mathbf{y}\ ] = -\sum_{i=1}^{n}\sum_{j=1}^{|\mathcal{S}_i|} w_{ij}\frac{\partial^2 l_X(\boldsymbol{\gamma}_g|\mathbf{x}^{(j)})}{\partial\boldsymbol{\gamma}_g\partial\boldsymbol{\gamma}_g^T}$ with the second derivatives given above, or written more explicitly

$$\left[\mathrm{E}\ [-\frac{\partial^2 l_X(\boldsymbol{\gamma}_g)}{\partial\boldsymbol{\gamma}_g\partial\boldsymbol{\gamma}_g^T}|\mathbf{x}_{obs},\mathbf{y}\ ]\right]_{kk} = \frac{1}{\gamma_k^2}\{\sum_{i=1}^{n}\sum_{j=1}^{|\mathcal{S}_i|} w_{ij}x_{k,i}^{(j)}\} + \frac{1}{(1-\sum_{i=1}^{(r-1)}\gamma_i)^2}\{\sum_{i=1}^{n}\sum_{j=1}^{|\mathcal{S}_i|} w_{ij}x_{r,i}^{(j)}\},$$

and

$$\left[\mathrm{E}[\ -\frac{\partial^2 l_X(\boldsymbol{\gamma}_g)}{\partial\boldsymbol{\gamma}_g\partial\boldsymbol{\gamma}_g^T}|\mathbf{x}_{obs},\mathbf{y}\ ]\right]_{kl} = \frac{1}{(1-\sum_{i=1}^{(r-1)}\gamma_i)^2}\{\sum_{i=1}^{n}\sum_{j=1}^{|\mathcal{S}_i|} w_{ij}x_{r,i}^{(j)}\},$$

where $x_{k,i}^{(j)}$ refers to the k^{th} element in $\mathbf{x}_{g,i}^{(j)}$, the j^{th} possible row vector of haplotype counts for subject i that is consistent with that subject's observed genotype data.

It can then be shown that the first term in Louis' formula for the observed Fisher information (the conditional expected value of the complete-data information given the observed data) can be written in matrix form as

$$\begin{bmatrix} X^TWVX & 0 \\ 0 & NG + n_r/(1-\sum_{i=1}^{(r-1)}\gamma_i)^2 J \end{bmatrix},$$

where X, V and W are the augmented matrices as defined in section 3.2.2, $N_{(r-1)\times(r-1)}$ is a diagonal matrix whose elements are the sums over all individuals of the weighted counts of the first $r-1$ haplotypes, n_r is the sum of the weighted numbers of the r^{th} haplotype, $G_{(r-1)\times(r-1)} = \mathrm{diag}(1/\gamma_k^2)$ and $J_{(r-1)\times(r-1)}$ is a matrix of 1's.

The second term in $\mathrm{I}(\boldsymbol{\beta},\boldsymbol{\gamma}_\mathrm{g})$ using Louis' formula, $\mathrm{cov}[\mathrm{S_c}(\theta)|\mathbf{x}_\mathrm{obs},\mathbf{y}]$, can be found by first noting that the n subjects are independent. Therefore $\mathrm{S_c}(\boldsymbol{\beta},\boldsymbol{\gamma}_\mathrm{g}) =$

$\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{S_{c,i}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_{\mathrm{g}})$ and

$$\begin{aligned}
\mathrm{cov}[\mathrm{S_c}(\boldsymbol{\beta}, \boldsymbol{\gamma}_{\mathrm{g}})|\mathbf{x}_{\mathrm{obs}}, \mathbf{y}] &= \mathrm{cov}\left[\sum_{i=1}^{n} \mathrm{S_{c,i}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_{\mathrm{g}})|\mathbf{x}_{\mathrm{obs}}, \mathbf{y}\right] \\
&= \sum_{i=1}^{n} \mathrm{cov}\left[\mathrm{S_{c,i}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_{\mathrm{g}})|\mathbf{x}_{\mathrm{obs}}, \mathbf{y}\right] \\
&= \sum_{i=1}^{n} \big\{\mathrm{E}[\mathrm{S_{c,i}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_{\mathrm{g}}), \mathrm{S_{c,i}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_{\mathrm{g}})^{\mathrm{T}}|\mathbf{x}_{\mathrm{obs}}, \mathbf{y}] - \\
&\qquad \mathrm{E}[\mathrm{S_{c,i}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_{\mathrm{g}})|\mathbf{x}_{\mathrm{obs}}, \mathbf{y}]\mathrm{E}^{\mathrm{T}}[\mathrm{S_{c,i}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_{\mathrm{g}})|\mathbf{x}_{\mathrm{obs}}, \mathbf{y}]\big\}
\end{aligned}$$

For those people whose covariate information is fully known, the variance is 0, therefore the sum is only over those n^* pseudo-individuals corresponding to subjects with ambiguous haplotype phase. Substituting the expected values for the score using the weight notation and letting $S_{c,ij}$ be the complete data score on the j^{th} covariate type compatible with the observed data for subject i gives

