A COMPARISON OF TWO LOGISTIC REGRESSION APPROACHES FOR CASE-CONTROL DATA WITH MISSING HAPLOTYPES

by

Mercedeh Ghadessi
B.Sc. in Applied Mathematics, Sharif University of Technology, 1988

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

SIMON FRASER UNIVERSITY
Summer 2005

APPROVAL

Name: Mercedeh Ghadessi

Degree: Master of Science

Title of project: A Comparison of two Logistic Regression Approaches for Case-Control Data with Missing Haplotypes

Examining Committee: Dr. Richard Lockhart
Chair

____

Dr. Brad McNeney
Senior Supervisor
Simon Fraser University

____

Dr. Jinko Graham
Senior Supervisor
Simon Fraser University

____

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

Date Approved: August 2, 2005

Abstract

In a case-control study, subjects are selected according to disease status and their risk factors are determined retrospectively. When risk factors are fully observed for all subjects, maximum-likelihood inference of disease associations may be obtained by applying prospective logistic regression to case-control data as though it were collected prospectively. We investigate the statistical properties of prospective maximum-likelihood (PML) inference of disease associations with risk factors known as haplotypes when haplotype phase is not fully observed in some subjects. We motivate applying PML to case-control data and compare PML to an estimating equation (EE) approach developed specifically for such data. We conduct limited simulations of case-control data to investigate the bias of PML and EE, both in estimated haplotype risks and in their standard errors. PML performed well in the simulation configurations we considered. By contrast, EE gave anticonservative inference when there was marked haplotype ambiguity.

Acknowledgements

Many people have made the completion of this program possible. I can never include everyone or thank each of them sufficiently. First and foremost, I am indebted to Dr. Richard Lockhart, Chair of Graduate Studies and the Graduate Committee, who gave me hope to pursue my studies. Dr. Brad McNeney and Dr. Jinko Graham, my educational supervisors, guided me with acumen and helped me find clarity, reach for high standards, and keep focused throughout the program. I cannot thank them enough. My research was made possible by the Department of Statistics and Actuarial Science, and IRMACS at Simon Fraser University giving me access to special laboratory facilities. I also owe profound thanks to Dr. Charmaine Dean, Dr. Denise Daly, Ms Kelly Burkett and iCAPTURE Research Group at St. Paul's Hospital who gave me the opportunity to carry out research under their supervision. I wish to acknowledge the generosity of external experts who made this project possible: Dr. Richard Duncan of Emory University regarding the Chaplin software, and Dr. Richard J. Laws of Fred Hutchinson Cancer Research Center in Seattle pertaining to the Hplus software. My fellow graduate students were always helpful and kind. Many thanks to: Suman, Simon, Darby, Wendell, Pritam, Sandy, Darcy, Maria, Karey, Crystal, David, Zhijian, Linnea, Chufang, Stephen, Amy, Matthew, and Jason N. I am deeply grateful to Ji-Hyung Shin who assisted me in the completion of my thesis. The management of Triant Technologies, especially Dr. Viday Vankayala, allowed me maximum work flexibility. Their support will always be remembered. I give my special thanks to Dr. Tali Conine for her love and support especially in difficult times, and my uncle Massoud who believed in me. Most of all, and always, I am grateful to my parents, Manizheh and Manouchehr, and my brother, Mehdi, whose selfless love made everything possible for me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter reviews briefly the underlying epidemiologic and genetic concepts used in the thesis and provides a brief overview of the thesis.

1.1 Epidemiologic Background

Two commonly used epidemiologic study designs are *cohort* and *case-control* designs. In a cohort or *prospective* study, risk factors and baseline characteristics are measured on disease-free subjects who are then observed over a defined follow-up time to see if they become diseased. The sample is required to represent the underlying population, and thus the population disease rate may be estimated. By contrast, in a case-control or *retrospective* study, subjects are selected according to their disease status, and their risk factors are determined retrospectively. The sample does not represent the underlying population, though diseased and disease-free subjects are representative of their sub-groups within the population. If a disease is rare, it is difficult to observe an acceptable number of incident cases in a cohort study without increasing the follow-up time considerably. It may also be economically and/or practically impossible to measure covariates for all members of a cohort. Hence, for rare diseases, a case-control study design is considered more suitable [6].

Throughout this thesis, we consider *logistic regression* models for genetic associations with a disease. Logistic regression is a standard model for binary disease status. The *logistic* function $\ell(t) = (1 + \exp(-t))^{-1}$ takes on values between 0 and 1 and so is a natural candidate for modelling probabilities of a disease. The S-shape of the function is consistent with a model in which the effects of risk factors, or covariates, on an individual's disease risk is minimal until some threshold [13]. Logistic regression models an individual's probability of disease ($D = 1$) given his or her row-vector of covariates x as $\text{pr}(D = 1 \mid x) = \ell(\beta_0 + x\beta)$, where β_0 is an intercept term and β is a column vector of parameters describing the log-odds-ratio. The *odds* of disease for an individual with a given set of risk factors is the ratio of the probability of developing a disease to the probability of staying disease-free; that is $\text{odds}(x) = \text{pr}(D = 1 \mid x)/(1 - \text{pr}(D = 1 \mid x))$. The *odds ratio* or OR is the ratio of odds of developing a disease for two different individuals given their risk factors. If the risk factors for the first individual are denoted x_1 and those for the second individual are denoted x_2, $\text{OR}(x_1, x_2) = \text{odds}(x_1)/\text{odds}(x_2)$. The *relative risk* (RR), also called the *risk ratio*, is the ratio of the probabilities of disease for two distinct individuals given their risk factors: $\text{RR}(x_1, x_2) = \text{pr}(D = 1 \mid x_1)/\text{pr}(D = 1 \mid x_2)$. Though the RR may be viewed as having a more natural interpretation than the OR, an important advantage of the OR is that it can be estimated from case-control data, while disease risks, and therefore the RR, can not [2]. However, for a rare disease, $1 - \text{pr}(D = 1 \mid x) \approx 1$, in which case the OR and RR are effectively the same. Since case-control studies are most appropriate for rare diseases, the inability to estimate disease risks from such data does not usually pose a problem. For rare diseases, investigators typically identify risk factors on the basis of their relative risks and this information is preserved in a case-control study.

1.2 Genetic Background

DNA is comprised of four different building blocks or nucleotide bases denoted by A, T, C, G, and is packaged into *chromosomes*, each of which has a linear DNA sequence. An organism's total DNA content is called its *genome*. All human cells except germ cells and red blood cells are *diploid* and thus carry pairs of chromosomes. These pairs are called *homologous chromosomes*. One member of each pair is inherited from the mother, while the other is inherited from the father. There are 22 pairs of *autosomal chromosomes* and a pair of *sex chromosomes* in diploid cells. *Haploid* cells such as germ cells or *gametes* (sperm and egg cells) have only a single copy of each chromosome.

The location of a gene or another identifiable point on the genome is called a *locus*. The DNA at a locus may come in a variety of forms or *alleles*. In diploid cells, an individual has two (possibly identical) alleles at each locus. The unordered pair of alleles (with respect to inheritance) that an individual has is the individual's *genotype* at this locus. Individuals who are *heterozygous* have a genotype in which the alleles are different. Individuals who are *homozygous* have a genotype with the same alleles. If the locus is one relating to a functional gene, the resulting potentially observable characteristic of the individual is the *phenotype* or *trait*. *Penetrance* is the probability of expressing a phenotype such as a disease, given a specific genotype.

Hardy-Weinberg equilibrium (HWE) defines the relationship between the frequency of genes in a population and the frequency of genotypes in individuals. Once the allele frequencies are stable in a population, the expected genotype frequencies in the offspring generation will be in equilibrium after one generation of random mating. For example, suppose A and a are the two alleles at a locus, and that $P(A) = p$ and $P(a) = 1 - p = q$ are their frequencies. Then, after one generation of random mating, the genotype frequencies should be in *Hardy-Weinberg proportions* (HWP) with $P(AA) = p^2$, $P(Aa) = 2pq$, and $P(aa) = q^2$. Numerous factors affect the frequency of alleles and genotypes at any particular locus and thus will perturb HWE. These factors include mutation, selection, migration and nonrandom mating [12].

Exchange of genetic material between homologous chromosomes (*cross-over*) occurs during *meiosis*, the process of gamete formation. Meiosis leads to four different *haploid* gametes (sperm or egg cells). If two loci are relatively far from each other on a chromosome, many crossing-over events are expected to occur between them in a meiosis. In this case, the expectation is that half of the gametes will be *recombinant*, or have alleles of different parental origin at the two loci. Hence, the *recombination frequency* between the two loci, or the ratio of the expected number of recombinants to the total number of observed gametes, is $r = 0.5$ and the loci are *unlinked*. The closer the two loci are, the fewer gametes are expected to be recombinant (so that $r < 0.5$). These loci are said to be *genetically linked*. Linkage analysis tracks the transmission of genetic material in gametes and is based on the principal that nearby loci on a chromosome tend to *co-segregate* to the next generation. For completely linked loci ($r = 0$), the parental origin of the allele at one locus on a gamete completely determines the parental origin of the allele at the other locus.

DNA alterations that involve a single base pair and are observed in at least 1% of the population are called *single nucleotide polymorphisms* (SNPs). SNPs are abundantly dispersed throughout the genome and form part of the natural genetic variation. They are diallelic and are therefore less informative on their own for studying linkage than loci having more alleles. In spite of this, SNPs are increasingly used in both genetic linkage and association studies due to their high frequency throughout the genome, low mutation rates, and easily automated detection. The high density of SNPs throughout the genome allows investigators to combine nearby SNPs, creating a potentially more informative pseudo-locus with variant forms called haplotypes. A *haplotype* is a sequence of alleles at several loci on the same chromosome. Each individual has two haplotypes, one on the paternal chromosome and the other on the maternal chromosome. If an individual is homozygous at all constituent loci, then his/her *dosage* of the corresponding haplotype is two. If an individual is heterozygous at one or more constituent loci, then his/her dosage of each of the corresponding haplotypes is one. Experimental determination or *phasing* of an individual's haplotypes can be technically challenging and expensive. The haplotype phase of an individual may be inferred from the individual's genotype data at the constituent loci if family members are also genotyped. However, genotyping family members in large-scale association studies is impractical.

1.3 Overview of Thesis

In this thesis, we consider the problem of assessing haplotypic risk factors for a disease using data from a case-control study in which haplotype phase information is missing for some individuals. Two approaches are compared: PML, prospective maximum likelihood developed for cohort data with uncertain haplotype phase (Burkett et al. [4]) and EE, an estimating equation approach developed specifically for case-control data (Zhao et al. [20]). Chapter 2 reviews both approaches, provides a theoretical justification for applying PML to case-control data, and explores the similarities and differences between the two approaches. Chapter 3 describes the results of a limited simulation study comparing statistical properties of the two approaches when there are no non-genetic covariates. The impact of haplotype ambiguity on bias in risk estimators and their standard errors is explored. Chapter 4 summarizes the main conclusions and discusses directions for future research.

Chapter 2

Methods

Burkett et al. [4] and Lake et al. [14] independently developed an EM algorithm for PML inference of haplotype risks from prospective or cross-sectional data on unphased genotypes, non-genetic attributes, and trait values in generalized linear models (GLMs) of trait penetrance. Similar methodology was developed by Stram et al. [17] for the special case of binary disease traits, where the GLM is a logistic regression. In this chapter, we investigate theoretically the validity of PML applied to case-control data and compare it to the EE approach of Zhao et al. [20] developed specifically for case-control data, after reviewing some general concepts.

