A GRAPHICAL TOOL FOR EXPLORING SNP-BY-ENVIRONMENT INTERACTION IN CASE-PARENT TRIOS

by

Linnea Duke
B.Sc. Joint Major in Biology and Chemistry,
University of Northern British Columbia, 2002
M.A. Criminology, Simon Fraser University, 2004

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
Spring 2007

APPROVAL

Name:	Linnea Duke
Degree:	Master of Science
Title of project:	A Graphical Tool for Exploring SNP-by-Environment Interaction in Case-Parent Trios

Examining Committee: Dr. Richard Lockhart
Chair

Dr. Jinko Graham
Senior Supervisor
Simon Fraser University

Dr. Brad McNeney
Simon Fraser University

Dr. Denise Daley
External Examiner
University of British Columbia

Date Approved: March 28, 2007

Abstract

We propose a data-smoothing method for exploring statistical interaction between a single nucleotide polymorphism (SNP) and non-genetic risk factors in case-parent trios. Our smoother can be used as a diagnostic tool for checking for the presence or the form of the interaction. The smoother arises from a case-only analysis conditional on parental genotypes. Conditioning on parental genotypes helps to protect against the false impression of interaction that traditional case-only analyses can give when genotypes and non-genetic risk factors are not independent in the population. We discuss the theoretical motivation for the smoother, and illustrate its use with simulated data. We show that the effect of the SNP would have been missed if the interaction suggested by the smoother had not been modelled.

Keywords: age-dependent genetic risks; exploratory data analysis; family-based association studies; gene-by-environment interaction; generalized additive models; single-nucleotide polymorphisms

Subject Terms: Genetics—Statistical methods; Genetic epidemiology—Statistical methods

Dedication

To Robert.

Acknowledgements

Many people were involved in my journey through the Statistics and Actuarial Sciences Department at Simon Fraser University. I will not be able to thank them all enough for their support, nor mention them all.

First and foremost, I would like to thank my supervisor, Dr. Jinko Graham whose support and patience helped me through the "growing pains" of starting in a new discipline. Her time, energy and encouragement during my three years in the MSc. program are greatly appreciated. I would also like to thank my Committee members, Dr. Brad McNeney and Dr. Denise Daley for their time and suggestions for future extensions of this work.

I also want to acknowledge the generosity and kindness of the other members of my research group. They always made me smile and cheered me up when I was blue: Kelly Burkett, Ji-Hyung (Jean) Shin, David Zamar, Zhijian (Charlie) Chen, Mercedeh Ghadessi and Sigal Blay. Thank you!

My research was supported and encouraged by NSERC, the Department of Statistics and Actuarial Science, and the IRMACS facility at Simon Fraser University. I also want to thank Dr. Carl Schwarz for sparking my interest in statistics, Dr. Randy Sitter for his support during that first REALLY scary semester, Dr. Richard Lockhart for helping me grasp the finer points of transformations, Dr. Derek Bingham for his encouragement during Stat 801, Mr. Robin Insley for his support and encouragement and Mr. Ian Bercovitz for introducing me to statistical consulting.

In addition, my fellow graduate students were always there with a smile and friendly word. Thank you: Amy, Carolyn, Chunfang, Cindy, Crystal, Darcy, Darby, Dean, Elizabeth, Eric S., Gurbakhshash, Jason N., John, Laura, Linda, Lucy, Maria, Mark, Matthew, Monica, Paul, Pritam, Ryan, Saman, Simon, Tony and Wendell.

I would also like to thank Sadika Jungic, Charlene Bradbury, Kelly Jay and Sylvia Holmes for their time and dedication.

As always, I am grateful to my parents, Christina and Glenn and my sisters Tamara and Katarina. I would also like to thank David and Ginette Johnstone, Julie Johnstone, Eric Johnstone and Claude Ouellette for their continued support and encouragement. Last but not least, I give my special thanks to Robert Johnstone, for his love, support and encouragement during all my endeavours.

Contents

Appendices

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Genetic Association Studies

Population-based association studies investigate the correlation between genetic variants, such as the various alleles of a gene, and a trait, such as disease status, at the population level (Cardon and Bell 2001). Clayton (2003) proposes three scenarios for explaining the presence of associations at the population level between the genotype at a particular locus and disease status. In the first scenario, the association arises because the locus of interest is the disease-predisposing locus and different genotypes have different levels of risk. In the second scenario, the locus of interest is not the disease-predisposing locus, but is physically linked to and in linkage disequilibrium with the locus of interest. Linkage disequilibrium (LD) is defined as the non-random association of genotypes at different loci (Freeman and Herron 2007), and is synonymous with gametic phase disequilibrium or allelic association (Cardon and Bell 2001). Finally, in the third scenario, the locus of interest is not linked to the disease-predisposing locus and the association is the result of confounding by population admixture or stratification. A "real" association is represented by the first two scenarios. In contrast, population associations that result from population stratification or admixture are of little interest with respect to the investigation of disease aetiology. As a result, it is important that these "spurious" associations are excluded from the study by design, analysis, or both (Clayton 2003).

The main goal of a candidate gene association study is to determine if the gene under study has a direct causal relationship with disease status. However, this goal is seldom possible in association studies involving unrelated individuals (e.g. case-control studies), because such studies are generally not able to distinguish between a gene having a direct causal effect on disease status, and a gene in linkage disequilibrium with the causal gene (Thomas 2004). Association studies alone are not enough to establish a causal link between a genetic variant and a disease; however, according to Thomas (2004), appropriately designed association studies are able to eliminate those associations that result from population stratification.

Ideally, association studies that are conducted on a population basis use well established epidemiological study designs such as case-control or cohort designs. With both cohort and case control study designs, careful attention must be paid to the selection of appropriate controls; for instance, ethnic origin must be carefully controlled through matching of cases and controls in order to avoid detection of spurious associations (Thomas 2004). In contrast, family-based association studies investigate the correlation between genetic variants and trait differences on the basis of the nuclear family. Therefore, the problems usually associated with population substructure are reduced or eliminated through the use of internal (i.e. family) controls (Lazzeroni and Lange 1998).

Family-based study designs are a popular strategy for protecting against spurious associations due to hidden population structure. In the simplest form of a family-based design, genotype information is collected from unrelated cases and their parents. Additional information on non-genetic factors, such as the age of the case may also be collected. In essence, the non-transmitted genotypes of the parents of the case are used as the reference, rather than the genotypes of controls from the general population. The focus of this project will be on case-parent trio data.

Several innovative methods have been proposed for the analysis of case-parent trio data. Spielman et al. (1993) introduced the transmission/disequilibrium test (TDT), based on scoring the transmissions of heterozygous parents to affected children. A convenient feature of the TDT is that it requires no knowledge of the *penetrance* or risk model for the disease. To increase the power to detect linkage and association between a genetic marker and a disease with a variable age of onset, Li and Fan (2000) partially specify the disease penetrance with a Cox proportional-hazards (PH) model for age at onset (Collett 2003). Their PH model leads to a score test for linkage and association that is robust to the form of the underlying baseline hazard rate, under the PH assumption. However, both the TDT and robust score test have limitations as they cannot incorporate age-dependent genetic risks (i.e. statistical interaction between age and the genotype) which imply non-PH. Since we are interested in exploring statistical interaction between age and genotype in our example dataset, we need exploratory and analytical methods that allow for the possibility of interaction. As we will discuss in Chapter 3, conditional logistic regression is the analytic method we chose to follow up on the results of our exploration of statistical interaction. If our smoother suggests that interaction is present, conditional logistic regression allows us to include interaction terms in the risk model.

1.2 Transmission/Disequilibrium Test (TDT)

The TDT is a family-based test for linkage in the presence of association or for association in the presence of linkage (Ewens and Spielman 2003). The TDT was originally proposed as a method for reducing or eliminating the detection of spurious associations resulting from population stratification (Spielman et al. 1993). The TDT achieves this goal by creating an internal control group via conditioning on the observed parental genotypes. Specifically, the marker alleles not transmitted to the affected children become the control set for the marker alleles transmitted to the affected children.

Throughout this discussion, consider a diallelic locus with alleles + and $-$. The null hypothesis for the TDT is a composite null hypothesis of no association or no linkage. The alternative hypothesis is the hypothesis of association and linkage. If the marker locus and the disease locus are not linked or not associated then, with respect to the genotypes of the child at the marker locus, our case-parent trios are just a random sample of trios from the population. In other words, in the absence of linkage or association, the fact that the trios are ascertained through a diseased case should not influence the genotype distribution of the marker in the cases. Thus, under the null hypothesis, the allelic transmissions of heterozygous parents to affected children should have the same distribution

	Non-Transmitted Allele		
Transmitted Allele	+	−	Total
+	n_{11}	n_{12}	$n_{1.}$
−	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$2n$

Table 1.1: Contingency table for the transmission/disequilibrium test.

as allelic transmissions of heterozygous parents to a random sample of children.

Table 1.1 summarizes all the allelic transmissions from parents to cases. The n_{11} and n_{22} entries of Table 1.1 represent transmissions from homozygous parents, and as a result, these transmissions can be discarded from the analysis because they are not informative. Such transmissions are not informative because there is no variation in what they transmit; homozygous parents are only able to transmit the allele for which they are homozygous. In contrast, the transmissions from heterozygous parents are informative because they can transmit either one of their two available alleles. According to Mendelian law, each allelic transmission from a heterozygote parent to a random child may be considered to be an independent Bernoulli trial with $p = 0.5$. Under the composite null hypothesis of no linkage or no association, the proband may be taken to be a random child. Under the more specific (restricted) null hypothesis of no linkage, all affected children within a family (except for monozygotic twins) may be taken to be random children. As illustrated in Table 1.1, there is a total of $n_{12} + n_{21}$ cases with heterozygous parents. Hence, under the null, n_{12} is binomially distributed with probability of success equal to 0.5, mean $\frac{(n_{12} + n_{21})}{2}$, and variance $\frac{(n_{12} + n_{21})}{4}$. Thus, the null hypothesis will be rejected if n_{12} differs too much from Mendelian expectations. The binomial test statistic can be standardized by centering with its mean and scaling with its standard deviation under the null hypothesis. This process produces a z-score, which using the normal approximation to the binomial, is asymptotically normal under the null hypothesis. Furthermore, squaring the z-score produces a test statistic that is asymptotically χ^2 with one degree of freedom. Therefore, the binomial test statistic can be written as:

$$TDT = \frac{(n_{12} - \frac{n_{12}+n_{21}}{2})^2}{(\frac{n_{12}+n_{21}}{4})} = \frac{(n_{12} - E[n_{12}])^2}{Var[n_{12}]}, \tag{1.1}$$

which after simplification, reduces to

$$TDT = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}.$$

However, since we are using a continuous distribution (i.e. normal) to approximate a discrete distribution (i.e. binomial), a continuity correction can improve the accuracy of the approximation. After such a continuity correction, the transmission/disequilibrium statistic becomes:

$$TDT = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}}. \tag{1.2}$$

1.3 A Log-Linear Model of Disease Risk

Self et al. (1991) propose a likelihood approach to the analysis of case-parent trio data. They begin with a statistical model that characterizes the multiplicative factor by which the risk of developing disease in individuals with one set of covariate values differs from the risk of developing disease in individuals with a different set of covariate values. With its focus on describing relative risk, their model takes the traditional epidemiologic approach to discovering the factors underlying disease aetiology. Disease risks depend only on the child's unordered genotype G_c and covariate values X and are log-linear in the sense that

$$P(D = 1|G_c = g, X = x) = k\exp(z(g)\beta + h(x)\gamma) \tag{1.3}$$

or

$$\log(P(D = 1 \mid G_c = g, X = x)) = \log(k) + z(g)\beta + h(x)\gamma,$$

where $D = 1$ if the child develops the disease during the period of the disease incidence study and is 0 otherwise, k is a proportionality constant, β and γ are unknown coefficients estimated from regression analysis, $h(x)$ is a user-specified function of x, and $z(g)$ codes the child's unordered genotype. For instance, two copies of the $-$ allele could be coded as 2, one copy of the $-$ allele could be coded as 1 and zero copies of the $-$ allele could be coded as 0. This coding method implies that risk changes by a multiplicative factor $\exp(\beta)$ with each additional copy of the minor allele. Therefore, if the values of X are fixed, the risk for heterozygotes is $\exp(\beta)$ times that for individuals homozygous for the major allele, and the risk for individuals that are homozygous for the minor allele is $\exp(\beta)$ times that for heterozygotes. One further point of interest regarding the risk model is that it does not include terms for statistical interaction between the SNP and the non-genetic risk factor. Moreover, if the non-genetic covariate X is age and the follow-up time for the disease-incidence study is short, the risk model (1.3) is well approximated by the Cox proportional hazards model for disease age-at-onset proposed by Li and Fan (2000), with baseline hazard rate $\lambda_0(x)$ proportional to $\exp(h(x)\gamma)$. To see why, suppose the study period is of length dx and note that $P(D = 1 \mid G = g, X = x)$ is the probability that a disease-free subject of age x with genotype g at the beginning of the study develops the disease by the end of the study, or

$$P(D = 1 \mid G = g, X = x) = P(\text{disease in } (x, x + dx) \mid \text{disease free at age } x, G = g).$$