$$\begin{aligned}
&\mathrm{cov} \quad [\ \mathrm{S}_c(\boldsymbol{\beta}, \boldsymbol{\gamma}_g) \mid \mathbf{x}_{obs}, \mathbf{y}\ ] \\
&= \sum_{i=1}^{n^*} \left\{ \sum_{j=1}^{|\mathcal{S}_i|} S_{c,ij}(\boldsymbol{\beta}, \boldsymbol{\gamma}_g), S_{c,ij}(\boldsymbol{\beta}, \boldsymbol{\gamma}_g)^T w_{ij} - (\sum_{j=1}^{|\mathcal{S}_i|} S_{c,ij}(\boldsymbol{\beta}, \boldsymbol{\gamma}_g) w_{ij})(\sum_{k=1}^{|\mathcal{S}_i|} S_{c,ik}(\boldsymbol{\beta}, \boldsymbol{\gamma}_g) w_{ik})^T \right\} \\
&= \sum_{i=1}^{n^*} \sum_{j=1}^{|\mathcal{S}_i|} S_{c,ij}(\boldsymbol{\beta}, \boldsymbol{\gamma}_g), S_{c,ij}(\boldsymbol{\beta}, \boldsymbol{\gamma}_g)^T w_{ij} - \sum_{i=1}^{n^*} \sum_{j=1}^{|\mathcal{S}_i|} \sum_{k=1}^{|\mathcal{S}_i|} S_{c,ik}(\boldsymbol{\beta}, \boldsymbol{\gamma}_g) S_{c,ij}(\boldsymbol{\beta}, \boldsymbol{\gamma}_g)^T w_{ij} w_{ik}.
\end{aligned}$$

Let S be a matrix whose rows are complete-data score vectors $S_{c,ij}$ for pseudo-individuals, with rows arranged so that the pseudo-individuals for subject i are in consecutive rows. Then, it can be shown that the covariance matrix can be written as

$$S^T W S - S^T W B W S,$$

where W is the diagonal matrix of weights corresponding to each individual and B is a block diagonal matrix of 1's with the number of rows and columns of each block equal to the number of haplotypes compatible with a given subject's observed covariates. For example, if there are three compatible covariate vectors for one person, and two

compatible for a second person, the block diagonal matrix would be

$$B = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

Note that for all matrices S, W and B, the number of rows is n^* since the summation is only over the pseudo-individuals for subjects with ambiguous haplotype phase.

The score matrix S can be broken up into columns for the derivatives with respect to the regression parameters and derivatives with respect to the covariate parameters. That is

$$S = \begin{bmatrix} \; S_{\boldsymbol{\beta}} \mid S_{\boldsymbol{\gamma}_g} \; \end{bmatrix}.$$

Using the logistic model as defined, and assuming HWE for the haplotypes,

$$S_{\boldsymbol{\beta}} = HX \text{ and } S_{\boldsymbol{\gamma}_g} = X_{gen}\tilde{G}$$

where X is the augmented matrix of covariates used in the regression model for the ambiguous individuals only and H is a diagonal matrix of elements $(y_i - p_{ij})$, where $p_{ij} = \Pr\{\mathrm{y_i} = 1|\mathbf{x}_\mathrm{i}^{(\mathrm{j})}, \boldsymbol{\beta}\}$ which corresponds to the first derivatives of the regression parameters. Note that in equation 3.1, each pseudo-individual's unweighted contribution to the derivative with respect to β_j is $(y_i - p_{ij})x_{ij}$. Since $[S_{\boldsymbol{\beta}}]_{ij}$ represents the i^{th} pseudo-individuals contribution to the derivative with respect to β_j, each element in the matrix is given by $(y_i - p_{ij})x_{ij}$ or in matrix notation $S_{\boldsymbol{\beta}} = HX$. The X_{gen} matrix has rows of length r which count the haplotypes for each augmented individual, and the matrix $\tilde{G}$ is

$$\begin{bmatrix} 1/\gamma_1 & 0 & \cdots & 0 \\ 0 & 1/\gamma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 1/\gamma_{r-1} \\ -1/(1-\sum_{i=1}^{(r-1)}\gamma_i) & -1/(1-\sum_{i=1}^{(r-1)}\gamma_i) & \cdots & -1/(1-\sum_{i=1}^{(r-1)}\gamma_i) \end{bmatrix}_{r\times(r-1)}.$$

The justification for $S_{\boldsymbol{\gamma}_g} = X_{gen}\tilde{G}$ can be seen by examining equation 3.2, which is the i^{th} pseudo-individual's contribution to the derivative with respect to γ_k. If the row vector $\mathbf{x}_{g,i}$, which is a count of each of the haplotypes, is multiplied by $\tilde{G}$, the corresponding row vector is i's contribution to the derivative for $\boldsymbol{\gamma}$. If each pseudo-individuals haplotype row vector is a row of the matrix X_{gen}, then $S_{\boldsymbol{\gamma}_g} = X_{gen}\tilde{G}$. Should any of the genetic modeling assumptions be changed, the variance need not be derived again, only the score matrix for the new parameterization.

Therefore, the observed Fisher information, found using Louis' formula, is

$$\begin{bmatrix} X^T W V X & 0 \\ 0 & NG + n_r/(1 - \sum_{i=1}^{(r-1)} \gamma_i)^2 J \end{bmatrix} - S^T(W - WBW)S.$$

The variance can now be estimated by substituting the estimated parameters and the weights from the final iteration of the EM algorithm into the matrix given above and calculating the inverse.