2.1 EM Algorithm for Missing Data Problems

The EM algorithm is phrased in terms of the observed, incomplete data Y and the unobserved, complete data Z from a statistical experiment. The complete data Z is usually viewed as the observed data augmented in some way. The algorithm is useful for maximizing an observed-data likelihood function $L_o(\theta; y)$ that may be difficult to maximize explicitly, but which has a corresponding complete-data likelihood $L_c(\theta; z)$ that is easily maximized [18]. In our context, we assume non-genetic covariates are fully observed so that the observed data Y will be a random vector of the disease status, unphased single-locus genotypes and non-genetic covariates on all subjects. Let $\mathrm{pr}(y; \theta)$ be the pdf or pmf of Y, and $l_o(\theta; y) = \log \mathrm{pr}(y; \theta)$ be the observed-data log-likelihood for parameters θ. The complete data Z are the unphased single-locus genotypes augmented by the *haplogenotypes* or underlying haplotype pairs on all individuals. Let $\text{pr}(z;\theta)$ be the pdf or pmf of Z, and $l_c(\theta;z) = \log \text{pr}(z;\theta)$ be the complete-data log-likelihood. The EM algorithm involves repeatedly maximizing the conditional expected value of the complete-data log-likelihood given the observed data. Specifically, starting from an initial guess $\theta^{(0)}$ of θ, iterate through the following expectation (E) and maximization (M) steps for $t = 0, 1, \ldots$:

E-step: Calculate $Q(\theta \mid \theta^{(t)}) = E_{\theta^{(t)}}\left[l_c(\theta;Z) \mid Y\right]$, where $\theta^{(t)}$ is the estimate of θ at iteration t, and $E_{\theta^{(t)}}$ denotes expectation under parameter value $\theta^{(t)}$.

M-step: Maximize $Q(\theta \mid \theta^{(t)})$ over θ to obtain $\theta^{(t+1)}$.

With each iteration of the EM algorithm, the log-likelihood either increases or stays the same [7]; i.e. $l_o(\theta^{(t+1)};y) \geq l_o(\theta^{(t)};y)$. This iteration stops when $\theta^{(t)}$ converges to a local maximum of the likelihood function. For example, stop when $l_o(\theta^{(t+1)};y) - l_o(\theta^{(t)};y)$ is small (preferred) or when $||\theta^{(t+1)} - \theta^{(t)}||$ is small.

2.2 Direct Maximization for Missing Data

Direct maximization of the log-likelihood by Newton-Raphson requires calculation of the score and hessian matrix. Formulas for the observed-data score and hessian in missing data problems were derived by Louis [15], and can be summarized as follows. Keep the same notation as before and, in addition, let

$$\dot{l}_o(\theta;y) = \frac{\partial}{\partial\theta} l_o(\theta;y) \quad \text{and} \quad \dot{l}_c(\theta;z) = \frac{\partial}{\partial\theta} l_c(\theta;z)$$

denote the observed- and complete-data scores, respectively. Let

$$I_o(\theta;y) = -\frac{\partial^2}{\partial\theta^2} l_o(\theta;y) \quad \text{and} \quad I_c(\theta;z) = -\frac{\partial^2}{\partial\theta^2} l_c(\theta;z)$$

denote minus the observed- and complete-data hessians, respectively. Here, I is intended to indicate the "observed information", not its expectation (the Fisher information). However, the term "observed" is also used to distinguish observed- from complete-data likelihoods which could lead to confusion. For brevity and to avoid confusion, refer to I_o and I_c as the observed-data information and complete-data information, respectively, rather than as the observed-data observed information and the complete-data observed information. Louis [15] showed that:

$$\dot{l}_o(\theta; y) = E_\theta\left[\dot{l}_c(\theta; Z) \mid Y = y\right], \tag{2.1}$$

and that

$$I_o(\theta; y) = E_\theta\left[I_c(\theta; Z) \mid Y = y\right] - V_\theta\left[\dot{l}_c(\theta; Z) \mid Y = y\right], \tag{2.2}$$

where V_θ denotes variance under parameter value θ. Equations (2.1) and (2.2) are known as *Louis' equations.* If the conditional expectations and variances in Louis' equations can be computed for given values of θ, then the likelihood may be maximized directly by Newton-Raphson in the usual way. That is, starting with an initial value $\theta^{(0)}$, perform a series of Newton-Raphson updates $\theta^{(t+1)} = \theta^{(t)} - I_o(\theta^{(t)}; y)^{-1}\dot{l}_o(\theta^{(t)}; y)$ until convergence.

2.3 Applications to Haplotype Risk Estimation

Both direct maximization using Louis' formulas and the EM algorithm require calculation of conditional expectations of functions of the complete data Z given the observed data Y. For direct maximization, the appropriate functions of Z are the complete-data scores and hessians. For maximization via the EM algorithm, the appropriate function is the complete-data log-likelihood. Hence, both approaches require expressions for the conditional distribution of Z given Y. Recall that the complete data are the observed data on disease status, single-locus genotypes and non-genetic covariates, augmented by the haplogenotypes. The complete data need only include as much information as required for the penetrance model, and so can be less detailed than the haplogenotypes. For example, if the penetrance model is comprised of a single multiplicative effect for a particular risk haplotype, the complete data need only indicate the number of copies of that haplotype. However, penetrance models such as the saturated haplogenotypes model or a multiplicative model that includes terms for all haplotypes (except a baseline haplotype) require the complete haplogenotype information. In addition, specifying haplogenotypes as the complete data leads to estimates of haplogenotypes frequencies, which are often of interest on their own. Hence we adopt the convention that the complete data include the haplogenotypes.

There are a finite number of haplogenotypes consistent with the observed single-locus genotypes. Hence, there are a finite number of probabilities that comprise the conditional distribution of the complete data given the observed data. Thus, conditional expectations of functions of the complete data given the observed data are weighted averages over the finite number of complete data values consistent with the observed data, where the "weights" are the conditional probabilities. Therefore, both the direct and EM approaches require calculation of *weights* for likelihood maximization. From the general description of EM, it can be seen that the weights are updated in the E-step, but then treated as fixed in the M-step when the weighted complete-data log-likelihood is maximized to update the parameter estimates. By contrast, in Newton-Raphson, the complete-data log-likelihood and the weights are both considered as functions of the current parameter values in a single update step.

2.4 PML and Case-Control Data

2.4.1 Previous Theoretical Results

Carroll et al. [5], hereafter referred to as CWW, state that applying a prospective approach to retrospective data typically gives unbiased parameter estimates and conservative or correct standard errors. However, a key assumption of CWW is that the estimating equations the estimators solve are *retrospectively unbiased* in the sense that their expected values under the retrospective sampling scheme are zero. CWW require mean zero for *all* sample sizes, a stronger condition than having mean zero as the sample size tends to infinity. Estimating equations with the weaker property of having mean zero as the sample size tends to infinity are *asymptotically unbiased.* Zhao et al. [20] showed that the estimating equations for their approach are asymptotically unbiased, but not retrospectively unbiased. We suspect that the PML estimating equations are also retrospectively biased because the weights, or conditional haplogenotype probabilities given the observed data on an individual, are only approximate for case-control data. If the weights were instead correctly specified, the resulting estimating equations would be retrospectively unbiased because they would correspond to the retrospective score functions. Score functions have expectation zero regardless of the sample size. A detailed investigation of retrospective unbiasedness for the PML estimating equations is beyond the scope of this project.

2.4.2 Previous Empirical Results

Stram et al. [17] applied PML to ated case-control data and concluded that bias in estimated haplotype risks increases with decreasing ability of single-locus genotypes to predict the underlying haplotypes (haplogenotypes). However, the design of their simulation study raises some questions. In particular, they appear to estimate risks for haplotypes of a subset of three of six possible loci at which single-locus genotypes are available. With single-locus genotypes at *only three loci*, the effect of a *six-locus risk haplotype* can be confounded with the effect of a non-risk haplotype. For example, based on the first, third and fifth loci, a six-locus risk haplotype 0-0-0-0-0-0 would be indistinguishable from non-risk haplotypes 0-1-0-1-0-1, 0-1-0-1-0-0, 0-1-0-0-0-1, 0-1-0-0-0-0, 0-0-0-1-0-1, 0-0-0-1-0-0 and 0-0-0-0-0-1. In this situation, we would expect the estimated effect of the three-locus risk haplotype 0-0-0 based on the first, third and fifth loci to be a weighted average of the odds-ratio of the six-locus risk haplotype 0-0-0-0-0-0 and the odds-ratios of the confounded non-risk haplotypes 0-1-0-1-0-1, 0-1-0-1-0-0, 0-1-0-0-0-1, 0-1-0-0-0-0, 0-0-0-1-0-1, 0-0-0-1-0-0 and 0-0-0-0-0-1. The estimated effect for the three-locus risk haplotype 0-0-0 would therefore be attenuated relative to that of the six-locus risk haplotype 0-0-0-0-0-0. In their simulations, the particular six-locus risk haplotypes for which they claimed a bias toward zero in estimated effect were in fact those haplotypes confounded with other non-risk haplotypes of non-trivial frequency when a reduced set of three loci were examined. Hence, the bias they observed may be due to the more fundamental problem of confounding rather than to the application of a prospective approach to retrospective data. In this thesis, we will examine both theoretically and through simulations the potential bias in haplotype risk estimators and their standard errors arising from applying PML to case-control data, in a context where confounding is not an issue.

In summary, doubts about Stram et al.'s empirical evidence motivated us to reconsider the bias arising from applying PML to case-control data. In the following sections, we investigate in more detail the theoretical justification for applying PML. Specifically, we examine a) whether prospective weights are sensible in the E-step, and b) whether maximizing the prospective weighted likelihood is reasonable in the M-step. To gain additional insight into PML versus EE, we derive the PML estimating equations for retrospective data and compare them to those of the EE approach.

2.5 Review of PML

Let D_i denote disease status with value 1 for diseased and 0 for non-diseased, G_i denote single-locus genotypes and A_i denote non-genetic attributes on the ith subject. These variables comprise the observed data on the ith subject. Suppose there are n independent subjects. Let H_i denote the haplogenotype of the ith subject. The observed data (D_i, G_i, A_i) augmented by H_i comprise the complete data on the ith subject. Since G_i is determined by H_i, the complete data may be written (D_i, H_i, A_i). Let D denote a vector of disease status indicators on all n subjects. Similarly, let G, A and H denote, respectively, the collections of single-locus genotypes, non-genetic variables and haplogenotypes on all n subjects. Let lower-case letters denote observed values. For example, d_i is the observed disease status on the ith subject and d denotes a vector of observed disease status indicators on all subjects. Let $X(H_i, A_i)$ be the covariate row-vector for the ith subject in the logistic model of penetrance:

$$\text{logit}P(D_i = 1 \mid H_i, A_i) = \beta_0 + X(H_i, A_i)\beta,$$

where β_0 is an intercept term and β is a column-vector of odds-ratio parameters. For cohort data it is cumbersome to separate β_0 from the odds ratio parameters, because β_0 can also be estimated. However, we do so now because such a separation will be necessary when considering case-control data, where β_0 can not be estimated. Let γ denote the parameters that describe the joint distribution of H and A. For notational convenience, suppose γ is a column-vector. The collection of all parameters in this problem is $\theta = (\beta_0, \beta^T, \gamma^T)^T$.