The hazard rate, $\lambda_g(x)$, for a disease-free subject of age x with genotype g is their instantaneous probability of developing the disease, or

$$\lambda_g(x) = \lim_{dx \downarrow 0} \frac{P(\text{disease in } (x, x + dx) \mid \text{disease free at age } x, G = g)}{dx}$$

Thus, for a short study period of length dx,

$$\frac{P(D = 1 \mid G = g, X = x)}{dx} \approx \lambda_g(x). \tag{1.4}$$

Substituting the right-hand side of expression 1.3 for $P(D = 1 \mid G = g, X = x)$ in expression 1.4 gives:

$$\frac{k \exp\left(z(g)\beta + h(x)\gamma\right)}{dx} \approx \lambda_g(x). \tag{1.5}$$

Comparing equation 1.5 to the Cox proportional hazards model

$$\lambda_g(x) = \lambda_0(x) \exp\left(z(g)\beta\right),$$

we see that the baseline hazard rate $\lambda_0(x) \approx \dfrac{k}{dx}\exp(h(x)\gamma)$. Finally, noting that k and dx are both constants with respect to x, we obtain the result that $\lambda_0(x)$ is approximately proportional to $\exp(h(x)\gamma)$.

Although Self et al. (1991) use a regression approach, their risk model is not a logistic regression model. In a logistic regression model, the *log-odds* of disease would be modelled as linear in the predictors. In contrast, the log-linear risk model used by Self et al. (1991) models the *log-risk* of disease as linear in the predictors. However, as will be shown in Chapter 2, the likelihood for the Self et al. (1991) model, and hence, the likelihood for our extension of the Self et al. (1991) model, is very similar to that of a conditional logistic regression model for matched case-control data (Hastie and Tibshirani 1990). As a result, we can exploit statistical software, such as the **clogit()** function in the R **survival** package, that is readily available for conditional logistic regression analysis.

The statistical model proposed by Self et al. (1991) was developed for estimating the relative risk of disease given a particular Human Leukocyte Antigen (HLA) haplotype. The Human Leukocyte Antigen complex or major histocompatibility complex (MHC) is the name given to a group of genes located on human chromosome six. The HLA genes code for cell-surface antigen-presenting proteins that are critical to the functioning of the immune system, and as a result are highly polymorphic. Since Self et al. (1991) are concerned with the risk associated with certain haplotypes in the HLA region of the genome, their risk model is based on the assumption of an infinitely polymorphic marker, where everyone is heterozygous for different alleles (haplotypes). In contrast, the risk model that we develop in Chapter 2 is based on a diallelic marker. The implications of this distinction will be discussed more fully in Chapter 2.

1.4 Generalized Additive Models

The following description of generalized additive models is based on work by Hastie and Tibshirani (1990). Consider the simplest case of the linear regression model, with n measurements of the response variable Y and n measurements of a single predictor variable,

X:

$$E(Y \mid X) = \alpha + X\beta, \tag{1.6}$$

where α and the β are regression parameters that must be estimated from the data. The goal of the linear regression model is to describe the dependence of $E(Y \mid X)$, as a linear function of the predictor, X. As long as the dependence of $E(Y \mid X)$ is linear, model 1.6 is extremely useful. When $E(Y \mid X)$ does not depend linearly on X, a common "fix" is to incorporate non-linear terms for the predictor into the linear regression model. For example, if the dependence of $E(Y)$ on X is believed to be quadratic, then an X^2 term can be added so that

$$E(Y \mid X) = \alpha + X\beta_1 + X^2\beta_2.$$

As before, α, β_1 and β_2 are regression parameters that must be estimated from the data. However, deciding what polynomial terms to include in the model can be difficult as the dependence of $E(Y \mid X)$ on X is often not easy to determine by simple inspection of a scatterplot of the data. Instead, smoothers may be used to allow the data itself to reveal the functional form of $E(Y \mid X)$.

Scatterplot smoothers are extremely useful because they reveal the true functional form of the data without imposing a rigid parametric model on the dependence between the response variable and the predictor variable. Since we are interested in describing the dependence between $E(Y \mid X)$ and X as flexibly as possible, the linear regression model 1.6 can be generalized as follows:

$$E(Y \mid X) = \alpha + f(X), \tag{1.7}$$

where $f(X)$ is an unspecified function.

By definition, a *smoother* is a tool that visually summarizes the dependence between the response variable Y and the predictor X by producing an estimate of the dependence that is less variable than the response itself. The visual summary of the dependence produced by a smoother is called a *smooth*, and therefore, the visual estimate of $f(X)$ from equation 1.7 is referred to as a smooth. Commonly used smoothers include the running mean (i.e. moving average) smoother and the loess (i.e. local polynomial regression fitting) smoother. Most smoothers employ local averaging, where a subset of the values of the response Y with associated values of the predictor X close to some target value of the predictor are used to produce an estimate of the mean response at the target value. We refer to a *neighbourhood* as the subset of (X, Y) pairs with predictor values close to the target value. Therefore, two obvious questions are: (i) How big should the neighbourhoods be? and (ii) How should the responses in each neighbourhood be averaged?

The main difference between types of smoothers is their method of averaging the response values within each neighbourhood. For instance, the running mean smoother gives equal weight to each value in the neighbourhood. In contrast, in loess, the response values in the neighbourhood are weighted according to their distance from the target value using a *tri-cube weight function* (Cleveland 1979). The size of the neighbourhood used for the smoothing is typically represented in the form of an adjustable *smoothing parameter.* For example, in loess, the smoothing parameter indicates the fraction of the data to include in the neighbourhood, and can take any value between zero and one. The smoothing parameter in loess is referred to as the *span.* Other smoothers, known as smoothing splines rely on a smoothing parameter referred to as the *target equivalent degrees of freedom*, but we will not consider these here.

A trade-off between bias and variance exists with respect to the chosen size of the neighbourhoods. Large neighbourhoods can produce estimates of $E(Y \mid X)$ with low variance but high bias and small neighbourhoods can produce estimates with small bias but high variance. Intuitively, this means that, in general, the larger the neighbourhoods the "smoother" the estimate of the dependence between $E(Y \mid X)$ and X.

Typically, we have more than one predictor variable of interest. Thus, model 1.7 can be extended to multiple predictors as follows:

$$E(Y \mid X_1, X_2) = \alpha + f_1(X_1) + f_2(X_2), \tag{1.8}$$

where $f(X_1)$ and $f(X_2)$ are both unspecified functions that can be estimated iteratively using a *backfitting* algorithm. For instance, given an estimate, $\hat{f}_1(X_1)$, of $f_1(X_1)$:

1. Estimate $f_2(X_2)$ by smoothing the residual $Y - \hat{f}_1(X_1)$ on X_2.

2. Refine the estimate of $f_1(X_1)$ by smoothing $Y - \hat{f}_2(X_2)$ on X_1.

3. Continue this process until the estimates of $f_2(X_2)$ and $f_1(X_1)$ no longer change between iterations.

The models discussed above (Equations 1.6, 1.7 and 1.8) are referred to as *additive* models. They are additive models because of the underlying assumption that the predictor variables are additive in their effects. In other words, once the model has been fit, we can examine the marginal effects of the predictors separately. As a result, additive models are approximations of the true response surface.

Equations 1.7 and 1.8 are both generalizations of the familiar linear regression model (1.6) because they do not impose a rigid parametric structure on the relationship between the response variable and the predictors. We can generalize these additive models further to accommodate non-normal responses as:

$$g\left(E(Y \mid X)\right) = \alpha + \sum_{j=1}^{p} f_j(x_j)$$

where g is the "link" function in a generalized linear model (McCullagh and Nelder 1989), $E\left(f_j(x_j)\right) = 0$ and the f_j's are univariate smooth functions, one for each of the p predictors. Scatterplot smoothers similar to those discussed above can be used to estimate each of the f_j's, and after the functional form of each of the f_j's has been determined, standard generalized linear modelling incorporating the appropriate terms for the predictors (e.g. polynomial terms) can be undertaken.

1.5 Thesis Overview

In this thesis, we propose a data-smoothing method for exploring statistical interaction between a SNP and a non-genetic risk factor in case-parent trios. Chapter 1 provides background information on genetic association studies and generalized additive models. Chapter 2 develops the smoothing approach and Chapter 3 applies the smoothing approach to a simulated dataset. Chapter 3 also follows up on the exploratory findings with likelihood-based tests. Finally, Chapter 4 summarizes the important results and discusses the possibilities for future research.

Chapter 2

Methods

This chapter describes the notation used throughout and also presents the statistical motivation for the likelihood-based smoother.

2.1 Notation

Let D and X be the case's disease status and non-genetic covariate values, respectively. In this study design, the disease status for the child (case) is $D = 1$. Also, let $G_p = ((M_1, M_2), (F_1, F_2))$ be the parental genotypes, where (M_1, M_2) and (F_1, F_2) denote, respectively, the unordered genotypes of the mother and father with respect to grandparental origin. There are four possible combinations $(M_1, F_1), (M_1, F_2), (M_2, F_1)$ and (M_2, F_2) of parental transmissions that could lead to the child. Following the notation of Self et al. (1991), we call this set of four possible combinations $*$. Further, let T_c be the specific allele combination inherited by the child. We assume that all parental genotypes are available.

Self et al. (1991) assumed an infinitely polymorphic marker rather than a diallelic marker, and so in their context T_c, which is the specific allele combination inherited by the child, could be observed. Consequently, their likelihoods are based on observed data T_c; in contrast, we have based our likelihoods on observed genotype data for a diallelic, single nucleotide polymorphism (SNP). As a result, we cannot directly observe T_c. With SNPs, only G_c, where G_c is the unordered genotype of the child with respect to parental origin, can be observed. Since parental genotypes G_p and parental transmissions T_c together imply the offspring genotype G_c, in our model we write $G_c = g(G_p, T_c)$.

We are interested in modeling the risk as function of the number of risk alleles that a child carries. In this context, we arbitrarily label the risk allele as the rarer, or minor allele, and denote it with a $-$ sign. The major, or common, allele is denoted with a $+$ sign. By convention, the major allele is listed first and the minor allele is listed second in heterozygotes. Given a diallelic locus, this means that a child can carry zero, one or two copies of the risk allele. Throughout, we code G_c as $Z(G_c) = (Z_1, Z_2)$ where $(Z_1, Z_2) = (0,0)$ if G_c includes two copies of the $+$ allele, $(Z_1, Z_2) = (1,0)$ if G_c includes one copy of the $+$ allele and one copy of the $-$ allele, and $(Z_1, Z_2) = (1,1)$ if G_c includes two copies of the $-$ allele.