Chapter 4

Analysis of a simulated data set

Logistic regression can be useful in finding the effect of genetic markers on the probability of getting a disease. SNP markers are becoming increasingly popular as genetic markers due to their abundance in the genome and the more cost-effective methods of genotyping. They are diallelic and are expected to have high linkage disequilibrium between adjacent markers in a candidate region. Hence, we consider the haplotypes as the covariates instead of genotypes at a single locus. Since haplotypes are not observed for all individuals, in chapter 3, the method of weights, an implementation of the EM algorithm for generalized linear models, was described for a logistic model with haplotypes as the missing covariates. The method of weights as originally formulated does not incorporate the biologically reasonable assumptions of HWE and independence of genetic and environmental covariates. We incorporated these assumptions, which are standard in genetic analyses, in order to improve the statistical efficiency and stability of the resulting parameter estimates. In addition, these assumptions simplify the implementation of the algorithm and calculation of standard errors. Under these assumptions, the standard errors of the regression coefficients were derived using Louis' formula for the observed Fisher information. In this chapter, the resulting EM procedure is applied to an association scan of a simulated candidate region for a complex disease.

The genetic data were simulated under a neutral coalescent with recombination (Hudson 1983), using a C program kindly provided by R.R. Hudson. The data are

from a hypothetical cohort study of a common complex, late-onset genetic disease in which 500 disease-free subjects between the ages of 50 and 70 years were sampled from a closed population isolate and followed for a short period of time. Genotype data for 196 SNP markers in a 5 cM candidate region were simulated for all individuals assuming HWE. The average spacing of the markers is 23.7 kilobase pairs (kbp). The susceptibility locus was removed from the dataset for analysis. The susceptibility allele frequency in the sample is 28%. The environmental data simulated for each individual was their age, gender and Body Mass Index (BMI). The environmental covariates were simulated independently of the genetic information and mimicked covariate values found in middle-aged and older individuals in North American populations. Affection status was generated randomly, with disease probability determined by a logit function of the number of susceptibility alleles, age, BMI, and a "polygene" effect. More information about both the population-genetic model and the risk model used to simulate the data can be found in Appendix A, which also provides an excerpt of the data.

For the analysis, haplotypes from adjacent pairs of loci were used as the genetic data. If the two alleles at each SNP locus are denoted by 1 and 0, the four possible haplotypes are 00, 01, 10 and 11. There are 10 possible multilocus genotypes and 9 possible combinations of single-locus genotypes. Only the double heterozygote 0/1,0/1 has unknown haplotype phase and there are two possible multilocus genotypes consistent with the single-locus genotypes: 00/11 and 01/10. A haplotype-dose model is fit, so if nXX represents the number of XX haplotypes, the model equation is:

$$\text{logit}(p) = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{gender} + \beta_3 \times \text{BMI} + \beta_4 \times n01 + \beta_5 \times n10 + \beta_6 \times n11.$$

The number of 00 haplotypes is $2 - n01 - n10 - n11$. There are 196 SNP loci in total, so moving along the haplotype, the first analysis is for the haplotypes from locus 1 and 2, the second from locus 2 and 3 and so on. In total, 195 analyses are done.

The weighted EM method was implemented using R (www.r-project.org). For each analysis, the single-locus data for the two loci was first converted into haplotypes. The dataset was augmented by the appropriate number of pseudo-individuals. Since

there are only two possible multilocus genotypes consistent with the observed genotype for the phase-unknown individuals, an extra row was added for these individuals with only the haplotype information different between the two rows.

For each locus pair, EM was performed using the haplotype and environmental covariates. The initial values of the regression parameters were found by first fitting the non-weighted logistic model to the augmented dataset. The haplotype frequencies are all initially set equal. A drawback to this initialization of the haplotype parameters versus estimating the initial values from the unambiguous individuals is that the parameters could be far from their true values, slowing down the EM convergence. However, these initial values are then used to calculate the weights that are the initial values in the E step, and usually less than 10 iterations were required for convergence. Using the augmented dataset and the associated weights, the R `glm` routine is used to calculate the new estimates of the regression parameters. The new haplotype parameters are calculated by summing up the weighted numbers of each haplotype and dividing by 1000 (two times the sample size). These values are then used to calculate the next set of weights, and the process continues until convergence. The standard errors are then calculated using the final set of weights, estimates and fitted values.

For 30 of the 195 locus sets, not all four haplotypes were observed in the individuals whose haplotype phase was known. This is probably partially due to the fact that the markers for these loci are less than 10 kbp apart. The distance is so small that few recombinations have occurred between the two loci, causing a greater allelic association and fewer haplotypes. In these cases, the EM algorithm will not converge because it is attempting to maximize a likelihood that is formulated assuming the existence of all 4 haplotypes in the population. We therefore make the reasonable simplifying assumption that if a haplotype is not observed in the non-ambiguous individuals, it does not exist in the population. Under these circumstances, the haplotype phase for the double heterozygote individuals will be known as well. Since with this assumption there is no missing haplotype information, the unweighted `glm` function was used on the logistic model to estimate the regression coefficients of the remaining haplotypes.

For each set of loci, we test whether there is a genetic effect from the haplotypes. The null hypothesis is that there is no haplotype effect or $\beta_4 = \beta_5 = \beta_6 = 0$. A Wald test was used to test this hypothesis and p-values were calculated using the asymptotic χ^2 distribution of the test statistic, with 3 degrees of freedom for the datasets with all haplotypes observed, and 2 or 1 degrees of freedom for those with one or two haplotypes not observed in the sample. Alternatively, a likelihood ratio test could be performed using the log-likelihoods at convergence for both models. A score test for a genetic effect described by Schaid *et al.* (2002) is another possibility. To visualize the relative departures from this hypothesis for each of the sets of loci, the negative base-10 logarithm of the p-value was plotted versus the midpoint position of the two loci on the chromosome, with the position measured in kbp relative to the true disease susceptibility locus (Figure 4.1).