2.5.1 E-step of the EM Algorithm

The conditional expectation of the complete-data log-likelihood given the observed data and the value of the parameters $\theta^{(t)}$ at iteration t is

$$Q(\theta \mid \theta^{(t)}) = E_{\theta^{(t)}}\left[l_c(\theta; D, H, A) \mid D = d, G = g, A = a\right] \qquad \text{where,}$$

$$\begin{aligned} l_c(\theta; D, H, A) &= \log \mathrm{pr}(D, H, A; \theta) \qquad \text{(independence of subjects)} \\ &= \sum_{i=1}^{n} \log \mathrm{pr}(D_i, H_i, A_i; \theta) \equiv \sum_{i=1}^{n} l_{ci}(\theta; D_i, H_i, A_i). \qquad \text{Hence,} \end{aligned}$$

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= \sum_{i=1}^{n} E_{\theta^{(t)}}\left[l_{ci}(\theta; D_i, H_i, A_i) \mid D = d, G = g, A = a\right] \\ &= \sum_{i=1}^{n} E_{\theta^{(t)}}\left[l_{ci}(\theta; D_i, H_i, A_i) \mid D_i = d_i, G_i = g_i, A_i = a_i\right] \qquad (2.3) \end{aligned}$$

The conditional expectations in equation (2.3) may now be written as weighted sums of complete-data log-likelihoods $l_{ci}(\theta)$ over complete-data values consistent with the observed data for an individual. Let $\mathcal{S}_i = \left\{h_i^k;\ k = 1, \ldots, K_i\right\}$ denote the set of haplogenotypes consistent with g_i; $i = 1, \ldots, n$. For $h_i^k \in \mathcal{S}_i$, let

$$\begin{aligned} w_{ik}(\theta^{(t)}) &= \mathrm{pr}(D_i = d_i, H_i = h_i^k, A_i = a_i \mid D_i = d_i, G_i = g_i, A_i = a_i\ ;\ \theta^{(t)}) \\ &= \mathrm{pr}(H_i = h_i^k \mid D_i = d_i, G_i = g_i, A_i = a_i\ ;\ \theta^{(t)}). \qquad (2.4) \end{aligned}$$

Replacing the conditional expectation in equation (2.3) with the appropriate weighted sum gives: $Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K_i} w_{ik}(\theta^{(t)}) l_{ci}(\theta; d_i, h_i^k, a_i)$, where $l_{ci}(\theta; d_i, h_i^k, a_i) = \log \mathrm{pr}(D_i = d_i, H_i = h_i^k, A_i = a_i; \theta)$. Recall that $\theta = (\beta_0, \beta^T, \gamma^T)^T$ where β_0 and β are from the penetrance model and γ is the parameter vector that describes the joint distribution of H and A. Write the weights in terms of θ as:

$$\begin{aligned}
w_{ik}(\theta^{(t)}) &= \mathrm{pr}(H_i = h_i^k \mid D_i = d_i, G_i = g_i, A_i = a_i \,; \theta^{(t)}) \\
&= \frac{\mathrm{pr}(H_i = h_i^k, D_i = d_i, G_i = g_i, A_i = a_i \,; \theta^{(t)})}{\mathrm{pr}(D_i = d_i, G_i = g_i, A_i = a_i \,; \theta^{(t)})} \\
&= \frac{\mathrm{pr}(H_i = h_i^k, D_i = d_i, A_i = a_i \,; \theta^{(t)})}{\mathrm{pr}(D_i = d_i, G_i = g_i, A_i = a_i \,; \theta^{(t)})} \\
&= \frac{\mathrm{pr}(H_i = h_i^k, D_i = d_i, A_i = a_i \,; \theta^{(t)})}{\sum_{l=1}^{K_i} \mathrm{pr}(H_i = h_i^l, D_i = d_i, A_i = a_i \,; \theta^{(t)})} \\
&= \frac{\mathrm{pr}(D_i = d_i \mid H_i = h_i^k, A_i = a_i \,; \beta_0^{(t)}, \beta^{(t)})\mathrm{pr}(H_i = h_i^k, A_i = a_i \,; \gamma^{(t)})}{\sum_{l=1}^{K_i} \mathrm{pr}(D_i = d_i \mid H_i = h_i^l, A_i = a_i \,; \beta_0^{(t)}, \beta^{(t)})\mathrm{pr}(H_i = h_i^l, A_i = a_i \,; \gamma^{(t)})}. \quad (2.5)
\end{aligned}$$

Calculation of disease probabilities $\mathrm{pr}(D_i = d_i \mid H_i = h_i^k A_i = a_i \,; \beta_0^{(t)}, \beta^{(t)})$ is straightforwad using the penetrance model and the estimates $\beta_0^{(t)}$ and $\beta^{(t)}$ from fitting a logistic regression model in the M-step at iteration t.

Calculation of joint probabilities $\mathrm{pr}(H_i = h_i^k A_i = a_i \,; \gamma^{(t)})$ is not as straightforward. As discussed by Horton and Laird [11] when the covariate sample space is large (e.g. when the non-genetic attributes are continuous), there may be very little information in the data to infer γ and hence to calculate such probabilities. One possible solution to the difficulties posed by continuous non-genetic attributes is to impose independence of genetic and non-genetic attributes where reasonable [3, 4, 14]. Note that such an assumption may not be reasonable if non-genetic attributes are ethnicity or body-mass index, for example. Under this independence assumption, let γ_h denote the parameters in the marginal distribution of H and γ_a denote the parameters in the marginal distribution of A. Then the formula for the weights simplifies to:

$$\begin{aligned}
w_{ik}(\theta^{(t)}) &= \frac{\mathrm{pr}(D_i = d_i \mid H_i = h_i^k A_i = a_i \,; \beta_0^{(t)}, \beta^{(t)})\mathrm{pr}(H_i = h_i^k \,; \gamma_h^{(t)})\mathrm{pr}(A_i = a_i \,; \gamma_a^{(t)})}{\sum_{l=1}^{K_i} \mathrm{pr}(D_i = d_i \mid H_i = h_i^l, A_i = a_i \,; \beta_0^{(t)}, \beta^{(t)})\mathrm{pr}(H_i = h_i^l \,; \gamma_h^{(t)})\mathrm{pr}(A_i = a_i \,; \gamma_a^{(t)})} \\
&= \frac{\mathrm{pr}(D_i = d_i \mid H_i = h_i^k A_i = a_i \,; \beta_0^{(t)}, \beta^{(t)})\mathrm{pr}(H_i = h_i^k \,; \gamma_h^{(t)})}{\sum_{l=1}^{K_i} \mathrm{pr}(D_i = d_i \mid H_i = h_i^l, A_i = a_i \,; \beta_0^{(t)}, \beta^{(t)})\mathrm{pr}(H_i = h_i^l \,; \gamma_h^{(t)})}
\end{aligned}$$

so that estimation of γ_a is not required. Henceforth, assume independence of H and A. The distribution of haplogenotypes, specified by the vector of haplogenotype frequencies γ_h, remains in the calculation of the weights. However, this distribution can not be estimated from single-locus genotype data, as illustrated by the following simple example involving two-locus haplotypes. Let the alleles at both loci be denoted 0 and 1 and let a haplotype with allele a_1 at the first locus and a_2 at the second locus be denoted a_1a_2. Individuals who are doubly heterozygous at the constituent SNPs can have either 00/11 or 01/10 as their haplogenotype. The proportion of subjects who are either 00/11 or 01/10 can be estimated by the proportion of double heterozygotes in the sample. However, neither of these two haplogenotypes are ever observed on their own unambiguously. Hence, there is no information in the data about the relative proportion of 00/11 versus 01/10 haplogenotypes. The frequencies of the 00/11 and 01/10 haplogenotypes are therefore not identifiable from single-locus genotype data. One possible solution to this identifiability problem [3, 4, 14] is to impose HWP for haplogenotype probabilities and model them in terms of haplotype frequencies. Under HWP, haplotypegenotype frequencies can be estimated from data on single-locus genotypes, based on information provided by haplotypes within haplogenotypes that are observed unambiguously. We assume HWP throughout and redefine γ_h to be the vector of haplotype frequencies.

2.5.2 M-step of the EM algorithm

The function of θ to be maximized is:

$Q(\theta \mid \theta^{(t)})$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{k=1}^{K_i} w_{ik}(\theta^{(t)}) l_{ci}(\theta; d_i, h_i^k, a_i) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K_i} w_{ik}(\theta^{(t)}) \log \mathrm{pr}(D_i = d_i, H_i = h_i^k, A_i = a_i; \theta) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K_i} w_{ik}(\theta^{(t)}) \log \Big[\mathrm{pr}(D_i = d_i \mid H_i = h_i^k, A_i = a_i; \beta_0, \beta)\mathrm{pr}(H_i = h_i^k; \gamma_h)\mathrm{pr}(A_i = a_i; \gamma_a)\Big] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K_i} w_{ik}(\theta^{(t)}) \log \mathrm{pr}(D_i = d_i \mid H_i = h_i^k, A_i = a_i; \beta_0, \beta) + \\
&\quad \sum_{i=1}^{n}\sum_{k=1}^{K_i} w_{ik}(\theta^{(t)}) \log \mathrm{pr}(H_i = h_i^k; \gamma_h) + \sum_{i=1}^{n}\sum_{k=1}^{K_i} w_{ik}(\theta^{(t)}) \log \mathrm{pr}(A_i = a_i; \gamma_a) \\
&\equiv Q(\beta_0, \beta \mid \theta^{(t)}) + Q(\gamma_h \mid \theta^{(t)}) + Q(\gamma_a \mid \theta^{(t)})
\end{aligned}
$$

Each of the three expressions on the last line of the above equation may be maximized separately. The first expression, $Q(\beta_0, \beta \mid \theta^{(t)})$, is a weighted log-likelihood from a logistic regression model of penetrance and can be maximized using standard statistical software. The second expression, $Q(\gamma_h \mid \theta^{(t)})$, is a weighted multinomial log-likelihood and so its maximization is also straightforward. The last expression, $Q(\gamma_a \mid \theta^{(t)})$, involves only the nuisance parameter γ_a which is not required for the weights and hence need not be estimated. The two assumptions made for PML inference of haplotype risks are: 1) independence of haplotypes and non-genetic attributes in the population, and 2) HWP of haplogenotypes in the population. Simulation studies suggest that PML inference is robust to moderate departures from the independence assumption [4] and to moderate departures from HWP [14].

2.6 Justification of PML for Case-Control Data

We shall argue that naive application of PML to case-control data entails maximization of a weighted retrospective log-likelihood in the M-step, with approximate weights obtained in the E-step. Hence any bias in risk estimators arises due to the weight approximation in the E-step.

In a case-control study, exposures are sampled conditional on disease status. In our context this means the observed data G and A are sampled conditional on D. The change in the sampling scheme requires slight changes in the notation used previously. Let D_{ij}, G_{ij}, H_{ij} and A_{ij} denote, respectively, the disease status, single-locus genotypes, haplogenotypes and non-genetic attribute for the jth subject in the ith disease group; $i = 0, 1,\ j = 1, \ldots, n_i$, where n_i is the number of subjects in disease group i. By definition $D_{ij} = i$. Let D, G, H and A denote the collections of disease status, single-locus genotypes, haplogenotypes and non-genetic attributes for all subjects, as before.

For cohort (or cross-sectional) data, the parameters in the model were $\theta = (\beta_0, \beta^T, \gamma^T)^T$, where β_0 is an intercept and β is a vector of odds-ratio parameters in the logistic penetrance model, and γ is a vector of parameters that describes the joint distribution of H and A. The case-control likelihood may be parametrized in terms of $\text{pr}(H_{0j}, A_{0j} \mid D_{0j} = 0)$ and β, where $\text{pr}(H_{0j}, A_{0j} \mid D_{0j} = 0)$ depends on θ. Accordingly, define the parameters in the case-control likelihood to be $\vartheta = [\beta, \text{pr}(H_{0j}, A_{0j} \mid D_{0j} = 0)]$ [16]. Under case-control sampling, write $l_o(\vartheta; g, a) = \log \text{pr}(G = g, A = a \mid D = d; \vartheta)$ for the observed-data log-likelihood and $l_c(\vartheta; h, a) = \log \text{pr}(H = h, A = a \mid D = d; \vartheta)$ for the complete-data log-likelihood.

2.6.1 A Variant Sampling Scheme

Several calculations in this section rely on a variant sampling scheme (VSS) used implicitly throughout the arguments of Prentice and Pyke [16]. This is a two-stage sampling design with random sampling of disease status and covariates, but in which the total number of subjects is fixed to n. The first step is Bernoulli sampling of disease status, with probability n_1/n of sampling a case and n_0/n of sampling a control. Thus, for a study with n subjects (i.e. n Bernoulli trials), the expected number of cases and controls sampled are n_1 and $n - n_1 = n_0$, respectively.