2.2 Risk Model

We use the following generalization of the Self et al. (1991) risk model presented in Chapter 1 to accommodate statistical interaction between the SNP and the non-genetic risk factor:

$$P(D = 1 \mid G_c = g_c, X = x) = \exp\left(k + z(g_c)\beta + h(x) + z(g_c)f(x)\right). \qquad (2.1)$$

In our generalization, k is the log-risk of disease in individuals with $z(g_c) = (0,0)$ and $h(x)$ $= 0$. In this model, β is a vector of regression parameters, with $\beta = (\beta_1, \beta_2)$ and $h(x)$ is the value of an unspecified function h when $X = x$. The function, $f(x) = (f_1(x), f_2(x))^T$, is a vector with unspecified scalar elements $f_1(x)$ and $f_2(x)$, where f_1 and f_2 are smooth functions.

Under this model, the risk for heterozygotes, at a fixed value of the non-genetic covariates, $X = x$, is $\exp(\beta_1 + f_1(x))$ times that of homozygotes for the major allele. Comparatively, the risk for individuals carrying two copies of the minor allele is $\exp(\beta_2 + f_2(x))$ times that of heterozygous individuals. Thus, under this model, both of the genotype relative risks can vary across the levels of the non-genetic covariate, x. In this sense, the model allows for *statistical interaction* between G_c and X. When $f_1(x) = f_2(x) \equiv 0$ for all x, there is no interaction. We have deliberately left $f = (f_1, f_2)^T$ unspecified. In this project, we propose to *let the data tell us its form* through exploratory plots that we will motivate in the next section.

2.3 Likelihoods Based on Transmission Data

If we assume that (i) G_c and X are conditionally independent given G_p, and (ii) there is Mendelian segregation, then after conditioning on the genotype information of the parents, the disease status of the child, and the non-genetic covariate values of the child, we obtain the following result:

$$P(T_c = t \mid G_p = g_p, X = x, D = 1) = \frac{\exp\left(z(g_c)\beta + z(g_c)f(x)\right)}{\sum_* \exp\left(z(g_*)\beta + z(g_*)f(x)\right)}. \tag{2.2}$$

In equation 2.2, $g_c = g(g_p, t)$ and $g_* = g(g_p, t_*)$, and the sum in the denominator is over the four possible values of T_c. We now derive equation 2.2. From the definition of conditional probability, the probability shown on the left-hand side of equation 2.2 can be expanded as follows:

$$P(T_c = t \mid G_p = g_p, X = x, D = 1) = \frac{P(T_c = t, G_p = g_p, X = x, D = 1)}{P(G_p = g_p, X = x, D = 1)}.$$

Since $\dot{\bigcup_*}\{T_c = t_*\} = \Omega$, the sample space, the event $\{G_p = g_p\}$ can be partitioned as

$$\{G_p = g_p\} = \dot{\bigcup_*} \{T_c = t_*, G_p = g_p\}, \tag{2.3}$$

so that

$$P(G_p = g_p, X = x, D = 1) = \sum_* P(T_c = t_*, G_p = g_p, X = x, D = 1).$$

We can therefore re-write

$$\begin{aligned}
&P(T_c = t \mid G_p = g_p, X = x, D = 1) \\
&= \frac{P(T_c = t, G_p = g_p, X = x, D = 1)}{\sum_* P(T_c = t_*, G_p = g_p, X = x, D = 1)} \\
&= \frac{P(D = 1 \mid T_c = t, G_p = g_p, X = x)P(T_c = t, G_p = g_p, X = x)}{\sum_* P(D = 1 \mid T_c = t_*, G_p = g_p, X = x)P(T_c = t_*, G_p = g_p, X = x)} \\
&= \frac{P(D = 1 \mid T_c = t, G_p = g_p, X = x)P(T_c = t \mid G_p = g_p, X = x)P(G_p = g_p, X = x)}{\sum_* P(D = 1 \mid T_c = t_*, G_p = g_p, X = x)P(T_c = t_* \mid G_p = g_p, X = x)P(G_p = g_p, X = x)} \\
&= \frac{P(D = 1 \mid T_c = t, G_p = g_p, X = x)P(T_c = t \mid G_p = g_p, X = x)}{\sum_* P(D = 1 \mid T_c = t_*, G_p = g_p, X = x)P(T_c = t_* \mid G_p = g_p, X = x)}.
\end{aligned} \tag{2.4}$$

Since T_c and X are assumed to be conditionally independent given G_p, equation 2.4 becomes

$$\frac{P(D = 1 \mid T_c = t, G_p = g_p, X = x)P(T_c = t \mid G_p = g_p)}{\sum_* P(D = 1 \mid T_c = t_*, G_p = g_p, X = x)P(T_c = t_* \mid G_p = g_p)}. \tag{2.5}$$

Under the assumption of Mendelian segregation of the SNP in the population, the above equation simplifies further to

$$P(T_c = t \mid G_p = g_p, X = x, D = 1) = \frac{P(D = 1 \mid T_c = t, G_p = g_p, X = x)}{\sum_* P(D = 1 \mid T_c = t_*, G_p = g_p, X = x)}.$$

As noted earlier in Section 2.1, the parental genotypes, G_p, along with their transmissions T_c to the child imply the child's unordered genotype G_c. Therefore, we end up with

$$P(T_c = t \mid G_p = g_p, X = x, D = 1) = \frac{P(D = 1 \mid G_c = g(g_p, t), X = x)}{\sum_* P(D = 1 \mid G_c = g(g_p, t_*), X = x)}. \quad (2.6)$$

Substituting equation 2.1 into equation 2.6 we obtain:

$$\begin{aligned} & P(T_c = t \mid G_p = g_p, X = x, D = 1) \\ = & \frac{\exp\left(k + z(g_c)\beta + h(x) + z(g_c)f(x)\right)}{\sum_* \exp\left(k + z(g_*)\beta + h(x) + z(g_*)f(x)\right)} \\ = & \frac{\exp(k)\exp\left(h(x)\right)\exp(z(g_c)\beta + z(g_c)f(x))}{\exp(k)\exp\left((h(x)\right)\sum_* \exp\left(z(g_*)\beta + z(g_*)f(x)\right)} \\ = & \frac{\exp\left(z(g_c)\beta + z(g_c)f(x)\right)}{\sum_* \exp\left(z(g_*)\beta + z(g_*)f(x)\right)}, \end{aligned}$$

which is the same expression as that presented in equation 2.2. Therefore, if we had an infinitely polymorphic genetic marker, so that the transmissions were always observable, the likelihood allowing for statistical interaction between G_c and X would be:

$$L(\beta, f) = \prod_{i=1}^{N} \frac{\exp\left(z(g_c)\beta + z(g_c)f(x)\right)}{\sum_* \exp\left(z(g_*)\beta + z(g_*)f(x)\right)}, \quad (2.7)$$

where N is the number of case-parent trios. Equation 2.7 could then used to derive the maximum-likelihood estimators for the genetic risk parameters β and f.

2.4 Likelihoods Based on Genotype Data

In the case-parent study design, the observed data consists of the genotype of the case, G_c, the genotypes of the parents, G_p and the non-genetic covariate values, X of the case. However, by conditioning on G_p and X we avoid the estimation of nuisance parameters, those parameters other than β_1, β_2, f_1 and f_2 in the risk model (2.1). Equation 2.2 provides a way to calculate the probability of obtaining G_c, conditioned on G_p and X. However, the expression $P(G_c = g_c \mid G_p = g_p, X = x, D = 1)$ represents what we will call the *unconditional* likelihood contributions of a trio with $G_p = g_p$ and $X = x$. The motivation for our smoother is to extract information about (β_1, f_1) by conditioning further on $\{G_c = ++$ or $+-\} \equiv \{Z_2 = 0\}$, and to extract information about (β_2, f_2) by conditioning further on $\{G_c = --$ or $+-\} \equiv \{Z_1 = 1\}$. With this in mind, we refer to the probabilities, $P(G_c = g_c \mid Z_2 = 0, G_p = g_p, X = x, D = 1)$ and $P(G_c = g_c \mid Z_1 = 1, G_p = g_p, X = x, D = 1)$ as the *conditional* likelihood contributions of a trio with $G_p = g_p$ and $X = x$. These conditional likelihood contributions provide the motivation for the exploratory plots presented in Chapter 3. However, we will first consider the unconditional likelihoods as they are required to derive the conditional likelihoods. These conditional likelihoods are discussed in Sections 2.6 and 2.7.

2.4.1 Unconditional Likelihood Tables

Tables 2.1 - 2.3 list the possible outcomes for T_c, the corresponding outcomes G_c, the codings $Z(G_c) = (Z_1, Z_2)$, and the probabilities associated with each outcome of T_c, as calculated using equation 2.2. These probabilities are first calculated up to a constant of proportionality, and then exactly, for each parental mating type. Of the six possible unordered parental mating types from a diallelic locus, only three lead to variation in G_c, and are therefore considered *informative*:

1. $G_p = (++, +-)$ or $(+-, ++)$,
2. $G_p = (--, +-)$ or $(+-, --)$, and
3. $G_p = (+-, +-)$.

G_p	T_c	G_c	(Z_1, Z_2)	Numerator of Equation 2.2	$P(T_c \mid G_p, X, D)$
$(++,+-)$	M_1F_1	$++$	$(0,0)$	1	$\frac{1}{2+2\exp(\beta_1+f_1(x))}$
	M_1F_2	$+-$	$(1,0)$	$\exp(\beta_1+f_1(x))$	$\frac{\exp(\beta_1+f_1(x))}{2+2\exp(\beta_1+f_1(x))}$
	M_2F_1	$++$	$(0,0)$	1	$\frac{1}{2+2\exp(\beta_1+f_1(x))}$
	M_2F_2	$+-$	$(1,0)$	$\exp(\beta_1+f_1(x))$	$\frac{\exp(\beta_1+f_1(x))}{2+2\exp(\beta_1+f_1(x))}$
$(+-,++)$	M_1F_1	$++$	$(0,0)$	1	$\frac{1}{2+2\exp(\beta_1+f_1(x))}$
	M_1F_2	$++$	$(0,0)$	1	$\frac{1}{2+2\exp(\beta_1+f_1(x))}$
	M_2F_1	$+-$	$(1,0)$	$\exp(\beta_1+f_1(x))$	$\frac{\exp(\beta_1+f_1(x))}{2+2\exp(\beta_1+f_1(x))}$
	M_2F_2	$+-$	$(1,0)$	$\exp(\beta_1+f_1(x))$	$\frac{\exp(\beta_1+f_1(x))}{2+2\exp(\beta_1+f_1(x))}$

Table 2.1: Unconditional likelihood information for first parental mating type, $G_p = (++,+-)$ or $(+-,++)$

G_p	T_c	G_c	(Z_1, Z_2)	Numerator of Equation 2.2	$P(T_c \mid G_p, X, D)$
$(--,+-)$	M_1F_1	$+-$	$(1,0)$	$\exp(\beta_1 + f_1(x))$	$\frac{1}{2+2\exp(\beta_2+f_2(x))}$
	M_1F_2	$--$	$(1,1)$	$\exp(\beta_1+\beta_2+f_1(x)+f_2(x))$	$\frac{\exp(\beta_2+f_2(x))}{2+2\exp(\beta_2+f_2(x))}$
	M_2F_1	$+-$	$(1,0)$	$\exp(\beta_1 + f_1(x))$	$\frac{1}{2+2\exp(\beta_2+f_2(x))}$
	M_2F_2	$--$	$(1,1)$	$\exp(\beta_1+\beta_2+f_1(x)+f_2(x))$	$\frac{\exp(\beta_2+f_2(x))}{2+2\exp(\beta_2+f_2(x))}$
$(+-,--)$	M_1F_1	$+-$	$(1,0)$	$\exp(\beta_1 + f_1(x))$	$\frac{1}{2+2\exp(\beta_2+f_2(x))}$
	M_1F_2	$+-$	$(1,0)$	$\exp(\beta_1 + f_1(x))$	$\frac{1}{2+2\exp(\beta_2+f_2(x))}$
	M_2F_1	$--$	$(1,1)$	$\exp(\beta_1+\beta_2+f_1(x)+f_2(x))$	$\frac{\exp(\beta_2+f_2(x))}{2+2\exp(\beta_2+f_2(x))}$
	M_2F_2	$--$	$(1,1)$	$\exp(\beta_1+\beta_2+f_1(x)+f_2(x))$	$\frac{\exp(\beta_2+f_2(x))}{2+2\exp(\beta_2+f_2(x))}$