From the plot, the peak over all loci pairs in the scan occurs for loci 86 and 87 (-80 kbp to -77.5 kbp from the true location) and it is distinct in its height from the other peaks. The chromosomal distance between these two markers is 2.5 kbp, a size so small that two haplotypes are unobserved. The tests at the sets of loci adjacent to the peak also give relatively high values. The haplotype spanning the disease-predisposing locus is 132 kbp, which is a relatively large distance. Since this distance is large, more recombinations are expected to have occurred between the two markers. This would break up the ancestral haplotype bearing the disease mutation, making it more probable that any single haplotype is no more likely to be associated with the disease than another.

The analysis was repeated removing the environmental covariates to see what effect adjusting for environmental covariates has in finding the genetic signal. Even though it is known that in many complex diseases environmental factors play a significant role in developing disease, many association analyses that use haplotypes, such as the χ^2 test described in section 2, do not take the environmental factors into account. Figure 4.2(a) is a plot of the $-\log_{10}$(p-value) from the Wald test for haplotype effect for the model without environmental covariates versus the midpoint of the haplotype position relative to the disease locus, and Figure 4.2(b) has superimposed the analyses with and without the environmental covariates to compare the two results.

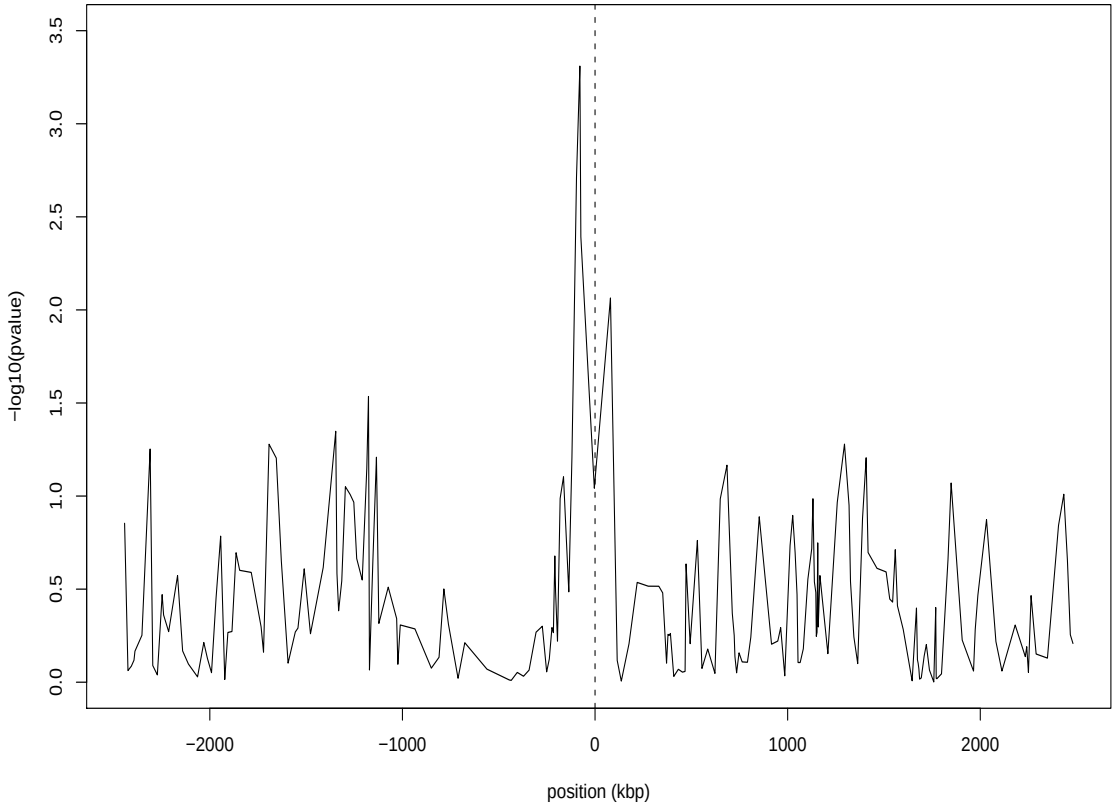

Figure 4.1: Genetic signal for haplotype effect versus chromosome position; model with environmental covariates.