In the second step, covariates are sampled (observed) from the appropriate conditional distributions of covariates given disease status. The conditional distributions in this second step are the same conditional distributions as in the true case-control sampling scheme. Under VSS, disease-status/covariate pairs are sampled jointly. However, given disease status, covariates are sampled independently. In contrast, under the basic stratified sampling of unmatched case-control studies, disease status is fixed rather than random and, conditional on disease status, covariates are sampled independently. Let pr_v denote probability densities or mass functions, as appropriate, under VSS and let pr denote probabily densities or mass functions under the true case-control sampling. From the description of VSS, $\text{pr}(H = h, A = a \mid D = d; \theta) = \text{pr}_v(H = h, A = a \mid D = d; \vartheta)$. Define the VSS hypothetical population to be a population with case frequency n_1/n and control frequency n_0/n. Then, VSS corresponds to random sampling from the VSS hypothetical population.

2.6.2 E-step

The conditional expectation of the complete-data log-likelihood given the observed data, disease status and the value of the parameters $\vartheta^{(t)}$ at iteration t is now

$Q(\vartheta \mid \vartheta^{(t)}) = E_{\vartheta^{(t)}}\left[l_c(\vartheta; H, A) \mid G = g, A = a, D = d\right]$, where:

$$\begin{aligned} l_c(\vartheta; H, A) &= \log \mathrm{pr}(H, A \mid D = d; \vartheta) \\ &= \sum_{i=0}^{1}\sum_{j=1}^{n_i} \log \mathrm{pr}(H_{ij}, A_{ij} \mid D_{ij} = i; \vartheta) \quad \text{(independence within disease groups)} \\ &\equiv \sum_{i=0}^{1}\sum_{j=1}^{n_i} l_{cij}(\vartheta; H_{ij}, A_{ij}). \end{aligned}$$

$$\begin{aligned} \text{Hence } Q(\vartheta \mid \vartheta^{(t)}) &= \sum_{i=0}^{1}\sum_{j=1}^{n_i} E_{\vartheta^{(t)}}\left[l_{cij}(\vartheta; H_{ij}, A_{ij}) \mid G = g, A = a, D = d\right] \\ &= \sum_{i=0}^{1}\sum_{j=1}^{n_i} E_{\vartheta^{(t)}}\left[l_{cij}(\vartheta; H_{ij}, A_{ij}) \mid G_{ij} = g_{ij}, A_{ij} = a_{ij}, D_{ij} = i\right] \qquad (2.6) \end{aligned}$$

Let $\mathcal{S}_{ij} = \left\{h_{ij}^k;\ k = 1, \ldots, K_{ij}\right\}$ denote the set of haplogenotypes consistent with the single-locus genotypes g_{ij}; $i = 0, 1,\ j = 1, \ldots, n_i$. For $h_{ij}^k \in \mathcal{S}_{ij}$, let

$$\begin{aligned} w_{ijk}(\vartheta^{(t)}) &= \mathrm{pr}(H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \mid G_{ij} = g_{ij}, A_{ij} = a_{ij}, D_{ij} = i;\ \vartheta^{(t)}) \\ &= \mathrm{pr}(H_{ij} = h_{ij}^k \mid G_{ij} = g_{ij}, A_{ij} = a_{ij}, D_{ij} = i;\ \vartheta^{(t)}) \qquad (2.7) \\ &= \frac{\mathrm{pr}(H_{ij} = h_{ij}^k, G_{ij} = g_{ij}, A_{ij} = a_{ij} \mid D_{ij} = i\ ; \vartheta^{(t)})}{\mathrm{pr}(G_{ij} = g_{ij}, A_{ij} = a_{ij} \mid D_{ij} = i\ ; \vartheta^{(t)})} \\ &= \frac{\mathrm{pr}(H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \mid D_{ij} = i\ ; \vartheta^{(t)})}{\mathrm{pr}(G_{ij} = g_{ij}, A_{ij} = a_{ij} \mid D_{ij} = i\ ; \vartheta^{(t)})} \\ &= \frac{\mathrm{pr}(H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \mid D_{ij} = i\ ; \vartheta^{(t)})}{\sum_{l=1}^{K_{ij}} \mathrm{pr}(H_{ij} = h_{ij}^l, A_{ij} = a_{ij} \mid D_{ij} = i\ ; \vartheta^{(t)})}. \qquad (2.8) \end{aligned}$$

Equation (2.7) establishes that case-control weights describe the same conditional probabilities as the cohort weights in equation (2.4). Equation (2.6) is then $Q(\vartheta \mid \vartheta^{(t)}) = \sum_{i=0}^{1}\sum_{j=1}^{n_i}\sum_{k=1}^{K_{ij}} w_{ijk}(\vartheta^{(t)}) l_{cij}(\vartheta; h_{ij}^k, a_{ij})$, where $l_{cij}(\vartheta; h_{ij}^k, a_{ij}) = \log \mathrm{pr}(H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \mid D_{ij} = i; \vartheta)$.

At iteration t, writing the weights in terms of quantities that are estimated in the M-step requires consideration of the variant sampling scheme (VSS). Recall that

$$\mathrm{pr}(H_{ij} = h_{ij}, A_{ij} = a_{ij} \mid D_{ij} = i\ ; \vartheta^{(t)}) = \mathrm{pr}_v(H_{ij} = h_{ij}, A_{ij} = a_{ij} \mid D_{ij} = i\ ; \vartheta^{(t)}).$$

Hence $\text{pr}(H_{ij} = h_{ij}, A_{ij} = a_{ij} \mid D_{ij} = i \,; \vartheta^{(t)})$

$$= \frac{\text{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}, A_{ij} = a_{ij} \,; \beta_{v0}^{(t)}, \beta^{(t)})\text{pr}_v(H_{ij} = h_{ij}, A_{ij} = a_{ij} \,; \gamma_v^{(t)})}{\text{pr}_v(D_{ij} = i)}, \tag{2.9}$$

where $\text{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}, A_{ij} = a_{ij} \,; \beta_{v0}^{(t)}, \beta^{(t)})$ is a logistic regression model with the same odds-ratio parameters β as the logistic model for a population sample, but with a different intercept β_{v0}; γ_v parameterizes the joint distribution of H and A under VSS; and $\text{pr}_v(D_{ij} = i)$ is the disease risk under VSS, which is n_i/n by definition [16]. The new parameter $\vartheta_v \equiv (\beta, \gamma_v)$ reparametrizes $\vartheta = [\beta, \text{pr}(H_{0j}, A_{0j} \mid D_{0j} = 0)]$, while β_{v0} is a function of β and γ_v [16]. Substituting equation (2.9) into the expression (2.8) for the weights gives:

$$w_{ijk}(\vartheta_v^{(t)}) = \frac{\text{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}^k, A_{ij} = a_{ij}; \beta_{v0}^{(t)}, \beta^{(t)})\text{pr}_v(H_{ij} = h_{ij}^k, A_{ij} = a_{ij}; \gamma_v^{(t)})}{\sum_{l=1}^{K_{ij}} \text{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}^l, A_{ij} = a_{ij}; \beta_{v0}^{(t)}, \beta^{(t)})\text{pr}_v(H_{ij} = h_{ij}^l, A_{ij} = a_{ij}; \gamma_v^{(t)})},$$

which is of the same form as equation (2.5) for the weights under population sampling, but with population probabilities replaced by probabilities under VSS. This expression for the weights suggests they can be calculated as though the case-control data were collected prospectively. We show $\text{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \,; \vartheta_v^{(t)})$ can be calculated with the estimates $\beta_{v0}^{(t)}$ and $\beta^{(t)}$ obtained in the M-step, by fitting a weighted logistic regression to the case-control data as though it were collected prospectively. However, calculation of $\text{pr}_v(H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \,; \gamma_v^{(t)})$ is not as straightforward.

As with data collected prospectively, there will be little information available to infer the joint distribution of H and A under VSS. Additionally, the distribution of haplogenotypes under VSS will not be identifiable from the data on single-locus genotypes. For prospectively collected data, the solution to these problems is to impose the assumptions of population HWP and independence of H and A. The expression for the weights then simplifies and depends only on the population haplotype frequencies, which can be estimated. Due to oversampling of the cases, assumptions such as HWP and independence of H and A that may be valid in the general population need not hold in the combined case-control sample (i.e., in the VSS hypothetical population), because the pooled case-control sample is not representative of the general population. Hence weights calculated for case-control data assuming HWP and independence of H and A may be incorrect, which could potentially lead to incorrect inference of haplotype risks. However, for prospectively collected data, modest mis-specification of the weights arising from incorrectly assuming HWP [14] and independence of H and A [4] appears to have little effect on maximum likelihood inference of haplotype risks. Hence, approximate weights for case-control data which incorrectly assume HWP and independence of H and A in the VSS hypothetical population may still lead to reasonable inference. Using the notation for case-control data, the approximate weights are:

$$\tilde{w}_{ijk}(\vartheta_v^{(t)}) = \frac{\mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \,; \beta_{v0}^{(t)}, \beta^{(t)})\tilde{\mathrm{pr}}_v(H_{ij} = h_{ij}^k \,; \gamma_{vh}^{(t)})}{\sum_{l=1}^{K_{ij}} \mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}^l, A_{ij} = a_{ij} \,; \beta_{v0}^{(t)}, \beta^{(t)})\tilde{\mathrm{pr}}_v(H_{ij} = h_{ij}^l \,; \gamma_{vh}^{(t)})},$$

where $\tilde{\mathrm{pr}}_v(H_{ij} = h \,; \gamma_{vh})$ is the haplogenotype frequency in the VSS hypothetical population that would obtain *if* HWP held, and γ_{vh} is a vector of haplotype frequencies in the VSS hypothetical population. We stress that haplogenotype frequencies in the hypothetical population are approximated by values that would obtain under HWP. In summary, we propose that the E-step for case-control data be carried out in exactly the same way as the E-step for cohort data. Consequently, the resulting risk estimators obtained from case-control data will not be maximum likelihood.

2.6.3 M-step

Substituting the reparametrization ϑ_v for ϑ in $Q(\vartheta \mid \vartheta^{(t)})$ and using the approximate weights, the function of ϑ_v to be maximized is now:

$$\begin{aligned}\tilde{Q}(\vartheta_v \mid \vartheta_v^{(t)}) &= \sum_{i=0}^{1}\sum_{j=1}^{n_i}\sum_{k=1}^{K_{ij}} \tilde{w}_{ijk}(\vartheta_v^{(t)}) l_{cij}(\vartheta_v; h_{ij}^k, a_{ij}) \\ &= \sum_{i=0}^{1}\sum_{j=1}^{n_i}\sum_{k=1}^{K_{ij}} \tilde{w}_{ijk}(\vartheta_v^{(t)}) \log \mathrm{pr}(H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \mid D_{ij} = i; \vartheta_v),\end{aligned}$$

which is a weighted retrospective log-likelihood. Similar arguments to those in Prentice and Pyke [16] may be used to show that this weighted retrospective likelihood is maximized by (i) $\hat{\beta}$ from a weighted logistic regression, fit as though the data were collected prospectively, and (ii) the empirical distribution $\hat{\gamma}_h$ of H and A from the case-control sample to which

(h_{ij}^k, a_{ij}) $(i = 0, 1,\ j = 1, \ldots, n_i,\ k = 1, \ldots, K_{ij})$ contributes mass $\tilde{w}_{ijk}(\vartheta_v^{(t)})/n$. This argument is sketched in Appendix C.

In the M-step the distribution of H and A is *not* modelled so that appropriate estimates of β are obtained given fixed weights. However, the approximation to the weights made in the E-step requires estimates of the haplotype frequencies γ_{vh} under VSS, rather than estimates of the joint distribution γ_v of H and A under VSS. For prospective data, sample haplotype frequencies are used to estimate population haplotype frequencies. Hence applying PML to case-control data will estimate the haplotype frequencies γ_{vh} in the pooled case-control sample by the corresponding empirical frequencies $\hat{\gamma}_{vh}$. Using the empirical haplotype frequencies $\hat{\gamma}_{vh}$ in the E-step is reasonable, because the approximate weights depend on the joint distribution γ_v of haplogenotypes H and non-genetic covariates A only through γ_{vh}. In summary, the M-step for case-control data may be carried out in exactly the same way as the M-step for cohort data, but inference is not maximum likelihood because the weights in E-step are only approximate.