Table 2.2: Unconditional likelihood information for second parental mating type, $G_p = (--,+-)$ or $(+-,--)$

G_p	T_c	G_c	(Z_1, Z_2)	Numerator of Equation 2.2	$P(T_c \mid G_p, X, D)$
$(+-,+-)$	M_1F_1	$++$	$(0,0)$	1	$\frac{1}{1+2\exp(\beta_1+f_1(x))+\exp(\beta_1+\beta_2+f_1(x)+f_2(x))}$
	M_1F_2	$+-$	$(1,0)$	$\exp(\beta_1+f_1(x))$	$\frac{\exp(\beta_1+f_1(x))}{1+2\exp(\beta_1+f_1(x))+\exp(\beta_1+\beta_2+f_1(x)+f_2(x))}$
	M_2F_1	$+-$	$(1,0)$	$\exp(\beta_1+f_1(x))$	$\frac{\exp(\beta_1+f_1(x))}{1+2\exp(\beta_1+f_1(x))+\exp(\beta_1+\beta_2+f_1(x)+f_2(x))}$
	M_2F_2	$--$	$(1,1)$	$\exp(\beta_1+\beta_2+f_1(x)+f_2(x))$	$\frac{\exp(\beta_1+\beta_2+f_1(x)+f_2(x))}{1+2\exp(\beta_1+f_1(x))+\exp(\beta_1+\beta_2+f_1(x)+f_2(x))}$

Table 2.3: Unconditional likelihood information for third parental mating type, $G_p = (+-,+-)$

The information presented in Tables 2.1 - 2.3 can be used to calculate the probability of obtaining a case with a particular genotype from a particular parental mating type. For example, if we restrict our attention to the first parental mating type, and if $G_p = (++, +-)$, the probability that we observe a case with genotype $G_c = (+-)$ can be calculated using the following information:

1. $\{T_c = (M_1, F_2)\} \Rightarrow \{G_c = +-\}$, and $\{T_c = (M_2, F_2)\} \Rightarrow \{G_c = +-\}$.
2. $\{G_c = +-\} \Leftrightarrow \{T_c = (M_1, F_2) \quad \dot{\cup} \quad T_c = (M_2, F_2)\}$.
3. $P(G_c = +-) = P\left(T_c = (M_1, F_2)\right) + P\left(T_c = (M_2, F_2)\right)$.

Thus, using the information provided in Table 2.1, $G_c = (+-)$ with probability

$$\frac{2\exp\left(\beta_1 + f_1(x)\right)}{2 + 2\exp\left(\beta_1 + f_1(x)\right)},$$

which simplifies to

$$\frac{\exp\left(\beta_1 + f_1(x)\right)}{1 + \exp\left(\beta_1 + f_1(x)\right)}.$$

The probability of observing the other possible levels of G_c can be calculated analogously using the appropriate information from Tables 2.1 - 2.3.

The exact probabilities $P(T_c = t \mid G_p = g_p, X = x, D = 1)$ for each of the three informative parental mating types are obtained by applying equation 2.2. Specifically, we divide the numerator of equation 2.2 for the specific T_c value of interest by the sum over all possible outcomes of T_c for that mating type. For example, consider the second row of Table 2.2. Given that $G_p = (--, +-)$, then for $T_c = (M_1, F_2)$, the entry for (Z_1, Z_2) must be (1,1). Consequently, the numerator of equation 2.2 is

$$\begin{aligned} &\exp\left(z_1\beta_1 + z_2\beta_2 + z_1 f_1(x) + z_2 f_2(x)\right) \\ &= \exp\left(1\beta_1 + 1\beta_2 + 1f_1(x) + 1f_2(x)\right) \\ &= \exp\left(\beta_1 + \beta_2 + f_1(x) + f_2(x)\right) \end{aligned}$$

The numerator of equation 2.2 is similarly obtained for the other three outcomes of T_c. Summing the numerators for all four possible outcomes of T_c, the denominator of equation

2.2 is therefore:

$$\begin{aligned} & \exp\left(\beta_1 + f_1(x)\right) + \exp\left(\beta_1 + \beta_2 + f_1(x) + f_2(x)\right) \\ + \quad & \exp\left(\beta_1 + f_1(x)\right) + \exp\left(\beta_1 + \beta_2 + f_1(x) + f_2(x)\right) \\ = \quad & 2\exp\left(\beta_1 + f_1(x)\right) + 2\exp\left(\beta_1 + \beta_2 + f_1(x) + f_2(x)\right). \end{aligned}$$

Dividing the numerator by the denominator, we obtain

$$\begin{aligned} & P(T_c = (M_1, F_2) \mid G_p = (--, +-), X = x, D = 1) \\ = \quad & \frac{\exp\left(\beta_1 + \beta_2 + f_1(x) + f_2(x)\right)}{2\exp\left(\beta_1 + f_1(x)\right) + 2\exp\left(\beta_1 + \beta_2 + f_1(x) + f_2(x)\right)} \\ = \quad & \frac{\exp\left(\beta_2 + f_2(x)\right)}{2 + 2\exp\left(\beta_2 + f_2(x)\right)}, \end{aligned}$$

as specified in the entry for $P(T_c \mid G_p, X, D)$ in the second row of Table 2.2.

2.5 Motivation for Smoothing

Conditioning on $Z_2 = 0$ and $Z_1 = 1$ motivates a smoothing strategy for exploring the functional form of f_1 and f_2, respectively. Specifically, we propose to fit two generalized additive logistic models

1. Z_1 on X, with an intercept term, in which the trios from the first mating type of $G_p = (++, +-)$, or $(+-, ++)$ have a zero offset and trios from the third mating type of $G_p = (+-, +-)$ have an offset of $\log(2)$.

2. Z_2 on X with an intercept term in which trios from the second mating type of $G_p = (+-, --)$ or $(--, +-)$ have zero offset and trios from the third mating type of $G_p = (+-, +-)$ have offset $-\log(2)$.

The resulting smoothed values of f_1 and f_2 will provide insight into the form of the statistical interaction between the child's genotype and non-genetic covariates. The *offset* referred to above is simply a predictor with a fixed coefficient that allows us to specify the correct conditional likelihood (Vittinghoff et al. 2005).

However, conditioning further on $Z_2 = 0$ and $Z_1 = 1$ results in a loss of information. For example, consider the case-parent trios from the third parental mating type, $G_p =$

$(+-,+-)$. By conditioning on the event $Z_2 = 0$, for inference on (β_1, f_1), we exclude case-parent trios with $G_c = (--)$ from the analysis. Without conditioning, trios with $G_p = (+-,+-)$ and $G_c = (--)$ would have contributed information to the likelihood. As a result, not all of the information on f_1 is incorporated into the smooth. Such conditioning is equivalent to combining all trios from the first mating type with trios having $G_c = (++)$ or $(+-)$ from the third mating type. The loss of information resulting from conditioning on $Z_1 = 1$ is analogous to the loss of information resulting from conditioning on $Z_2 = 0$. Conditioning on $Z_1 = 1$ amounts to combining all trios from the second mating type with trios from the third mating type having $G_c = (+-)$ or $(--)$.

2.6 A Smoother for f_1 by Conditioning on $Z_2 = 0$

2.6.1 First Parental Mating Type, $G_p = (++,+-)$ or $(+-,++)$

Using the probabilities presented in Table 2.1 and conditioning on $\{G_c = (++) \cup (+-)\}$, or $\{Z_2 = 0\}$, we obtain

$$\begin{aligned}
&P(G_c = (+-) \mid Z_2 = 0, G_p = (++,+-) \cup (+-,++), X = x, D = 1) \\
&= P(G_c = (+-) \mid G_p = (++,+-) \cup (+-,++), X = x, D = 1) \\
&= \frac{2\exp(\beta_1 + f_1(x))}{2 + 2\exp(\beta_1 + f_1(x))} \\
&= \frac{\exp(\beta_1 + f_1(x))}{1 + \exp(\beta_1 + f_1(x))}
\end{aligned}$$

and

$$\begin{aligned}
&P(G_c = (++) \mid Z_2 = 0, G_p = (++,+-) \cup (+-,++), X = x, D = 1) \\
&= P(G_c = (++) \mid G_p = (++,+-) \cup (+-,++), X = x, D = 1) \\
&= \frac{2}{2 + 2\exp(\beta_1 + f_1(x))} \\
&= \frac{1}{1 + \exp(\beta_1 + f_1(x))}.
\end{aligned}$$

The two values of G_c are synonymous with values of Z_1. Thus, for these trios we can rewrite the conditional likelihood contribution, given $Z_2 = 0$ as

$$\begin{aligned} & P(Z_1 = z_1 \mid Z_2 = 0, G_p = (++,+-) \cup (+-,++), X = x, D = 1) \\ = \;& P(Z_1 = z_1 \mid G_p = (++,+-) \cup (+-,++), X = x, D = 1) \\ = \;& \frac{\exp\left(z_1[\beta_1 + f_1(x)]\right)}{1 + \exp\left(\beta_1 + f_1(x)\right)}. \end{aligned}$$

This expression has the same form as the likelihood contribution from a generalized additive logistic model for Z_1 on X, with β_1 as the intercept. We lose no information from these trios by conditioning on $Z_2 = 0$ as $Z_2 = 0$ for all trios from the first mating type.

2.6.2 Second Parental Mating Type, $G_p = (--,+-)$ or $(+-,--)$

Referring to the probabilities presented in Table 2.2 and conditioning on $Z_2 = 0$, we obtain

$$P(G_c = (+-) \mid Z_2 = 0, G_p = (--,+-) \cup (+-,--), X = x, D = 1) = 1.$$

Therefore, these trios make no contribution to the conditional likelihood for (β_1, f_1). However, these trios would not have contributed any information about these parameters anyway, even if we did not condition on $Z_2 = 0$ because

$$P(G_c = +- \mid G_p = (--,+-) \cup (+-,--), X = x, D = 1) = \frac{1}{1 + \exp\left(\beta_2 + f_2(x)\right)}.$$

Thus, no further information is lost for these trios by conditioning on $Z_2 = 0$.