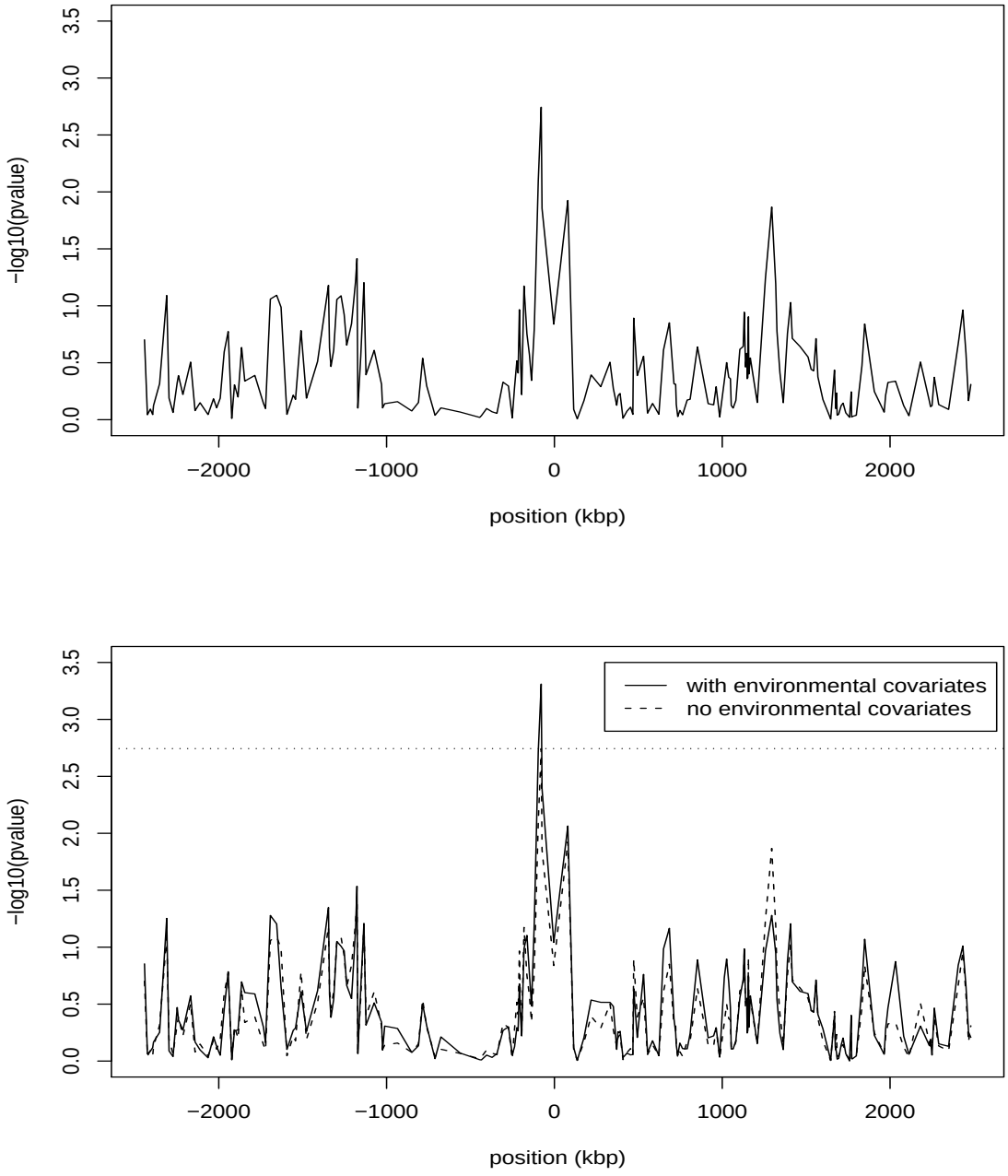

Figure 4.2: (a)Genetic signal for haplotype effect versus chromosome position; model without environmental covariates. (b) The analysis with the environmental covariates and without superimposed. The horizontal line is the maximum signal for the analysis with no environmental covariates.

In the analysis without environmental covariates, the peak again occurs for the 86th and 87th markers. However, the peak height is lower than in the analysis that adjusted for environmental factors. A second high peak occurs at the haplotype for markers 145 and 146, 1272.5 kbp to 1317.5 kbp from the true location. Without knowing where the disease locus is actually located, both regions might be considered to have strong enough signals relative to the other regions to pursue further studies on. In the analysis with the environmental factors however, this second peak is not present, meaning that the association in that region is explained by correcting for environmental factors. Therefore, adjusting for environmental factors had two benefits in this simulated dataset. The height of the true peak was increased, making it less likely for the genetic signal to be missed, and the false peak was decreased to such a point where it would not be considered a signal.

In fitting the logistic model along the haplotype for adjacent locus pairs, a total of 195 tests are performed. Since multiple tests are done, the p-value at each locus pair is lower than it should be if the number of tests is taken into account, making a Type I error more likely. However, the scan is a way to narrow down the candidate region to areas where the disease locus is more likely to be found so that those searching for the cause of the disease-predisposing locus do not waste resources on unlikely regions. The p-value can be considered as another measure of the genetic signal. From the results of a candidate gene scan such as this, the candidate region of 5 cM has now been reduced to a much smaller region. The Bonferroni corrected p-value for the analysis including environmental covariates would be about .096, and 0.35 for the analysis which does not correct for environmental covariates. However, this is likely too conservative since tests are not independent; tests are positively correlated owing to linkage disequilibrium. By contrast, a much stronger indication of the genetic signal was obtained from permutation tests based on 650 permutations of the affection status, which yielded p-values of 0.03 for the analysis including environmental covariates and 0.12 for the analysis without environmental covariates.

The results of the analysis with haplotypes can also be compared to the results

using a single-locus genetic covariate and the environmental covariates in the regression. A Wald test was used to test for the genetic effect of each of the 196 loci. Figure 4.3 shows $-\log_{10}$(p-value) versus the position along the chromosome (a) in the single-locus analysis and (b) the haplotype and single-locus results superimposed. For both analyses, the peaks are at the same location and are about the same height. This is not an unexpected result; Kaplan and Morris (2001) found that the single-locus and haplotype based χ^2 tests often give concordant results in simulation studies based on a single susceptibility allele at the disease-predisposing locus, as in this case. The difference in the two analyses for this scan of a candidate region seems to occur in the non-signal regions. The haplotype model seems to have smoothed out some of the peaks in the regions far away from the actual disease locus. The potential smoothing effect that using haplotypes has on the p-values was also noted in Akey *et al.* (2001).

In conclusion, including haplotypes and environmental covariates in a logistic regression model to find association between a haplotype and disease has some advantages. When compared to the single-locus logistic regression model, the haplotype analysis was as good in finding the true peak signal, but was better at smoothing out the false peaks. When compared to the logistic regression without environmental covariates, the haplotype analysis with environmental factors had a higher peak close to the true disease locus and removed strong signals farther from the disease locus. Therefore, for this simulated gene-scan data, the haplotype model analysis that adjusts for environmental factors decreased the false-positive rate and increased power.