To recap, when PML is naively applied to case-control data, weights in the E-step are approximated under the incorrect assumptions of HWP for H and independence of H and A in the VSS population. The M-step properly maximizes the resulting weighted retrospective log-likelihood. In prospective or cross-sectional studies of populations with departures from HWP and from independence of H and A, the weights in the E-step would also be approximations. However, simulation studies of the prospective approach suggest these approximate weights do not lead to substantial bias in haplotype risk estimators [4,14]. Hence, we might hope that approximate weights in the case-control context would also work reasonably well. We will investigate this hypothesis in the simulation study.

2.7 PML versus EE for Case-Control Data

Due to heavy oversampling of cases in a case-control study of a rare disease, the PML estimators of haplotype frequencies are obviously incorrect. We thus consider estimating equations for β only. We shall find that in these estimating equations the only point of difference between approaches is how they approximate the weights. In the E-step of PML,

$$\tilde{Q}(\vartheta_v \mid \vartheta_v^{(t)}) = \sum_{i=0}^{1} \sum_{j=1}^{n_i} \sum_{k=1}^{K_{ij}} \tilde{w}_{ijk}(\vartheta_v^{(t)}) \; l_{cij}(\vartheta_v; h_{ij}^k, a_{ij}),$$

where $\vartheta_v = (\beta, \gamma_v)$, and γ_v parametrizes $\mathrm{pr}_v(H_{ij} = h, A_{ij} = a)$. In the M-step, we maximize $\tilde{Q}(\vartheta_v \mid \vartheta_v^{(t)})$ over ϑ_v to find $\vartheta_v^{(t+1)}$. The corresponding estimating equation for $\beta^{(t+1)}$ is therefore

$$\frac{\partial}{\partial \beta} \tilde{Q}(\vartheta_v \mid \vartheta_v^{(t)}) \Big|_{\beta^{(t+1)}} = \sum_{i=0}^{1} \sum_{j=1}^{n_i} \sum_{k=1}^{K_{ij}} \tilde{w}_{ijk}(\vartheta_v^{(t)}) \; \frac{\partial}{\partial \beta} l_{cij}(\vartheta_v; h_{ij}^k, a_{ij}) \Big|_{\beta^{(t+1)}} = 0.$$

Upon convergence, the resulting estimates $\hat{\vartheta}_v$ solve an estimating equation for β that is analogous to Louis' equation for the score function for β [see equation (2.1)]

$$0 = \sum_{i=0}^{1} \sum_{j=1}^{n_i} \sum_{k=1}^{K_{ij}} \tilde{w}_{ijk}(\hat{\vartheta}_v) \; \frac{\partial}{\partial \beta} l_{cij}(\vartheta_v; h_{ij}^k, a_{ij}) \Big|_{\hat{\beta}},$$

where

$$\tilde{w}_{ijk}(\hat{\vartheta}_v) \quad = \quad \frac{\mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \; ; \hat{\beta}_{v0}, \hat{\beta}) \tilde{\mathrm{pr}}_v(H_{ij} = h_{ij}^k \; ; \hat{\gamma}_{vh})}{\sum_{l=1}^{K_{ij}} \mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}^l, A_{ij} = a_{ij} \; ; \hat{\beta}_{v0}, \hat{\beta}) \tilde{\mathrm{pr}}_v(H_{ij} = h_{ij}^l \; ; \hat{\gamma}_{vh})},$$

$\tilde{\mathrm{pr}}_v(H_{ij} = h \; ; \gamma_{vh})$ is the haplogenotype frequency in the VSS hypothetical population that would obtain *if* HWP held, and γ_{vh} is a vector of haplotype frequencies in the VSS hypothetical population. The weights used are approximations to the correct weights:

$$w_{ijk}(\vartheta_v) = \frac{\mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}^k, A_{ij} = a_{ij}; \beta_{v0}, \beta) \mathrm{pr}_v(H_{ij} = h_{ij}^k, A_{ij} = a_{ij}; \gamma_v)}{\sum_{l=1}^{K_{ij}} \mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}^l, A_{ij} = a_{ij}; \beta_{v0}, \beta) \mathrm{pr}_v(H_{ij} = h_{ij}^l, A_{ij} = a_{ij}; \gamma_v)}.$$

In the EE approach, the estimating equations for β are:

$$0 = \sum_{i=0}^{1} \sum_{j=1}^{n_i} \sum_{k=1}^{K_{ij}} \tilde{\tilde{w}}_{ijk}(\hat{\theta}) \; \frac{\partial}{\partial \beta} l_{cij}(\theta; h_{ij}^k, a_{ij}) \Big|_{\hat{\beta}},$$

where $\tilde{\tilde{w}}_{ijk}(\theta)$ denotes their approximation to the true weights, as described below. These estimating equations are clearly very similar in form to those of PML. Equation (2.7) shows that the weights for case-control data are the same conditional probabilities as the weights for cohort data in equation (2.4). Writing the weights for case-control data in equation

(2.7) in terms of the prospective parameters $\theta = (\beta_0, \beta^T, \gamma^T)^T$ rather than ϑ and expanding these conditional probabilities in the manner outlined in equation (2.5) leads to:

$$\begin{aligned} w_{ijk}(\theta) &= \mathrm{pr}(H_{ij} = h_{ij}^k \mid G_{ij} = g_{ij}, A_{ij} = a_{ij}, D_{ij} = i\ ;\theta) \\ &= \frac{\mathrm{pr}(D_{ij} = i \mid H_{ij} = h_{ij}^k, A_{ij} = a_{ij}\ ; \beta_0, \beta)\mathrm{pr}(H_{ij} = h_{ij}^k, A_{ij} = a_{ij}\ ; \gamma)}{\sum_{l=1}^{K_{ij}} \mathrm{pr}(D_{ij} = i \mid H_{ij} = h_{ij}^l, A_{ij} = a_{ij}\ ; \beta_0, \beta)\mathrm{pr}(H_{ij} = h_{ij}^l, A_{ij} = a_{ij}\ ; \gamma)}. \end{aligned}$$

Assuming independence of H and A and HWP in the population, Zhao et al. [20] and Stram et al. [17] simplify these weights to:

$$w_{ijk}(\theta) = \frac{\mathrm{pr}(D_{ij} = i | H_{ij} = h_{ij}^k, A_{ij} = a_{ij}; \beta_0, \beta)\mathrm{pr}(H_{ij} = h_{ij}^k\ ; \gamma_h)}{\sum_{l=1}^{K_{ij}} \mathrm{pr}(D_{ij} = i | H_{ij} = h_{ij}^l, A_{ij} = a_{ij}; \beta_0, \beta)\mathrm{pr}(H_{ij} = h_{ij}^l; \gamma_h)}. \tag{2.10}$$

To obtain their approximation $\tilde{\tilde{w}}_{ijk}(\theta)$ to the weights, Zhao et al. [20] first approximate the penetrance $\mathrm{pr}(D_{ij} = i \mid H_{ij} = h_{ij}^k, A_{ij} = a_{ij}\ ; \beta_0, \beta)$ under a rare disease assumption, substitute this approximation into equation (2.10), and find that β_0 cancels out of the expression. Such cancellation is convenient because β_0 cannot be estimated from case-control data without knowledge of the population disease risk [5]. Moreover, since population haplotype frequencies γ_h cannot be estimated conveniently from a biased sample, Zhao et al [20] again invoke the rare disease assumption and use the control sample to approximate the frequencies. Since the rare disease assumption is invoked both in approximating the penetrance and the population haplotype frequencies, we speculate that for case-control studies of more common diseases, the approximation $\tilde{\tilde{w}}_{ijk}$ from EE may provide a poorer fit to the true weights than the approximation $\tilde{w}_{ijk}$ from PML.

Chapter 3

Simulation Study

There are two specific aims of the simulation study: 1) to assess the bias of PML and EE haplotype risk estimators and their standard errors when these methods are applied to case-control data, and 2) to compare the efficiency of PML and EE to true maximum-likelihood inference (ML) for estimating haplotype risks when there are no non-genetic covariates in the penetrance model, but haplogenotypes are ambiguous. For penetrance models with non-genetic attributes, no maximum likelihood procedure has been developed. We used the software packages hapassoc [4], Hplus [20] and Chaplin [8] for PML, EE and ML inference, respectively.

3.1 Design

We considered haplotypes comprised of 3 SNPs in two separate simulation configurations which differed in haplotype frequencies. The choice of haplotype frequencies was motivated by the observations of Stram et al. [17]. These authors noted an apparent bias in haplotype risk estimates from PML when the ability of the single-locus genotypes to predict the number of copies of a risk haplotype h, as measured by R_h^2 (Appendix B), was $\leq 80\%$. We investigated the bias of haplotype risk estimators at one value of R_h^2 just under 80% (78.75%) and one value well below 80% (59.78%). Haplotype frequencies for the two configurations are given in Table 3.1, and have been deliberately set to achieve the desired R_h^2 values. We have observed that haplotype frequencies are not typically this uniform. The haplotype labelled h_1 (SNP1 = 0, SNP2 = 0, SNP3 = 0) was chosen to be the risk haplotype. It was necessary to choose the frequency of h_1 to be less than at least one of the other haplotypes because of a limitation in the Hplus software that prevents calculation of haplotype risk estimates for the most frequent haplotype in the sample. (Hplus automatically chooses the most frequent haplotype in the sample to be the baseline haplotype in the logistic regression model.) The haplotype frequencies of the first and second set of simulations give $R^2_{h_1}$ of 78.75 % and 59.78 %, respectively. Note that $R^2_{h_i}$, $i \neq 1$, does not have any impact on the results, since h_i, $i \neq 1$, are non-risk haplotypes.

In both configurations, the risk haplotype was simulated to have a multiplicative effect on disease risk, increasing the odds of being affected by a factor of $\exp(0.7) \approx 2$ (i.e. $\beta_{h_1} = 0.7$) for every copy of the risk haplotype that replaces a baseline haplotype. The intercept term in the penetrance model for each configuration was chosen so that the probability of developing disease over the study period was 0.0009. This corresponds to a two-year study of a disease with an annual incidence rate of 45 per hundred thousand people per year, consistent with Scandinavian rates of Type 1 Diabetes.

For each set of simulations, haplogenotypes and disease status were generated for a sample of size 1.5×10^6 individuals. From this sample, 1000 cases and 1000 controls were sub-sampled randomly from within each disease class and were recorded as data. The large number of cases and controls was chosen to limit the well-known finite-sample bias of maximum-likelihood inference in logistic regression [10], which could cloud bias comparisons. A penetrance model with multiplicative effects for each of the 7 non-baseline haplotypes was then fit to the data with hapassoc, Hplus and Chaplin. (The Hplus software package does not allow specification of *any other* penetrance model, in fact.) For each set of simulations, 1000 data sets were generated to obtain the empirical distribution of haplotype risk estimators, their standard errors, estimates of 95% coverage probabilities, and power.

Table 3.1: Haplotype Frequencies in Configuration 1 (Left) and 2 (Right)

label	haplotype	p_h	label	haplotype	p_h
h_1	000	$p_1 = .23$	h_1	000	$p_1 = .07$
h_2	001	$p_2 = .27$	h_2	001	$p_2 = .93/7$
h_3	010	$p_3 = .15$	h_3	010	$p_3 = .93/7$
h_4	011	$p_4 = .10$	h_4	011	$p_4 = .93/7$
h_5	100	$p_5 = .10$	h_5	100	$p_5 = .93/7$
h_6	101	$p_6 = .05$	h_6	101	$p_6 = .93/7$
h_7	110	$p_7 = .05$	h_7	110	$p_7 = .93/7$
h_8	111	$p_8 = .05$	h_8	111	$p_8 = .93/7$

3.2 Results

In the tables and figures summarizing the results, risk estimates for haplotype h_1 are compared to their true value of $\beta_{h_1} = 0.7$. Similarly, the empirical means of their associated standard errors are compared to the standard deviations of risk estimates over simulation replicates. The standard deviations of the risk estimates are considered to be the true values that the standard errors are estimating (within simulation error). In the figures, mirror-image distributions are drawn to compare the results of hapassoc and Hplus (above the horizontal axis) against the "gold standard" Chaplin (below the horizontal axis).