2.6.3 Third Parental Mating Type, $G_p = (+-,+-)$

Table 2.3 shows that there are two possible outcomes for G_c when $G_p = (+-,+-)$. These two outcomes for G_c correspond to $Z_1 = 0$ or $Z_1 = 1$. By referring to Table 2.3 and conditioning on $Z_2 = 0$, we obtain

$$\begin{aligned}
&P(G_c = (+-) \mid Z_2 = 0, G_p = (+-,+-), X = x, D = 1) \\
&= \frac{P(G_c = +-, Z_2 = 0 \mid G_p = (+-,+-), X = x, D = 1)}{P(Z_2 = 0 \mid G_p = (+-,+-), X = x, D = 1)} \\
&= \frac{P(G_c = (+-) \mid G_p = (+-,+-, X = x, D = 1)}{P(G_c = (++) \cup (+-) \mid G_p = (+-,+-), X = x, D = 1)} \\
&= \frac{P(G_c = (+-) \mid G_p = (+-,+-), X = x, D = 1)}{P(G_c{=}(++) \mid G_p{=}(+-,+-),X{=}x,D{=}1) + P(G_c{=}(+-) \mid G_p{=}(+-,+-),X{=}x,D{=}1)} \\
&= \frac{2\exp(\beta_1 + f_1(x))}{1 + 2\exp(\beta_1 + f_1(x))} \\
&= \frac{\exp(\log(2) + \beta_1 + f_1(x))}{1 + \exp(\log(2) + \beta_1 + f_1(x))}.
\end{aligned}$$

When $Z_2 = 0$, G_c must be either $(++)$ or $(--)$. Hence,

$$\begin{aligned}
&P(G_c = (++) \mid Z_2 = 0, G_p = (+-,+-), X = x, D = 1) \\
&= 1 - P(G_c = (+-) \mid Z_2 = 0, G_p = (+-,+-), X = x, D = 1) \\
&= \frac{1}{1 + \exp(\log(2) + \beta_1 + f_1(x))}.
\end{aligned}$$

Thus, the conditional likelihood contribution of a case-parent trio with $G_p = (+-,+-)$, given $Z_2 = 0$ can be re-written as

$$\begin{aligned}
&P(Z_1 = z_1 \mid Z_2 = 0, G_p = (+-,+-), X = x, D = 1) \\
&= \frac{\exp(z_1[\log(2) + \beta_1 + f_1(x)])}{1 + \exp(\log(2) + \beta_1 + f_1(x))}.
\end{aligned} \tag{2.8}$$

Equation 2.8 is also of the same form as the likelihood contribution from a generalized additive logistic model for Z_1 on X with intercept β_1. However, equation 2.8 contains an offset term, $\log(2)$. For this parental mating type, some information about β_1 and f_1 is lost by conditioning on $Z_2 = 0$. First, as noted in Section 2.5, trios with $G_c = (--)$ that would have contributed to the unconditional likelihood are excluded. Second, even for those trios with $Z_2 = 0$, only partial information about β_1 and f_1 is used. Referring to Table 2.3, we see that without the conditioning, such trios would have contributed

$$\begin{aligned}
&\frac{1}{1 + 2\exp(\beta_1 + f_1(x)) + \exp(\beta_1 + \beta_2 + f_1(x) + f_2(x))} && \text{if } G_c = ++\text{, and} \\
&\frac{2\exp(\beta_1 + f_1(x))}{1 + 2\exp(\beta_1 + f_1(x)) + \exp(\beta_1 + \beta_2 + f_1(x) + f_2(x))} && \text{if } G_c = +-.
\end{aligned}$$

These unconditional likelihood contributions can be combined into a single expression

$$\frac{\exp(z_1[\log 2 + \beta_1 + f_1(x)])}{1 + \exp(\log 2 + \beta_1 + f_1(x)) + \exp(\beta_1 + \beta_2 + f_1(x) + f_2(x))} \tag{2.9}$$

that is analogous (but not equal) to the conditional likelihood contribution given in equation 2.8.

2.7 A Smoother for f_2 by Conditioning on $Z_1 = 1$

To explore the functional form of f_2 we use a similar approach, except this time we condition on $\{G_c = +- \text{ or } --\} \equiv \{Z_1 = 1\}$, rather than on $Z_2 = 0$. The derivation of the conditional likelihoods for the f_2 smoother is analogous to the derivation of the conditional likelihoods for the f_1 smoother. Consequently, we skip the details and present only the final likelihoods.

2.7.1 First Parental Mating Type, $G_p = (++, +-)$ or $(+-, ++)$

Referring to Table 2.1 and conditioning on $Z_1 = 1$, we obtain

$$P(G_c = +- \mid Z_1 = 1, G_p = (++, +-) \cup (+-, ++), X = x, D = 1) = 1.$$

2.7.2 Second Parental Mating Type, $G_p = (+-, --)$ or $(--, +-)$

Referring to Table 2.2 and conditioning on $Z_1 = 1$, we obtain

$$\begin{aligned} P(G_c = (+-) \mid Z_1 = 1, G_p = (--, +-) \cup (+-, --), X = x, D = 1) \\ = \frac{1}{1 + \exp(\beta_2 + f_2(x))}. \end{aligned}$$

and

$$\begin{aligned} P(G_c = -- \mid Z_1 = 1, G_p = (--, +-) \cup (+-, --), X = x, D = 1) \\ = \frac{\exp(\beta_2 + f_2(x))}{1 + \exp(\beta_2 + f_2(x))}. \end{aligned}$$

The outcomes $G_c = (+-)$ and $G_c = (--)$, are synonymous with values of Z_2; therefore, the conditional likelihood contribution of the second mating type, given $Z_1 = 1$ can be written as

$$\frac{\exp(z_2[\beta_2 + f_2(x)])}{1 + \exp(\beta_2 + f_2(x))}.$$

2.7.3 Third Parental Mating Type, $G_p = (+-,+-)$

As shown in Table 2.3, $(Z_1, Z_2) = (0, 0)$ for $G_c = (++)$; therefore,

$$P(G_c = (++) \mid Z_1 = 1, G_p = (--,+-) \cup (+-,--), X = x, D = 1) = 0.$$

As a result, trios with $G_c = (++)$ and parents in the third mating type do not contribute anything to the conditional likelihood. In contrast, by referring to Table 2.3, we obtain

$$\begin{aligned}
&P(G_c = +- \mid Z_1 = 1, G_p = (+-,+-), X = x, D = 1)\\
&= \frac{2\exp(\beta_1 + f_1(x))}{2\exp(\beta_1 + f_1(x)) + \exp(\beta_1 + \beta_2 + f_1(x) + f_2(x))}\\
&= \frac{1}{1 + \frac{1}{2}\exp(\beta_2 + f_2(x))}\\
&= \frac{1}{1 + \exp(-\log(2) + \beta_2 + f_2(x))}.
\end{aligned}$$

When $Z_1 = 1$, G_c must be either $(+-)$ or $(--)$. Hence,

$$\begin{aligned}
&P(G_c = (--) \mid Z_1 = 1, G_p = (+-,+-), X = x, D = 1)\\
&= 1 - P(G_c = (+-) \mid Z_1 = 1, G_p = (+-,+-), X = x, D = 1)\\
&= \frac{\frac{1}{2}\exp(\beta_2 + f_2(x))}{1 + \frac{1}{2}\exp(\beta_2 + f_2(x))}\\
&= \frac{\exp(-\log(2) + \beta_2 + f_2(x))}{1 + \exp(-\log(2) + \beta_2 + f_2(x))}.
\end{aligned}$$

The two values of G_c from the third parental mating type, $G_p = (+-,+-)$ are synonymous with values of Z_2; therefore, we can re-write the conditional likelihood contribution from trios of the third parental mating type, given $Z_1 = 1$ as

$$\begin{aligned}
&P(Z_2 = z_2 \mid Z_1 = 1, G_p = (+-,+-), X = x, D = 1)\\
&= \frac{\exp(z_2[-\log(2) + \beta_2 + f_2(x)])}{1 + \exp(-\log(2) + \beta_2 + f_2(x))}. \qquad (2.10)
\end{aligned}$$

The present chapter discussed the theoretical development of our method for investigating genotype-by-environment interaction in case-parent trios. In the next chapter, we demonstrate how this method can be applied using a simulated dataset.

Chapter 3

Application

This chapter describes the simulation procedure used to generate the simulated dataset, **diabdat**. This chapter also presents descriptive summaries of the simulated dataset. Following the descriptive summaries, we apply our smoothing approach to explore the statistical interaction present in the simulated dataset. Finally, we confirm the results of the exploratory analysis with likelihood-based tests.

3.1 Simulation Procedure

The simulated dataset consists of 600 case-parent trios. Genotypes were coded using a $+$ to represent the major allele, and a $-$ to represent the minor allele. The 600 simulated case-parent trios were randomly sampled from a subdivided population of infinite size, comprised of two non-mixing subpopulations in Hardy-Weinberg equilibrium. The single nucleotide polymorphism (SNP) had a minor allele frequency of 0.3 in the first subpopulation, and 0.5 in the second subpopulation.

From each infinite subpopulation, 1,000,000 individuals were randomly sampled to form 500,000 mating pairs. For each mating pair, one child was generated according to Mendelian segregation probabilities. Each child was then randomly assigned a value for the dietary factor. In the first subpopulation, the distribution of the dietary factor was normal with a mean of 15 and a standard deviation of one. In contrast, the distribution of the dietary factor in the second subpopulation was normal with a mean of ten and standard deviation of one. Therefore, the dietary factor was simulated independently of age and gender, but the dietary factor and SNP genotypes were correlated due to their mutual dependence on subpopulation membership.

The parent-child trios from the two subpopulations were then combined into one large dataset, representing a random sample of 1,000,000 trios from the overall population. Age and gender were then simulated for each child in the dataset. The age distribution for the children was normal with a mean of 40 years and a standard deviation of 15 years, and was simulated independently of the child's SNP genotype, gender and the dietary factor. Although rare, any ages greater than 100 and less than zero were re-distributed uniformly between zero and 20 years. The gender of each child was simulated to be male or female with equal probability. The gender of the children was assigned independently of the child's SNP genotype, age and the dietary factor. Male cases were coded with a one and females were coded with a zero.

The probability of disease in the combined population was set to be $\sim 1.7/1000$, in order to mimic the yearly incidence of type 1 diabetes in northern European populations such as Sweden (Å. Lernmark, personal communication). Given the small probability of disease, a large number of children, and therefore a large number of mating pairs had to be simulated to generate our final dataset of 600 case-parent trios.

The next phase of the simulation was to assign disease status to each of the children present in the data set. The disease status of a child was represented using a zero or one; a disease status of zero represented an unaffected child and a disease status of one represented an affected child. The case genotype information, $G_c = g$, was coded as follows:

1. $z_1(g) = 1$ if $g = (+-)$ or $(--)$, else $z_1(g) = 0$, and
2. $z_2(g) = 1$ if $g = (--)$, else $z_2(g) = 0$.

The log-risk model used to simulate the cases was

$$\begin{aligned}&\log P(D = 1 \mid G_c = g, A = a, S = s)\\ = \;& \log(.03) + \log(1.1)(z_1 + z_2) - \log(1.04)(a - 13)^2\\ + \;& \log(1.01)(a - 13)^2 I(a)s - 0.1(z_1 + z_2)(a - 13), \end{aligned} \tag{3.1}$$

mgeno	dgeno	cgeno	age	sex	diet
$+-$	$+-$	$+-$	11.9	1	15.6
$++$	$+-$	$+-$	20.2	1	14.3
$++$	$+-$	$+-$	13.7	0	10.9
$++$	$+-$	$++$	14.0	1	14.1
$--$	$+-$	$+-$	16.1	0	7.5

Table 3.1: First five rows of the **diabdat** dataset.

where A represents the age of the case in years, and S represents the gender of the case. In equation 3.1, $I(a) = 1$ if $a > 13$ and 0 otherwise and log(.03) is the log-risk in (++) females aged 13 years. Equation 3.1 was used to assign a disease status to each child in the combined dataset. Children with disease status of zero and their parents were eliminated from the dataset. Trios in which both parents were homozygous were also eliminated. Six hundred case-parent trios were selected at random from the remaining trios to produce **diabdat**. Therefore, the resulting dataset consists of the diseased child (case), the genotypes of his/her parents, and the case's age, gender and dietary factor value. Table 3.1 shows the first five rows of **diabdat**. A description of the contents of each column is presented below:

1. *mgeno*, the mother's genotype,

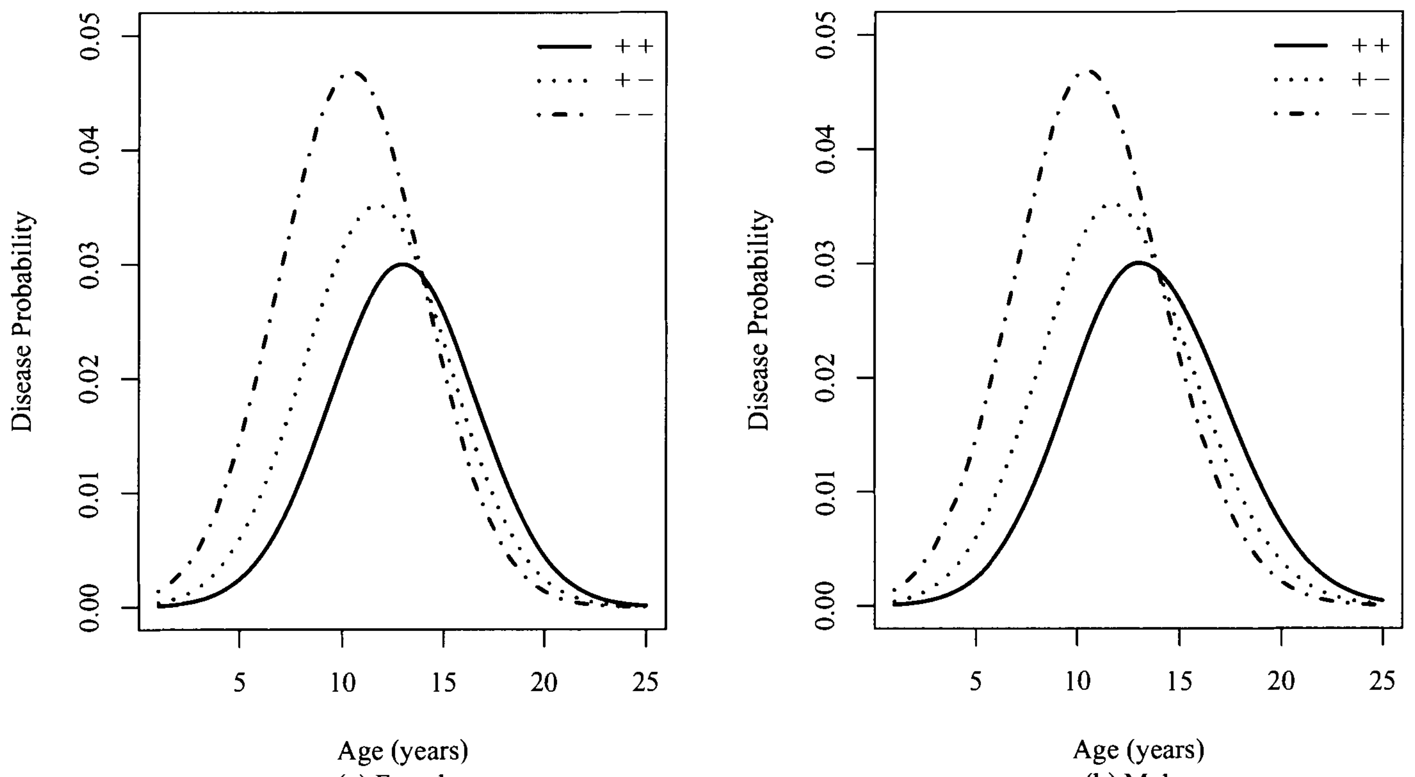
2. *dgeno*, the father's genotype,

3. *cgeno*, the child's (case's) genotype,

4. *age*, the age-of-onset for the child in years,

5. *sex*, the gender of the child, coded as a 1 for males and a 0 for females

6. *diet*, the child's dietary factor value.



(a) Females

(b) Males

Figure 3.1: Theoretical risk curves, as a function of age, by genotype for (a) females, and (b) males. The age range corresponds to that observed for cases in our simulated dataset.

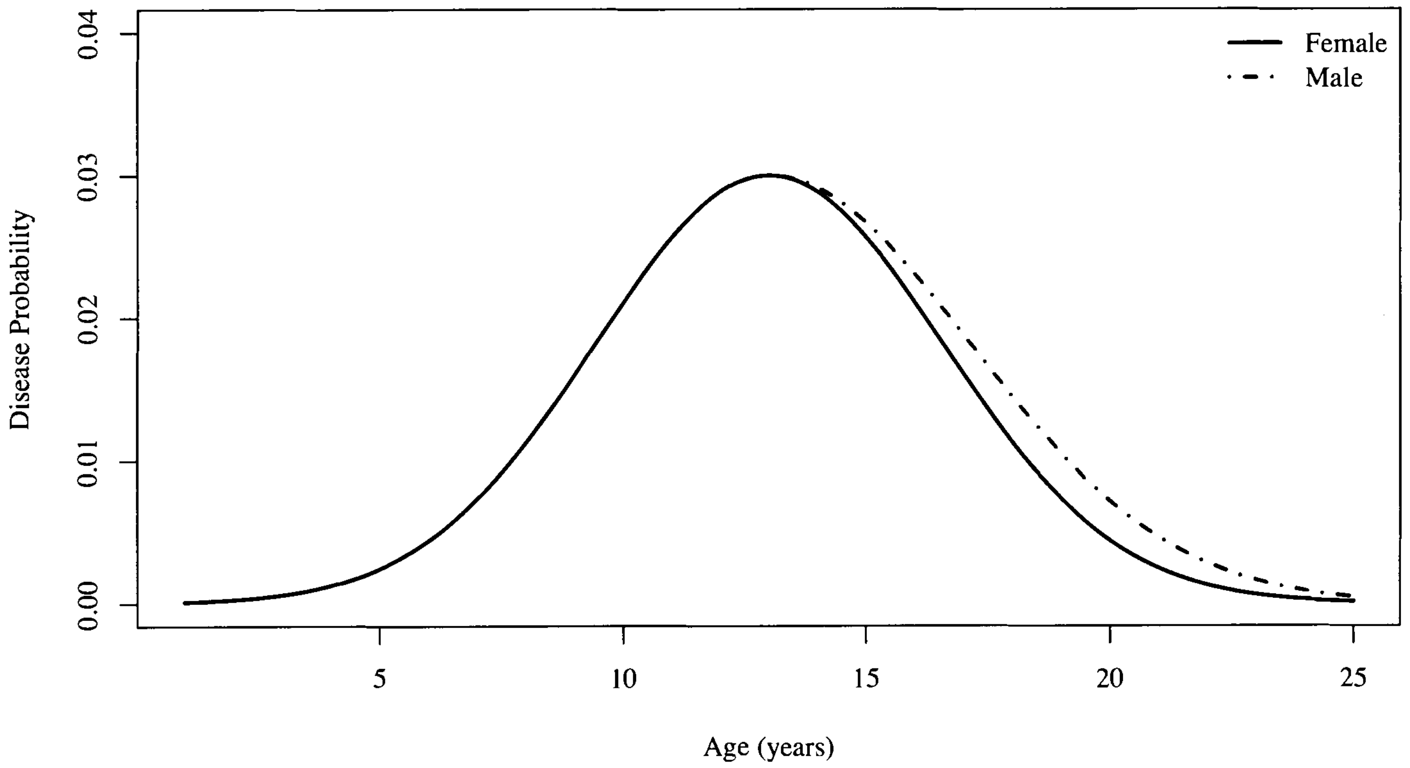


Figure 3.2: Theoretical risk curves for the (++) genotype, as a function of age, by gender. The age range corresponds to that observed for cases in our simulated dataset.

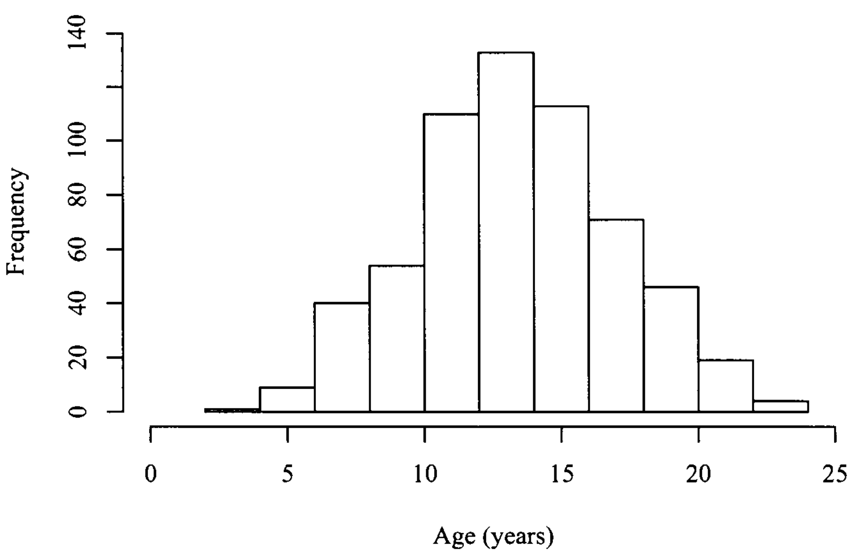


Figure 3.3: Age-of-onset distribution for the cases in the simulated dataset.

The theoretical disease risk curves as a function of age, for different gender and genotype combinations are presented in Figure 3.1. The curves reflect the peak incidence of type 1 diabetes around puberty. They also mimic the male preponderance among older-onset patients, observed in the Swedish population (Å. Lernmark, personal communication). For example, as shown in Figure 3.2 for the (++) genotype, the risk curve for males is higher than that for females at age > 15 years.

3.2 Descriptive Summaries

3.2.1 Marginal Distributions in Cases

The proportions of males and females among the sampled cases were approximately equal; of the 600 cases, 314 (52%) were male and 286 (48%) were female. Figure 3.3 illustrates the age-of-onset distribution for the **diabdat** dataset. The age-of-onset for the simulated data set had a range of 3.2 - 22.9 years of age, with a mean age of 13.4 years and standard deviation of 3.6 years. The dietary factor distribution for the cases is presented in Figure

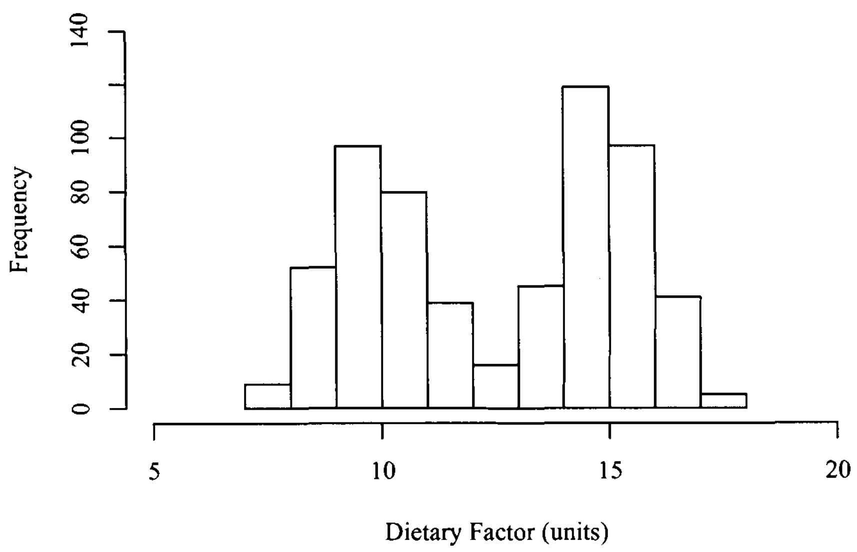


Figure 3.4: Distribution of the dietary factor for the cases in the simulated dataset.

3.4. As expected, the case distribution of the dietary factor is bimodal, with peaks centred on the subpopulation mean values of ten and 15. Overall, the range of the dietary factor was 7.1 - 17.6 units, with a mean of 12.6 units. Ignoring gender, age and the dietary factor, 188/600 $\approx$ 31% of the cases had zero copies of the risk allele, 283/600 $\approx$ 47% of the cases had one copy of the risk allele and 129/600 $\approx$ 22% of the cases had two copies of the risk allele.