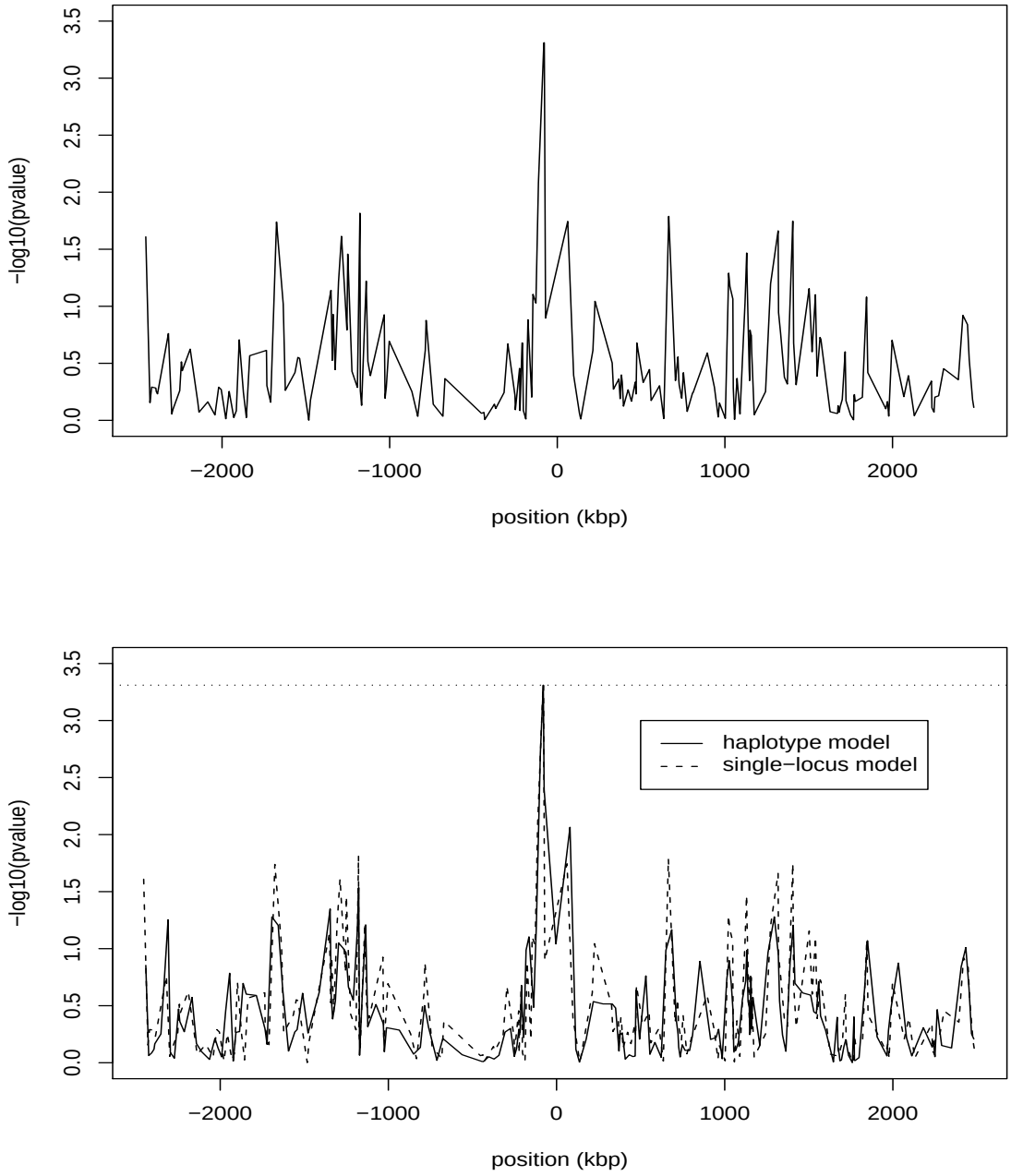

Figure 4.3: (a)Genetic signal for locus effect versus chromosome position. (b) Haplotype model and single-locus model results superimposed. Horizontal line indicates maximum signal from single-locus analysis.

Chapter 5

Summary and conclusions

Gene mapping for complex diseases is more complicated than for Mendelian diseases. Complex diseases are multi-factorial– there may be several environmental and genetic factors that contribute to the disease. For the genetic component of the disease, there may be multiple genes or mutations which contribute to the disease, or multiple genes that cause an indistinguishable disease phenotype. Thus, association methods which account for the contribution of one gene may not be sufficient to find the disease-predisposing gene without controlling for other important factors.

It has been suggested that the power of association methods can be improved by using haplotypes rather than single loci (Akey *et al.* 2001; Daly *et al.* 2001; Zöllner and von Haeseler 2001). Haplotypes are only observed if there is genotypic information on family members or if high cost laboratory techniques are used. In the case of association studies, neither of these approaches are feasible. For this reason, methods of estimating the haplotypes and the haplotype frequencies, such as Clark's method (1990), the EM algorithm (Long *et al.* 1995; Hawley and Kidd 1995; Excoffier and Slatkin 1995) and Bayesian approaches (Stephens *et al.* 2001; Niu *et al.* 2002) have been developed. The statistical approaches have been shown to be good at estimating the haplotypes and frequencies, but the consequences of using imputed haplotypes in further statistical analyses has not been studied. Since further analysis will not include the uncertainty associated with the reconstructed haplotypes, one might be concerned about any conclusions based on the results. Furthermore,

available information on disease status and environmental covariates is not used in the haplotype reconstructions.