3.2.1 First Set of Simulations

The risk haplotype h_1 estimated the most frequent haplotype among controls in 19 of the 1000 simulation replicates generated under the first configuration by Hplus. Consequently, the investigation of Hplus is based on only 981 simulation replicates. The investigation of hapassoc and Chaplin is based on all 1000 replicates. Results for haplotype risk estimators are summarized in Table 3.2 and Figure 3.1. Bias of the hapassoc and Chaplin estimators is within simulation error of zero (see Appendix A for a review of simulation error), while bias of the Hplus estimator slightly exceeds simulation error.

Results for standard errors are summarized in Table 3.3 and Figure 3.2. All methods show a slight upward (conservative) bias in the standard error estimator (0.00075, 0.00041, 0.00659 for Chaplin, hapassoc, and Hplus, respectively). Bias in the hapassoc estimator (0.00041) is of the same order as the finite-sample bias in the Chaplin estimator (0.00075).

Table 3.2: Haplotype Risk Estimates - Configuration 1

method	mean	bias	simulation error
Chaplin	0.69795	-0.00205	0.00681
hapassoc	0.70462	0.00462	0.00712
Hplus	0.70721	0.00721	0.00704

Table 3.3: Standard Error of Haplotype Risks - Configuration 1

method	mean($\overline{se}$)	$\hat{\beta}$ std dev	bias	simulation error
Chaplin	0.10836	0.10761	0.00075	0.00014
hapassoc	0.11297	0.11256	0.00041	0.00021
Hplus	0.11686	0.11026	0.00659	0.00102

In contrast, bias in the Hplus estimator is noticeably larger than that of the other two methods (0.00659). Additionally, the empirical distribution of standard errors for Hplus is clearly more spread out and has a heavier right tail than the empirical distributions for either hapassoc or Chaplin (see Figure 3.2).

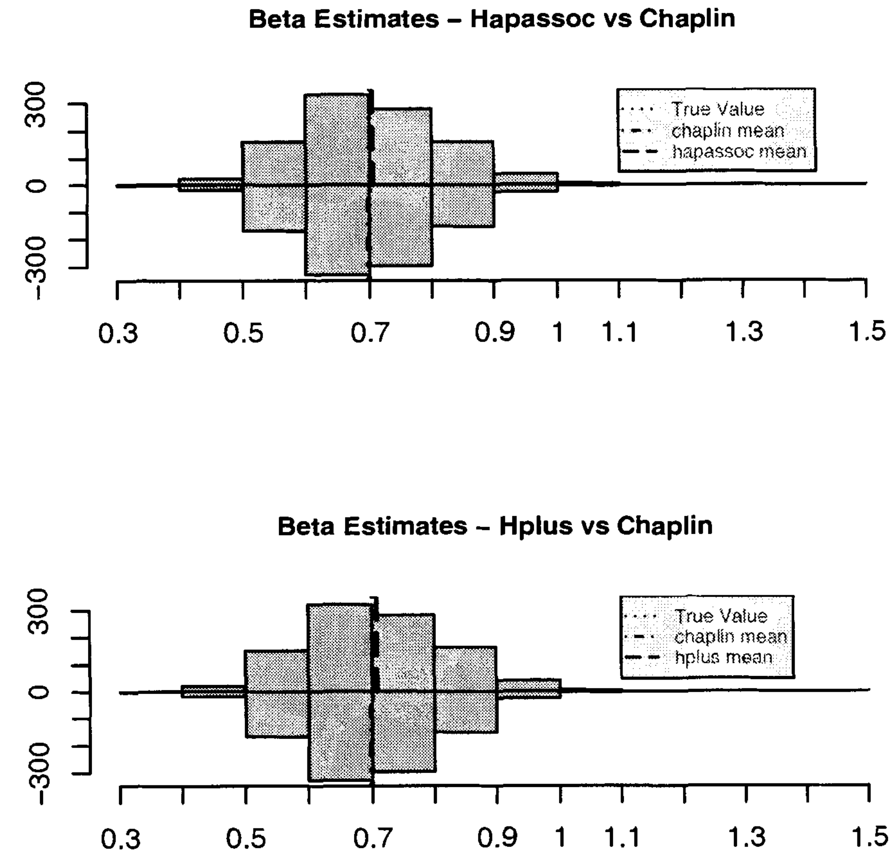
All methods had empirical power of 100% to detect the effect of the risk haplotype (i.e. the null hypothesis of no effect was rejected for all simulated data sets). The 95% confidence intervals for all 3 methods had slightly conservative coverage probabilities of ~96%.

3.2.2 Second Set of Simulations

For all 3 methods, the investigation of statistical properties is based on 1000 simulation replicates. Results for haplotype risk estimators are summarized in Table 3.4 and Figure 3.3. Those of hapassoc and Chaplin appear unbiased, but the Hplus estimator appears to be biased upward.

The results for standard errors are summarized in Table 3.5 and Figure 3.4. The methods all show bias, with Chaplin and Hplus biased downwards (anticonservative) and hapassoc biased upwards (conservative). As with the first configuration, the magnitude of bias in the hapassoc estimator is comparable to the finite-sample bias in the Chaplin estimator. As before, bias is greatest for the Hplus estimator, which is again the most variable of the 3 estimators.

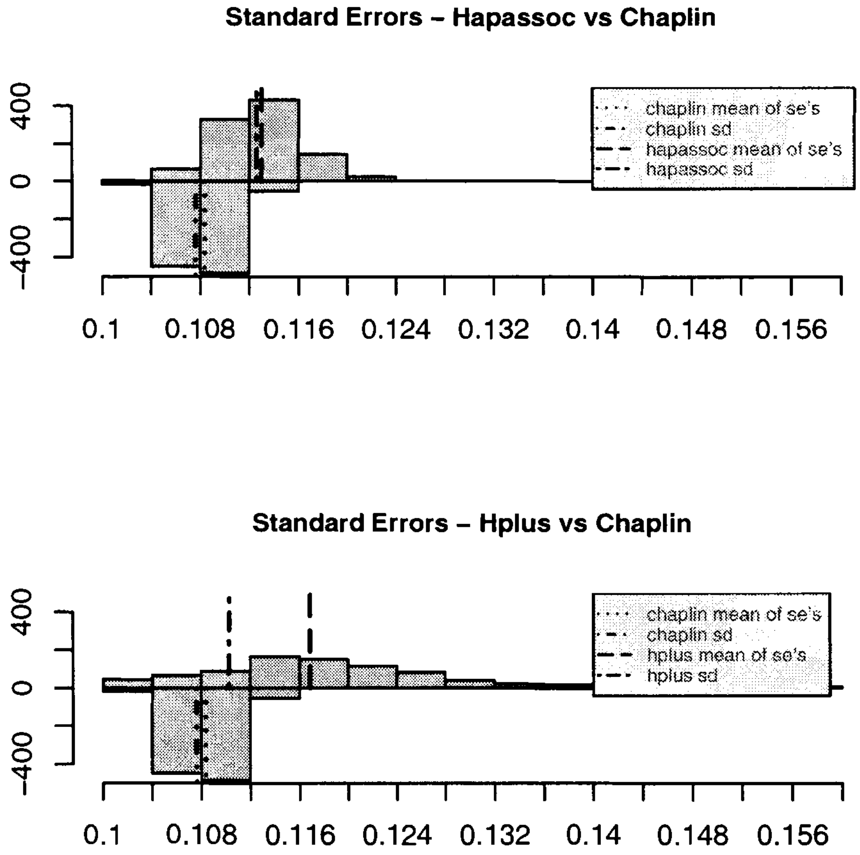
Figure 3.1: Distribution of Haplotype Risk Estimates - Configuration 1


Beta Estimates - Hapassoc vs Chaplin

Beta Estimates - Hplus vs Chaplin

The anticonservative estimates and standard errors from Hplus combine to give an anticonservative coverage probability of only 83.6% for the 95% confidence interval. Coverage probabilites from hapassoc and Chaplin are 94.5% and 94.2%, respectively, and within simulation error. The empirical power of Chaplin to detect the effect of the risk haplotype is 97.1%. Standard large-sample theory suggests that true maximum-likelihood inference (i.e. Chaplin) should be the most efficient of the 3 methods, and so it is not surprising that the 97.1% power of Chaplin is greater than the 92.7% power of hapassoc. The power of Hplus is not relevant given the anticonservative behaviour of the Hplus haplotype risk estimators,

Figure 3.2: Distribution of Standard Errors - Configuration 1



Standard Errors – Hapassoc vs Chaplin

Standard Errors – Hplus vs Chaplin

standard errors and coverage probabilities.

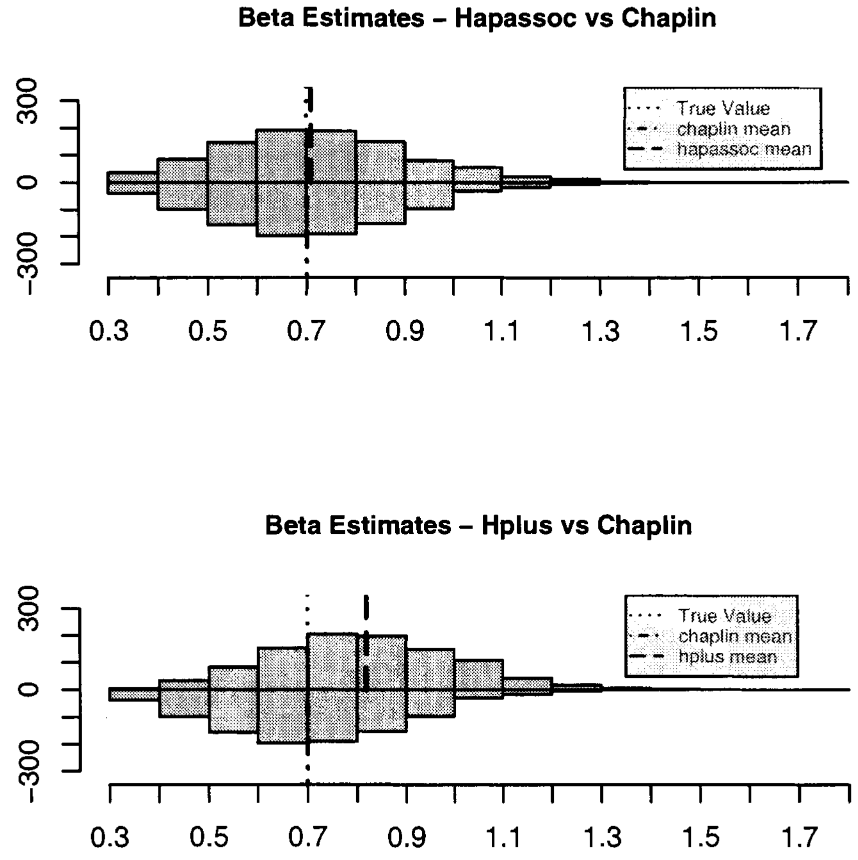
Table 3.4: Haplotype Risk Estimates - Configuration 2

method	mean	bias	simulation error
Chaplin	0.70066	0.00066	0.01214
hapassoc	0.70913	0.00913	0.01341
Hplus	0.81807	0.11807	0.01177

Table 3.5: Standard Error of Haplotype Risks - Configuration 2

method	mean($\overline{se}$)	$\hat{\beta}$ std dev	bias	simulation error
Chaplin	0.18641	0.19191	-0.00550	0.00054
hapassoc	0.21404	0.21209	0.00195	0.00061
Hplus	0.15976	0.18608	-0.02631	0.00139