3.2.2 Joint Distributions in Cases

We next looked at pairwise distributions of the risk factors in order to assess the possibility of their association in cases. Such an association would be consistent with statistical interaction between the risk factors *provided they are independent in the population* (Piegorsch et al. 1994).

Figure 3.5 presents the age-of-onset distribution for the cases by gender and shows that male cases have a slightly older mean age-of-onset than female cases. The mean age and

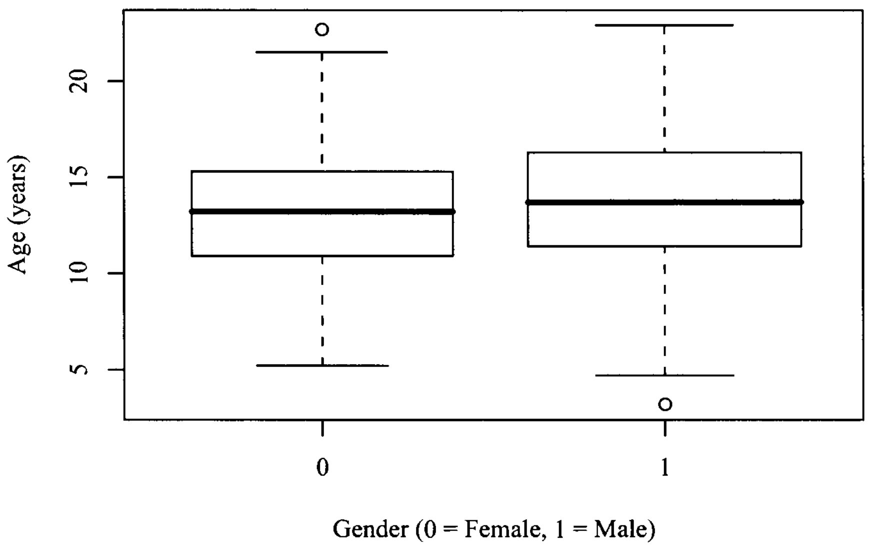


Figure 3.5: Case age-of-onset distribution by gender in the simulated dataset.

standard deviation for the female cases was 13.12 ± 3.45 years. In comparison, the mean age and standard deviation for the male cases was 13.72 ± 3.78. A t-test revealed that the mean age-of-onset for the male cases was statistically different from the mean age-of-onset for the female cases ($\alpha = 0.05$, $df = 598$, $t = -2.007$, $p\text{-}value = 0.045$). Such an association between age-at-onset and gender was simulated through an interaction term in the risk model (equation 3.1), and leads to a male preponderance among older-onset cases (e.g. ≥ 15 years of age). To further verify that this preponderance was reflected in our data, the male and female cases were grouped into two age categories: < 15 years of age and ≥ 15 years of age (Table 3.2). As expected, a goodness-of-fit test of independence revealed that the greater number of males in the ≥ 15 age category was statistically significant ($\alpha = 0.05$, $df = 1$, $\chi^2 = 4.57$, $p\text{-}value = 0.032$).

Figure 3.6 presents the distribution of the dietary factor for the cases as a function of gender in the simulated dataset. Based on Figure 3.6, the dietary factor does not appear to be associated with the gender of the case. No association between the dietary factor and gender was simulated in the data. Table 3.3 shows the gender distribution within each

	Age-of-Onset		
Gender	$< 15(\%)$	$\geq 15(\%)$	Total
Male	197(49)	117(58.5)	314
Female	203(51)	83(41.5)	286
Total	400(100)	200(100)	600

Table 3.2: Male preponderance in older onset cases in the simulated dataset.

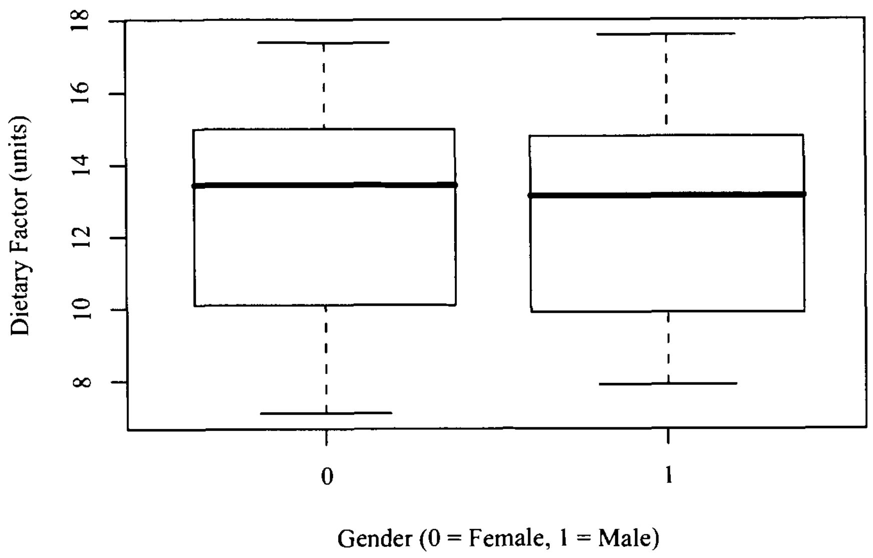


Figure 3.6: Case dietary factor distribution by gender in the simulated dataset.

	Genotype			
Gender	+ − (%)	+ + (%)	− − (%)	Total
Male	144 (51)	98(52)	72(56)	314
Female	139 (49)	90(48)	57(44)	286
Total	283 (100)	188 (100)	129(100)	600

Table 3.3: Gender distribution for each of the three genotypes.

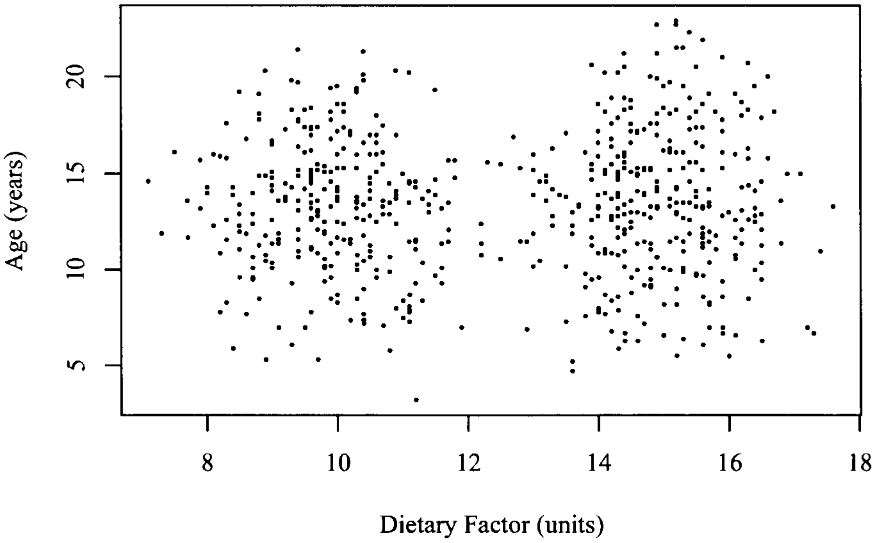


Figure 3.7: Scatterplot of case age-of-onset *versus* the dietary factor in the simulated dataset.

of the three genotypes. No association between gender and genotype was modeled in the simulated dataset. As expected, a chi-square goodness-of-fit test confirmed this lack of association in Table 3.3 ($\alpha = 0.05$, *df* $= 2$, $\chi^2 = 0.868$, *p-value* $= 0.648$).

Figure 3.7 is a scatterplot of the case age-of-onset *versus* the case dietary factor. The scatterplot reveals two slightly overlapping ellipsoids of roughly equal size and dispersion centred at the subpopulation mean values of ten and 15 of the dietary factor. Figure 3.7 indicates correctly that there is no association between case age-of-onset and the dietary factor in the simulated dataset.

Figure 3.8 shows the distribution of case ages as a function of case genotype. The

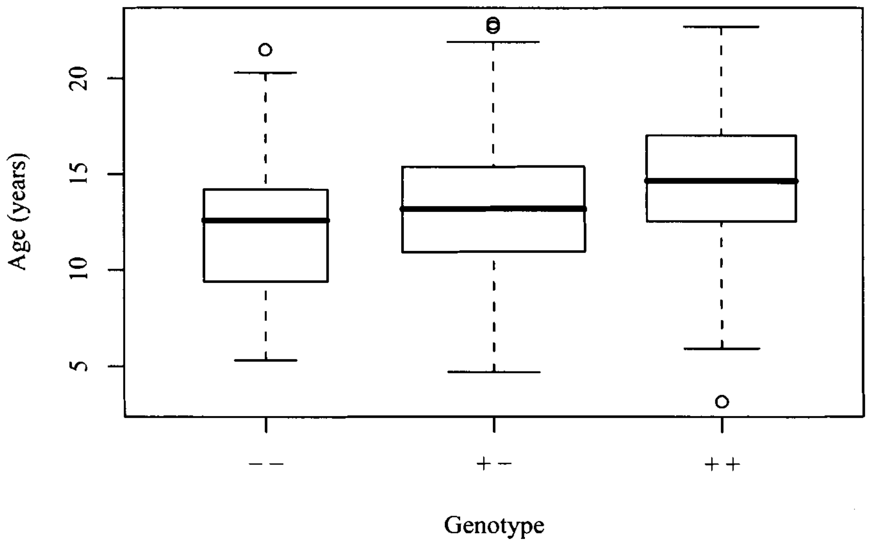


Figure 3.8: Case age-of-onset distribution by genotype in the simulated dataset.

median age-of-onset for cases with genotypes (++), (+−) and (−−) was 14.7 years, 13.2 years, and 12.6 years, respectively. A one-way analysis of variance indicated that the mean age-of-onset for the cases within each genotype were significantly different ($F_{2,597} = 21.869$, *p-value* < 0.001). A plot of the residuals *versus* predicted values (not shown) indicated that the normality and equal variance assumptions were reasonable, despite the slight skewing visible in Figure 3.8. Based on the ANOVA results, the age-of-onset was associated with genotype in cases, with the age-of-onset decreasing for increasing number of copies of the risk allele. Such an association is consistent with the statistical interaction between these risk factors modeled in equation 3.1.

As illustrated in the boxplots presented in Figure 3.9, diet appears to vary substantially by genotype. The median value of the dietary factor for cases with genotypes (++), (+−) and (−−), was 11.3 units, 13.1 units and 14.3 units, respectively. However, it is important to emphasize that no interaction between genotype and the dietary factor was modeled in our simulated dataset (see equation 3.1). Indeed, the dietary factor was not even a risk factor for disease. Therefore, this is a false association due to the fact that the dietary

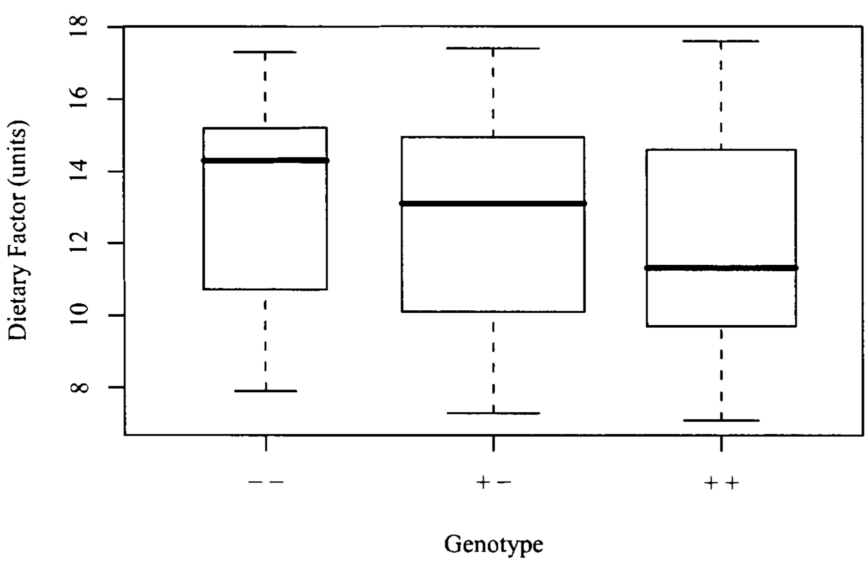


Figure 3.9: Case dietary factor distribution by genotype in the simulated dataset.

factor and minor allele frequency both varied as a function of subpopulation membership. Thus, if subpopulation membership is not taken into account, a researcher could be misled by a case-only analysis that suggests interaction between diet and genotype.