Our method involves a logistic regression model of association between a marker genotype and disease in a cohort sample. In order to use haplotypes as covariates in the regression, a likelihood method for generalized linear models with missing covariates called the method of weights is proposed. With this EM implementation, the haplotype values for each individual are not imputed. Instead, the weights account for the uncertainty of the haplotype phase. In the calculation of the weights, we have included a population-genetic model of genotype frequencies by assuming Hardy-Weinberg equilibrium, and we have assumed independence of the genetic and environmental covariates. This makes the estimation in the M step more efficient and stabilizes the estimated values. Using EM to estimate the coefficients directly, versus estimating the haplotype covariates and treating them as observed as is traditionally done, allows us to calculate standard errors using Louis' formula (1982) and correct for the uncertainty in the haplotypes explicitly.

In the simulated gene-scan dataset, both single-loci and haplotypes were used as covariates. Both analyses showed the most evidence for association around marker 86 and an association region of about 200 kbp. In fact, for both analyses the height of the strongest genetic signal is about the same. The difference in the two analyses on the simulated dataset is seen in the other regions, where the haplotype analysis has smoothed out some of the other peaks of genetic signal.

Perhaps the most important consequence of using logistic regression for association studies is the flexibility of the modeling that it allows. Not only can environmental factors easily be included in the model, but interaction between the genetic and environment effects can also be modeled. Since it is believed that these interactions are significant to disease development, this is an important benefit of this method. Not only can environmental covariates be included, but other unlinked genes which are known to also have an effect on the disease can easily be included as covariates in the analysis as well.

The effect of adjusting for environmental covariates in the analysis was shown using the simulated gene scan data. In the logistic regression analysis that only

included haplotypes, the association close to the true disease locus was not as strong as that from the analysis that included environmental information, and a false peak was present. One would not be as confident in the true peak, since it is not much higher than the false one.

Overall, the results of the data analysis suggest that the proposed approach to association mapping is more powerful than other approaches which do not include environmental factors. However, these results are based on only the one dataset that was simulated with a genetic model. The performance of the method should be assessed on data simulated under a variety of scenarios including violation of the Hardy-Weinberg assumption and the assumption of independence of genetic and environmental covariates. Additionally performance should be assessed using real genotype data.

The simulated data was of a gene-scan from a cohort study, so in this specific example, the focus was on using logistic regression to detect the association between disease and marker loci after adjusting for non-genetic covariates. The method is also applicable for genotype data on a few markers from a candidate gene. In this case, it can be used for detailed modeling of gene-by-environment and gene-by-gene interactions and to estimate odds ratios for disease-influencing haplotypes.

An important technical issue concerns the applicability of the logistic regression model to case-control data. We have made covariate distributional assumptions by using HWE as well as independence between the environmental and genetic covariates to formulate the likelihood. Although it is beyond the scope of this work, the application of this method to case-control studies should be examined. The key issue is that the classic justification for fitting a prospective logistic regression to retrospective (case-control) data (Prentice and Pyke 1979) assumed no modeling of the distribution of covariates.

As proof-of-principle, the algorithm was implemented for haplotypes of two loci only, but it is possible and potentially more useful to implement the algorithm so that there are more loci in a haplotype. For 3 diallelic loci, there are already 8 possible haplotypes, and 7 genetic variables in the regression model. Due to the biological process underlying haplotype creation, as well as sampling, it is likely that if the

number of haplotypes is high, not all haplotypes will be seen both in the population and in the sample. For our method to converge, these frequencies will have to be constrained to 0. As with using EM to estimate haplotype frequencies, our method is likely limited to a small number of diallelic loci. Building in more population genetic theory, as in Stephens *et al.* (2001), into the formulation of the likelihood could improve the number of loci that this method can handle.

Changes can be made to the regression model to allow for non-binary disease outcomes. The method of weights is appropriate for any generalized linear model, and has been extended to other regression models including those for survival data. See Horton and Laird (1999) for a general review of the method of weights and a discussion of its extensions to other types of data. Changes can also be made to the haplotype-effect model. By counting the number of copies of a particular haplotype, we are using a haplotype-dose model. This means that a person with more copies of the haplotype containing the disease mutation has a higher chance of developing the disease than a person with fewer copies of the haplotype. Specifically, an individual's odds of being affected changes by the same multiplicative factor for each additional copy of a haplotype. Recessive and dominant genetic risk models are also possible. For a recessive risk model, a person with two copies of a susceptibility allele is much more likely to develop the disease than if they have at least one allele that is not the disease-predisposing allele. Assuming that the disease-predisposing allele is associated with a putative "disease haplotype", this may be recoded as

$$x = \begin{cases} 1 & \text{homozygote for disease haplotype} \\ 0 & \text{heterozygote for disease haplotype or homozygote normal} \end{cases}.$$

For a dominant risk model, having one or two copies of the susceptibility haplotype leads to the same predisposition to disease. This can be coded as

$$x = \begin{cases} 1 & \text{has disease haplotype} \\ 0 & \text{does not have disease haplotype} \end{cases}.$$

Implementing either of these two different codings is not difficult and since these are the traditional inheritance models, they are probably worth trying on real datasets.