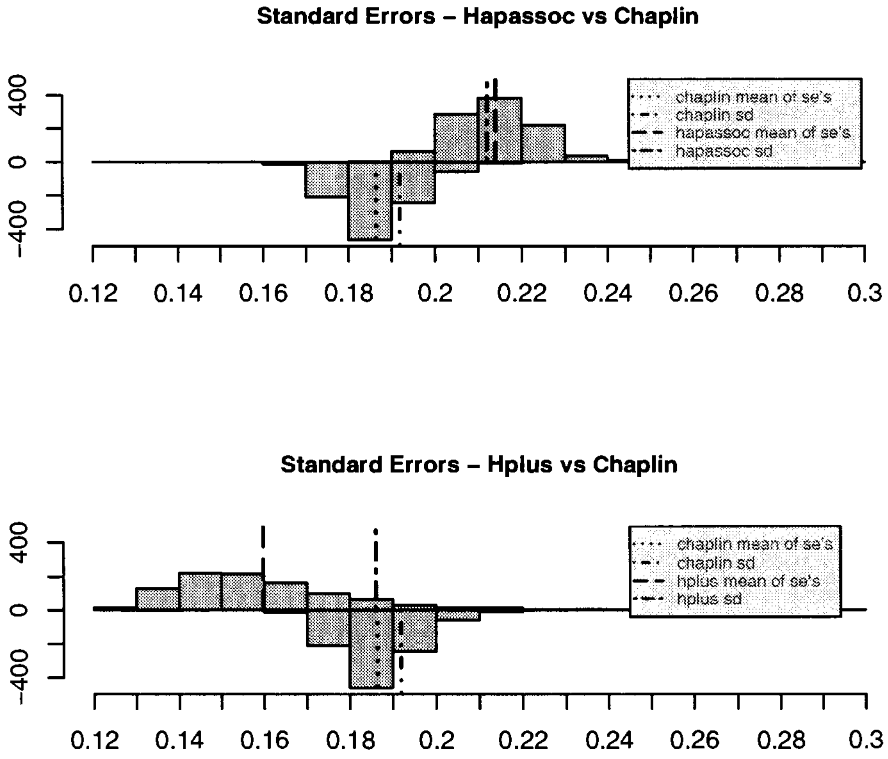
Figure 3.3: Distribution of Haplotype Risk Estimates - Configuration 2



Beta Estimates - Hapassoc vs Chaplin

Beta Estimates - Hplus vs Chaplin

Figure 3.4: Standard Error Distributions - Simulation 2



Standard Errors – Hapassoc vs Chaplin

Standard Errors – Hplus vs Chaplin

Chapter 4

Conclusions and Future Work

We have considered the problem of haplotype risk inference from case-control data in which haplotype phase information is missing for some subjects. We have provided a theoretical justification for applying PML to case-control data, and have explored similarities and differences between PML and EE. Statistical properties of the two approaches were compared by simulation for penetrance models with one risk haplotype and no non-genetic covariates. Two simulation configurations with different R_h^2 were considered. In contrast to the conclusions of Stram et al. [17], PML provided unbiased haplotype risk estimates for $R_h^2 < 80\%$. However, $R_h^2 < 80\%$ lead to EE-estimates of haplotype risk that were too large. In our simulations, standard errors of haplotype risk estimates were biased for both PML and EE. Bias increased as R_h^2 decreased and was largest for the EE approach. The distribution of EE-standard errors was the most variable, possibly due to the empirical nature of this estimator. For a lower R_h^2 of 59.78%, EE but not PML lead to anticonservative inference (e.g. EE coverage probabilities of 83.6% for 95% confidence intervals). Future work includes a comparison of PML and EE for other models of disease penetrance, such as those which include non-genetic risk factors, but software for true maximum-likelihood inference in this situation is not currently available. However, this future work is hampered by the Hplus software (EE), which currently does not permit specification of a non-multiplicative penetrance model.

Appendix A

Simulation Error

In this thesis, simulation was used to investigate statistical properties of two different methods of haplotype risk inference. For each of a large number of simulated data sets, haplotype risk estimates and their standard errors, coverage of 95% confidence intervals and the outcome of the hypothesis test of no genetic effect on disease were calculated and recorded. Mean values of such quantities over simulation replicates provide estimates of bias, coverage probabilities, and power, respectively. The simulation-based estimates are subject to error, which may be quantified as follows. Let $\widehat{\mu}$ be a random variable with unknown mean μ and variance σ^2. Suppose we estimate μ by taking an average, $\overline{\widehat{\mu}}$, of a large number N of independent realizations of $\widehat{\mu} : \widehat{\mu}_1 \cdots \widehat{\mu}_N$. By the Central Limit Theorem, $\overline{\widehat{\mu}} \dot{\sim} N(\mu, \sigma^2/N)$, where $\dot{\sim}$ denotes "approximately distributed as". The *Monte Carlo* or *simulation error* in the estimate, $\overline{\widehat{\mu}}$, of μ is $2\sqrt{\widehat{\sigma^2}/N}$, where $\widehat{\sigma^2}$ is some estimate of the variance of $\widehat{\mu}$. If $|\overline{\widehat{\mu}} - \mu_0|$ is larger than $2\sqrt{\widehat{\sigma^2}/N}$, we would reject $H_0 : \mu = \mu_0$ at the 5% level, because the approximate 95% confidence interval for μ, $\overline{\widehat{\mu}} \pm 2\sqrt{\widehat{\sigma^2}/N}$ does not cover μ_0.

Appendix B

The R_h^2 Measure

$R_h^2 = \frac{\mathrm{V}\{\mathrm{E}[\delta_h(H)|G]\}}{\mathrm{V}[\delta_h(H)]}$ measures how well the single-locus genotype data G predicts the "dosage" δ_h (=0,1,2) of haplotype h in an individual [17]. Let H be the haplogenotype, and p_h, p_H and p_G be, respectively, the population frequencies of h, H and G. We consider haplotypes specified for our simulation configurations (see Table 3.1). Assuming HWP for haplogenotypes, the haplotypes within an individual are independent. Hence, $\delta_{h(H)}$ is binomial with "success" probability p_h and number of trials 2, so that $\mathrm{E}[\delta_h(H)] = 2p_h$ and $\mathrm{V}[\delta_h(H)] = 2p_h(1-p_h)$. Also,

$$\mathrm{V}\{\mathrm{E}[\delta_h(H) \mid G]\} \;=\; \mathrm{E}\{\mathrm{E}[\delta_h(H) \mid G]^2\} - \mathrm{E}[\delta_h(H)]^2 = \left\{\sum_G \mathrm{E}[\delta_h(H) \mid G]^2 p_G\right\} - (2p_h)^2.$$

Since h_1 is the risk haplotype, we are interested in $R_{h_1}^2$. To calculate $\sum_G \mathrm{E}[\delta_{h_1}(H) \mid G]^2 p_G$, we need only consider haplogenotypes H that contain at least one h_1 or that contain no h_1 but lead to single-locus genotypes G consistent with H's that contain an h_1. So, we obtain:

$$\sum_G \mathrm{E}[\delta_{h_1}(H) \mid G]^2 p_G = 4p_1^2 + 2(p_1p_2 + p_1p_3 + p_1p_5) +$$
$$2\left[\frac{(p_1p_4)^2}{p_1p_4 + p_2p_3} + \frac{(p_1p_6)^2}{p_1p_6 + p_2p_5} + \frac{(p_1p_7)^2}{p_1p_7 + p_3p_5} + \frac{(p_1p_8)^2}{p_1p_8 + p_2p_7 + p_3p_6 + p_4p_5}\right].$$

And so,

$$R_{h_1}^2 = \frac{\mathrm{V}\{\mathrm{E}[\delta_{h_1}(H) \mid G]\}}{\mathrm{V}[\delta_{h_1}(H)]} = \frac{\{\sum_G \mathrm{E}[\delta_{h_1}(H) \mid G]^2 p_G\} - 4p_1^2}{2p_1(1-p_1)} = num/den,$$

where $den = 1 - p_1$ and,

$$num = p_2 + p_3 + p_5 + \frac{p_1p_4^2}{p_1p_4+p_2p_3} + \frac{p_1p_6^2}{p_1p_6+p_2p_5} + \frac{p_1p_7^2}{p_1p_7+p_3p_5} + \frac{p_1p_8^2}{p_1p_8+p_2p_7+p_3p_6+p_4p_5}.$$

Appendix C

Retrospective Log-Likelihoods

In this appendix we show that (i) the weighted retrospective log-likelihood $\tilde{Q}(\vartheta_v \mid \vartheta_v^{(t)})$ (defined in section 2.6.3) is maximized with respect to β by fitting a weighted logistic regression as though the data were collected prospectively; and (ii) the maximizer with respect to γ_v is an empirical distribution that puts mass $\tilde{w}_{ijk}(\vartheta_v^{(t)})/n$ at each (h_{ij}^k, a_{ij}). These results are a slight extension of Prentice and Pyke [16], which states that (i) the retrospective log-likelihood (no weights) is maximized with respect to the odds-ratio parameters by fitting a logistic regression as though the data were collected prospectively; and (ii) the maximizer with respect to the covariate distribution under VSS is an empirical distribution that puts mass $1/n$ at each covariate value observed in the case-control sample. We start with a review of the Prentice and Pyke argument, and then discuss the extension to accommodate weights.

C.1 No Missing Data

Assuming fully observed haplogenotypes, the case-control log-likelihood is: $l(\vartheta) = \sum_{i=0}^{1} \sum_{j=1}^{n_i} \log \mathrm{pr}(H_{ij} = h_{ij}, A_{ij} = a_{ij} \mid D_{ij} = i, \vartheta)$. Restating equation (2.9),

$$\begin{aligned}
&\mathrm{pr}(H_{ij} = h_{ij}, A_{ij} = a_{ij} \mid D_{ij} = i \,; \vartheta) \\
&\qquad = \frac{\mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}, A_{ij} = a_{ij} \,; \beta_{v0}, \beta)\mathrm{pr}_v(H_{ij} = h_{ij}, A_{ij} = a_{ij} \,; \gamma_v)}{\mathrm{pr}_v(D_{ij} = i)},
\end{aligned}$$

where $\mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}, A_{ij} = a_{ij}\ ; \beta_{v0}^{(t)}, \beta^{(t)})$ is a logistic regression, with population odds-ratio parameters β but with a different intercept β_{v0}; $\mathrm{pr}_v(H_{ij} = h_{ij}, A_{ij} = a_{ij}\ ; \gamma_v^{(t)})$ is parametrized by γ_v; and $\mathrm{pr}_v(D_{ij} = i) = n_i/n$. Therefore, up to a constant term, the reparametrized log-likelihood is:

$$\begin{aligned} l(\vartheta_v) &= \sum_{i=0}^{1}\sum_{j=1}^{n_i} \log \mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}, A_{ij} = a_{ij}\ ; \beta_{v0}, \beta) \\ &+ \sum_{i=0}^{1}\sum_{j=1}^{n_i} \log \mathrm{pr}_v(H_{ij} = h_{ij}, A_{ij} = a_{ij}\ ; \gamma_v). \end{aligned}$$

The parameters $(\beta_{v0}, \beta, \gamma_v)$ are constrained by the fact that the joint distribution of haplogenotypes and attributes given disease status integrates to one, or by:

$$\sum_{h\in\mathcal{H}} \int_{a\in\mathcal{A}} \mathrm{pr}(H_{ij} = h, A_{ij} = a \mid D_{ij} = i; \beta_{v0}, \beta, \gamma_v) da = 1 \quad i = 0, 1 \tag{C.1}$$

where $\mathcal{H}$ is the set of all haplogenotypes and $\mathcal{A}$ is the set of all attribute values. It can be shown that if equation (C.1) holds for one value of i (e.g. $i = 1$), it holds for the other. Therefore, equation (C.1) describes a single constraint

$$\sum_{h\in\mathcal{H}} \int_{a\in\mathcal{A}} \mathrm{pr}(H_{1j} = h, A_{1j} = a \mid D_{1j} = 1; \beta_{v0}, \beta, \gamma_v) da = 1.$$

Hence, from equation (2.9), an equivalent expression for the constraint is:

$$\sum_{h\in\mathcal{H}} \int_{a\in\mathcal{A}} \mathrm{pr}_v(D_{1j} = 1 \mid H_{1j} = h, A_{1j} = a\ ; \beta_{v0}, \beta) \mathrm{pr}_v(H_{1j} = h, A_{1j} = a\ ; \gamma_v) = \frac{n_1}{n}. \tag{C.2}$$

Though maximization of the log-likelihood $l(\vartheta_v)$ is subject to the constraint of equation (C.2), Prentice and Pyke [16] maximize without regard to the constraint. They then show that the unconstrained maximizer $\hat{\vartheta}_v = (\hat{\beta}_{v0}, \hat{\beta}, \hat{\gamma}_v)$ satisfies the constraint and hence is the MLE. Details are as follows.

The factorization of the reparametrized log-likelihood obtained when the constraint on $\vartheta_v = (\beta_{v0}, \beta, \gamma_v)$ is ignored implies that maximization with respect to β_{v0} and β is obtained by maximizing the first term of $l(\vartheta_v)$:

$$\sum_{i=0}^{1}\sum_{j=1}^{n_i} \log \mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}, A_{ij} = a_{ij}\ ; \beta_{v0}, \beta),$$

which is of the same form as a log-likelihood from a prospective logistic regression. The maximizers $(\hat{\beta}_{v0}, \hat{\beta})$ may be obtained in the usual way as the solution to a set of score equations.

Maximization of the log-likelihood with respect to γ_v is obtained by maximizing the second term of $l(\vartheta_v)$:

$$\sum_{i=0}^{1}\sum_{j=1}^{n_i} \log \mathrm{pr}_v(H_{ij} = h_{ij}, A_{ij} = a_{ij} \; ; \gamma_v).$$

Assuming the joint distribution of H and A is not modelled, this expression is a non-parametric log-likelihood whose maximizer $\hat{\gamma}_v$ is the usual empirical distribution function that puts mass $1/n$ at each data point (h_{ij}, a_{ij}) observed in the case-control sample (e.g. van der Vaart [19], pages 402-403).

The unconstrained maximizers $\hat{\vartheta}_v$ can be seen to satisfy the constraint in equation (C.2) by comparing the score equation for β_{v0} evaluated at $\hat{\vartheta}_v$ to the constraint evaluated at $\hat{\vartheta}_v$. The score equation for β_{v0} evaluated at $\hat{\vartheta}_v$ is:

$$0 = \frac{\partial}{\partial \beta_{v0}} l(\vartheta_v)\bigg|_{\hat{\vartheta}_v} = n_1 - \sum_{i=0}^{1}\sum_{j=1}^{n_i} \frac{\exp\left\{\hat{\beta}_{v0} + X(h_{ij}, a_{ij})\hat{\beta}\right\}}{1 + \exp\left\{\hat{\beta}_{v0} + X(h_{ij}, a_{ij})\hat{\beta}\right\}}. \tag{C.3}$$

The constraint (C.2) evaluated at $\hat{\vartheta}_v$ is:

$$\sum_{i=0}^{1}\sum_{j=1}^{n_i} \mathrm{pr}_v(D_{1j} = 1 \mid H_{1j} = h_{ij}, A_{1j} = a_{ij} \; ; \hat{\beta}_{v0}, \hat{\beta})\frac{1}{n} = \frac{n_1}{n},$$

because an integral with respect to an empirical distribution function is a sum over the observed data points, with each data point weighted by $1/n$. The above expression for the constraint simplifies to

$$\sum_{i=0}^{1}\sum_{j=1}^{n_i} \frac{\exp\left\{\hat{\beta}_{v0} + X(h_{ij}, a_{ij})\hat{\beta}\right\}}{1 + \exp\left\{\hat{\beta}_{v0} + X(h_{ij}, a_{ij})\hat{\beta}\right\}} - n_1 = 0,$$

which is the same as equation (C.3). Hence the joint maximizer $\hat{\vartheta}_v$, which satisfies the score equation (C.3) by definition, also satisfies the constraint of equation (C.2). The conclusion is that (i) the maximizer $\hat{\beta}$ with respect to the odds-ratio parameters of interest is obtained by solving score equations from a logistic regression, as though the data were collected prospectively; and (ii) the maximizer $\hat{\gamma}_v$ with respect to γ_v is the empirical distribution that puts mass $1/n$ at each value (h_{ij}, a_{ij}) observed in the case-control sample.

C.2 Weighted Case

The argument for the weighted case is very similar and so in places only a sketch of the details is given. The weighted retrospective log-likelihood is, up to a constant term,

$$\begin{aligned}\tilde{Q}(\vartheta_v \mid \vartheta_v^{(t)}) &= \sum_{i=0}^{1}\sum_{j=1}^{n_i}\sum_{k=1}^{K_{ij}} \tilde{w}_{ijk}(\vartheta_v^{(t)}) \log \mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \,; \beta_{v0}, \beta) + \\ &\quad \sum_{i=0}^{1}\sum_{j=1}^{n_i}\sum_{k=1}^{K_{ij}} \tilde{w}_{ijk}(\vartheta_v^{(t)}) \mathrm{pr}_v(H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \,; \gamma_v).\end{aligned}$$

Maximization of $\tilde{Q}(\vartheta_v \mid \vartheta_v^{(t)})$ is subject to the same constraint on $\vartheta_v = (\beta_{v0}, \beta, \gamma_v)$ as in the unweighted problem (see equation C.2). As in the unweighted problem, the approach in the weighted problem is to maximize without regard to the constraint and then show that the unconstrained maximizer $\vartheta_v^{(t+1)} = (\beta_{v0}^{(t+1)}, \beta^{(t+1)}, \gamma_v^{(t+1)})$ satisfies the constraint.

Maximization with respect to β_{v0} and β, ignoring the constraint on $\vartheta_v = (\beta_{v0}, \beta, \gamma_v)$, is achieved by maximizing

$$\sum_{i=0}^{1}\sum_{j=1}^{n_i}\sum_{k=1}^{K_{ij}} \tilde{w}_{ijk}(\vartheta_v^{(t)}) \log \mathrm{pr}_v(D_{ij} = i \mid H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \,; \beta_{v0}, \beta).$$

The maximizers $(\beta_{v0}^{(t+1)}, \beta^{(t+1)})$ may be obtained as the solution to prospective weighted logistic regression score equations. Maximization of the log-likelihood with respect to γ_v is achieved by maximizing

$$\sum_{i=0}^{1}\sum_{j=1}^{n_i}\sum_{k=1}^{K_{ij}} \tilde{w}_{ijk}(\vartheta_v^{(t)}) \mathrm{pr}_v(H_{ij} = h_{ij}^k, A_{ij} = a_{ij} \,; \gamma_v).$$

Assuming the joint distribution of H and A is not modeled, this expression is a weighted non-parametric log-likelihood. Similar arguments to those in the unweighted problem show that the maximizer $\gamma_v^{(t+1)}$ is the empirical distribution function that puts mass $\tilde{w}_{ijk}(\vartheta_v^{(t)})/n$ at each (h_{ij}^k, a_{ij}).

The unconstrained maximizers $\vartheta_v^{(t+1)}$ can be seen to satisfy the constraint in equation (C.2) by comparing the score equation for β_{v0} evaluated at $\vartheta_v^{(t+1)}$ to the constraint evaluated at $\vartheta_v^{(t+1)}$. The score equation for β_{v0} evaluated at $\vartheta_v^{(t+1)}$ is:

$$0 = \frac{\partial}{\partial \beta_{v0}} \tilde{Q}(\vartheta_v \mid \vartheta_v^{(t)}) \Big|_{\vartheta_v^{(t+1)}} = n_1 - \sum_{i=0}^{1} \sum_{j=1}^{n_i} \sum_{k=1}^{K_{ij}} \tilde{w}_{ijk}(\vartheta_v^{(t)}) \frac{\exp\left\{\beta_{v0}^{(t+1)} + X(h_{ij}^k, a_{ij})\beta^{(t+1)}\right\}}{1 + \exp\left\{\beta_{v0}^{(t+1)} + X(h_{ij}^k, a_{ij})\beta^{(t+1)}\right\}}. \tag{C.4}$$

The constraint (C.2) evaluated at $\vartheta_v^{(t+1)}$ is:

$$\sum_{i=0}^{1} \sum_{j=1}^{n_i} \sum_{k=1}^{K_{ij}} \mathrm{pr}_v(D_{1j} = 1 \mid H_{1j} = h_{ij}^k, A_{1j} = a_{ij} \,; \beta_{v0}^{(t+1)}, \beta^{(t+1)}) \frac{\tilde{w}_{ijk}(\vartheta_v^{(t)})}{n} = \frac{n_1}{n}$$

or

$$\sum_{i=0}^{1} \sum_{j=1}^{n_i} \sum_{k=1}^{K_{ij}} \tilde{w}_{ijk}(\vartheta_v^{(t)}) \frac{\exp\left\{\beta_{v0}^{(t+1)} + X(h_{ij}^k, a_{ij})\beta^{(t+1)}\right\}}{1 + \exp\left\{\beta_{v0}^{(t+1)} + X(h_{ij}^k, a_{ij})\beta^{(t+1)}\right\}} - n_1 = 0,$$

which is the same as equation (C.4). Hence the unconstrained joint maximizer $\vartheta_v^{(t+1)}$ satisfies the constraint of equation (C.2).

The conclusion is that (i) the maximizer $\beta^{(t+1)}$ with respect to the odds-ratio parameters of interest is obtained by solving score equations from a weighted logistic regression, as though the data were collected prospectively; and (ii) the maximizer $\gamma_v^{(t+1)}$ with respect to γ_v is the empirical distribution that puts mass $\tilde{w}_{ijk}(\vartheta_v^{(t)})/n$ at each (h_{ij}^k, a_{ij}).

Bibliography

[1] Breslow NE. Statistics in epidemiology:The case-control study. *Journal of the American Statistical Association*, 91, No 433:14–28, 3 1996.

[2] Breslow NE & Day NE. *Statistical Methods in Cancer Research. Volume I - The Analysis of Case-Control Studies.* International Agency for Research on Cancer (IARC Scientific Publications No. 32, Lyon, United Kingdom, 1980.

[3] Burkett K. Logistic Regression with Missing Haplotypes. Master's thesis, Simon Fraser University, Burnaby, BC, December 2002.

[4] Burkett K, McNeney B, Graham J. A note on inference of trait association with SNP haplotypes and other attributes in generalized linear models. *Human Heredity*, 57:200–206, 6 2004.

[5] Carroll RJ, Wang S, Wang CY. Prospective analysis of logistic case-control studies. *Journal of American Statistical Association*, 90, No. 429:157–169, 1995.

[6] Clayton D, Hills M. *Statistical Models in Epidemiology.* Oxford University Press, New York, 1993.

[7] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[8] Epstein MP. Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics*, 73:1316–1329, 2003.

[9] F. Hoffmann - La Roche Ltd. *Roche Genetics Education Program Interactive CD, version 4.0.0* . Roche Genetics, Switzerland, 2004.

[10] Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*, 80:27–38, 1993.

[11] Horton NJ, Laird NM. Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Resaerch*, 8:37–50, 1998.

[12] Khoury MJ, Beaty TH , Cohen BH. *Fundamentals of Genetic Epidemiology* . Oxford University Press, New York, 1993.

[13] Kleinbaum, DG. *Logistic Regression: A Self-learning Text.* Springer Verlag, New York, 1994.

[14] Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Human Heredity*, 55:56–65, 5 2003.

[15] Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, B 44(2):226–233, 1982.

[16] Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

[17] Stram DO, Leigh Pearce C, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity*, 55:179–190, 1973.

[18] Thompson EA. *Statistical Inference From Genetic Data on Pedigrees.* Institute of Mathematical Statistics, Beachwood, Ohio, 2000.

[19] van der Vaart AW. *Asymptotic Statistics.* Cambridge University Press, New York, 1998.

[20] Zhao LP, Li SS, Khalid N. A method for the assessment of disease association with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *The American Society of Human Genetics*, 72:1231–1250, 2003.