In conclusion, the method presented here for association mapping has potential advantages over current approaches. First, it allows for environmental covariates, possibly continuous, to be included and adjusted for. It also allows for detailed modeling of gene-environment and gene-gene interactions. The haplotype data is not imputed, and although the haplotype is not explicitly estimated, all data on an individual is used for determining the weight to be assigned to each haplotype configuration. If desired, the final weights can be used to infer haplotypes in a way that incorporates information of the disease status and environmental covariates. However, the importance of these advantages to the overall utility of this method can only be demonstrated when it is applied to more data.

Appendix A

The coalescent and the simulated data

SNP genotypes, non-genetic covariates and final disease status were simulated for 500 individuals in a cohort study of a late-onset complex disease. The randomly-paired chromosomes of individuals in the cohort were generated from a neutral coalescent with recombination (Hudson 1983) using a C-program kindly provided by R.R. Hudson. The coalescent (Kingman 1982) is a population-genetic model of ancestry for selectively neutral loci that provides a useful tool for simulating a sample of chromosomes under a variety of population genetic parameters. The coalescent with recombination allows for recombination events along the chromosome. Demographic parameters for the coalescent were chosen to reflect a population which had grown exponentially from an effective size of 5000 individuals 200 generations ago, to a current effective size of 2000000 individuals. Prior to 200 generations ago, the population was of constant effective size 5000. Effective sizes reflect the number of reproducing individuals under an idealized population-genetic model of reproduction and are reasonable for a modern European population. SNPs were simulated in a 5cM candidate region with 2000 segregating sites. Segregating loci were declared to be SNPs if the minor allele frequency in the sample was at least 20%. Of the 197 resulting SNPs, one from the middle of the 5cM region was chosen to be the primary susceptibility locus,

Affection Status	Gender	Age	BMI	SNP1	SNP2	SNP3	···	SNP195	SNP196
0	0	63.7	29.2	1,1	0,0	1,1	···	1,1	0,1
0	0	67.5	19	0,0	0,1	0,1	···	0,0	0,1
0	0	53.9	24.1	1,1	0,0	1,1	···	1,1	1,1
1	0	51.9	25.2	1,1	0,0	1,1	···	0,1	0,1
0	1	59.2	28.3	0,1	0,1	0,1	···	1,1	1,1

Table A.1: Subset of the simulated data set.

with susceptibility allele frequency 28% in the sample. Values of non-genetic covariates "gender", "age" and "body mass index" (BMI) were simulated independently of the genetic information and of each other. The distributions of these covariates mimicked those of gender, age and BMI in middle-aged and older North American populations. Affection status was generated randomly based on probabilities specified by the logistic regression model

$$logit(p) = -1 + 0.6 * \text{nsusc} + 0.3 * \text{age} + 0.6 * \text{BMI} + 0.3 * \text{pgene}.$$

Age and BMI were taken to be continuous covariates. Covariates have been centered by their sample mean and scaled by their sample standard deviation to provide an indication of the relative importance of effects. In the equation, p is the conditional probability of becoming affected during the study period given the covariate information, $logit(p)$ is the log-odds $\log[p/(1-p)]$ of becoming affected, "nsusc" is the standardized number of susceptibility alleles at the primary susceptibility locus and "pgene" is a standardized version of a uniformly distributed index between 0 and 10 indicating up to ten susceptibility alleles at other unlinked loci (polygenes) besides the major susceptibility locus. The sample prevalence of the disease was 28%. After generating disease status, information on the susceptibility locus was discarded from the dataset. Thus the data to be analyzed was single-locus genotypes for 196 SNPs, and information on non-genetic covariates gender, age, and BMI. An excerpt of the data is given in table A.1.

Bibliography

Akey J, Jin L and Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9: 291-300.

Ceppelini R, Siniscalco M and Smith CAB (1955) The estimation of gene frequencies in a random mating population. Annals of Human Genetics 20: 97-115.

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 2: 111-122.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ and Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229-232.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39(1):1-22.

Excoffier L and Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12: 921-927.

Fallin D, Cohen A, Essioux L, Chuymakov I, Blumenfeld M, Cohen D and Schork NJ (2001) Genetic Analysis of Case/Control Data Using Estimated Haplotype Frequencies: Application to APOE Locus Variation and Alzheimer's Disease. Genome Res 11: 143-151.

Fallin D and Schork NJ (2000) Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data. Am J Hum Genet 67: 947-959.

Hawley M and Kidd K (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86: 409-411.

Horton NJ and Laird NM (1999) Maximum likelihood analysis of generalized linear models with missing covariates. Statistical Methods in Medical Research 8: 37-50.

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theor Pop Biol 23:183-201

Ibrahim JG (1990) Incomplete data in generalized linear models. Journal of the American Statistical Association 85: 765-769.

Judson R and Stephens JC (2001) Notes form the SNP vs haplotype front. Pharmacogenics 2:7-10.

Kaplan N and Morris R (2001) Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. Genetic Epidemiology 20:432-457.

Kingman JFC (1982) The coalescent. Stochastic Processes 13:235-248

Long J, Williams R and Urbanek M (1995) An E-M Algorithm and Testing Strategy for Multiple-Locus Haplotypes. Am J Hum Genet 56: 799-810.

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B 44(2): 226-233.

Morris RW and Kaplan NL (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genetic Epidemiology 23: 221-233.

Niu T, Qin ZS, Xu X and Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70: 157-169.

Prentice RL and Pyke R (1979) Logistic disease incidence models and case-control studies. Biometrika 66: 403-411.

Schaid DJ, Rowland CM, Times DE, Jacobson RM and Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425-434.

Stephens M, Smith NJ and Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68: 978-979.

Zöllner S and von Haeseler A (2000) Coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. Am J Hum Genet 66: 615-628.