3.2.3 Parent-of-Origin Effects

Parent-of-origin effects were not simulated in this dataset. As expected, when the proportion of times the risk allele was transmitted to the affected child by heterozygous mothers was compared to the same proportion for heterozygous dads, no parent-of-origin effects were apparent. Specifically, the risk allele was transmitted $159/324 = 49.1\%$ of the time in heterozygous mothers compared to $148/294 = 50.3\%$ of the time in heterozygous fathers.

3.3 Transmission/Disequilibrium Test (TDT)

As discussed in Chapter 1, the TDT statistic may be constructed from the allelic transmissions of heterozygous parents. Under the null hypothesis of no linkage or association,

parents	$+ + * + -$		$- - * + -$		$+ - * + -$		
n (%)	268 (44.6)		154 (25.7)		178 (29.7)		
child	$++$	$+-$	$--$	$+-$	$++$	$+-$	$--$
observed	139	129	80	74	49	80	49
expected	134	134	77	77	44.5	89	44.5
$\chi^2(df)$	0.373 (1)		0.234 (1)		1.82 (2)		
p-value	0.541		0.629		0.402		

Table 3.4: Conditional genotype frequencies in cases and expectations under no association.

	Non-Transmitted Allele		
Transmitted Allele	$+$	$-$	Total
$+$	268	391	659
$-$	387	154	541
Total	655	545	1200

Table 3.5: Contingency table for the transmission/disequilibrium test in the **diabdat** dataset.

the transmissions of heterozygote parents to the cases should have the same distribution as transmissions to a random sample of children. Table 3.4 presents the observed conditional genotype frequency in cases *versus* the expected conditional genotype frequency in cases under the null hypothesis of no linkage or no association. According to Table 3.4 heterozygous parents transmit the minor allele, $-$ to the cases $387/778 = 49.7\%$ of the time. Table 3.5 can be constructed from the data presented in Table 3.4. The test statistic in equation 1.2 can be calculated using the cell counts presented in Table 3.5:

$$\frac{(\mid 391 - 387 \mid -1)^2}{(391 + 387)} = \frac{9}{778} = 0.0116$$

Thus, the TDT provides no evidence for linkage and association between the SNP and disease (*p-value* = 0.91).

3.4 Exploring Interaction with Transmission-Based Plots

Investigations of non-infectious disease aetiology are often based on the assumption that the disease is caused by alleles interacting with environmental factors. Umbach and Weinberg (2000) consider the situation where a diallelic locus and a binary exposure both influence disease susceptibility. One proposed method of investigating such gene-by-environment interaction is an extension of the TDT which tests for differences in the transmission rate of

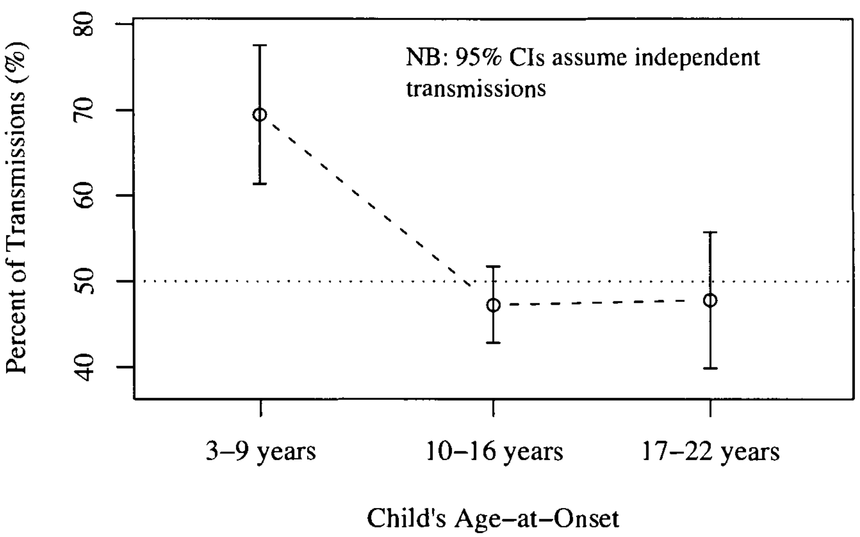


Figure 3.10: Percent times heterozygous parents transmitted the minor allele to the child, by child's age-at-onset.

the risk allele from heterozygous parents to exposed *versus* unexposed cases. This method uses an equal-transmission-rate null hypothesis as a proxy for the no-interaction hypothesis. However, such an approach is not generally valid (Umbach and Weinberg 2000). First, if there is an association between allele frequency and exposure levels in the general population, gene-by-environment interaction may be detected even if such interaction does not exist. Second, a test based on transmission rates ignores the fact that parents have been ascertained through their affected child and assumes statistical independence of alleles transmitted by parents who are both heterozygous. However, for some disease models, this assumption may not hold under the alternative hypothesis of linkage and association.

Figures 3.10 and 3.11 illustrate the basic idea behind the approach. Figure 3.10 was generated by dividing up the age range of the cases into three approximately equal-sized groups. Then, for each of the three age groups the proportion of times that a heterozygous parent transmitted the minor allele to their child was calculated and plotted as a percentage. The error bars shown in the plot were constructed using binomial probabilities, and assume

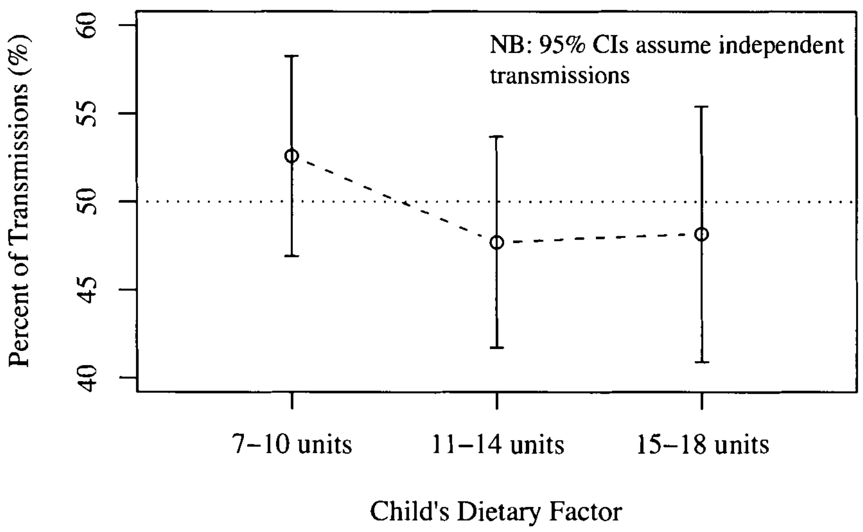


Figure 3.11: Percent times heterozygous parents transmitted the minor allele to the child, by child's level of the dietary factor.

statistical independence of the alleles that are transmitted by parents in case-parent trios with both parents heterozygous. As noted previously, transmissions of these alleles at the candidate locus are not, in general, statistically independent. Thus, the error bars shown may be too narrow. Figure 3.10 suggests the risk allele is transmitted more often than would be expected under Mendelian segregation when the case has a young age-at-onset but not when the case has an older age-at-onset. Hence, Figure 3.10 is very suggestive of an interaction between the SNP genotype and the age. This result is not unexpected given that genotype by age interaction was modeled in our simulated dataset. Therefore, the interaction suggested by Figure 3.10 is valid.

Figure 3.11 was obtained in the same manner as Figure 3.10 and shows transmission rates of heterozygous parents by the levels of the child's dietary factor. There is a weak suggestion that the risk allele is transmitted more often than would be expected when the case has low levels of the dietary factor but not when the case has higher levels. Genotype-by-diet interaction was not modeled, but allele frequency and dietary factor exposure levels are associated in our simulated dataset. Therefore, Figure 3.11 may be misleading because a transmission-based approach was used to explore gene-by-environment interaction.

With this in mind, we would like to create a visual representation of genotype-by-environment interaction that is valid more generally, so long as the probability models are specified correctly. The smoother motivated in Chapter 2 should provide a valid visual representation of genotype-by-environment interaction because it is likelihood-based.

3.5 Exploring Interaction with Our Smoother

We implemented our smoother in the **trioplot()** function (see Appendix A) and applied it to the simulated dataset. The panel on the left of Figure 3.12 gives the smooth for f_1 and the panel on the right gives the smooth for f_2, with both smooths centred by their average in the dataset. In the absence of statistical interaction, $f_1 = f_2 \equiv 0$ and the resulting smooth is expected to be a roughly horizontal line through zero. The series of short vertical lines along the x-axis in the two panels is called the *rugplot* and it indicates the locations of the observed data. The data values are randomly jittered to break ties. The dotted lines

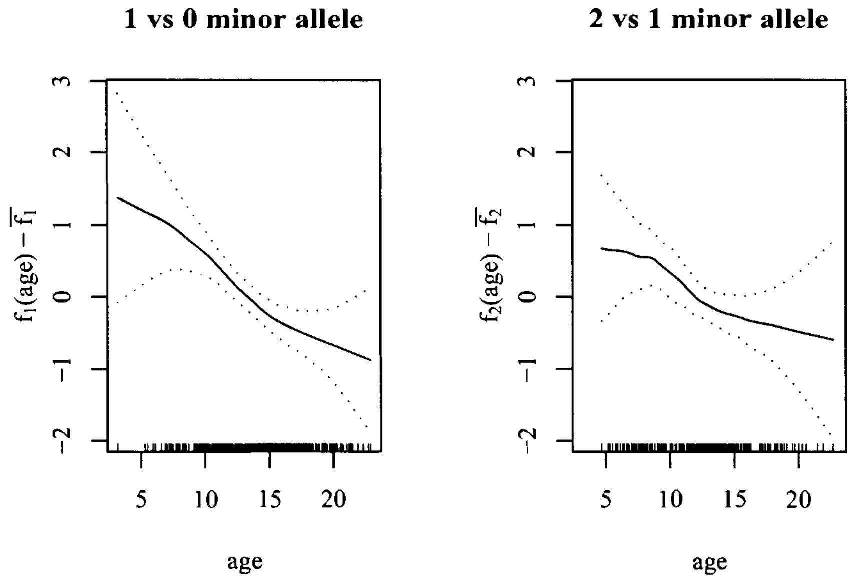


Figure 3.12: Smooths of age specific log relative risks from the simulated dataset.

above and below the estimated smooth are the pointwise 95% confidence limits which are obtained by adding and subtracting 2 × the standard error from each fitted value (Hastie 1993).

The smooth in Figure 3.12 suggests that statistical interaction between the SNP and age is present. Such interaction is expected. As will be demonstrated in the next section, incorporating an interaction term into the risk model is important for detecting an association between the SNP and the disease. Figure 3.12 also indicates that data at the youngest and oldest ages are sparse which, together with a smoothing window of only half the regular size, explains the wide confidence limits at these ages in both panels. In contrast to the SNP-by-age interaction evident in Figure 3.12, the smooth in Figure 3.13 suggests no statistical interaction between the SNP and the dietary factor as a horizontal line through zero can be drawn within bounds of the pointwise 95% confidence limits for the smooth. This result is not unexpected given that no statistical interaction between the SNP and the dietary factor was simulated.

The **trioplot()** function can also be used to explore SNP interaction with a nominal

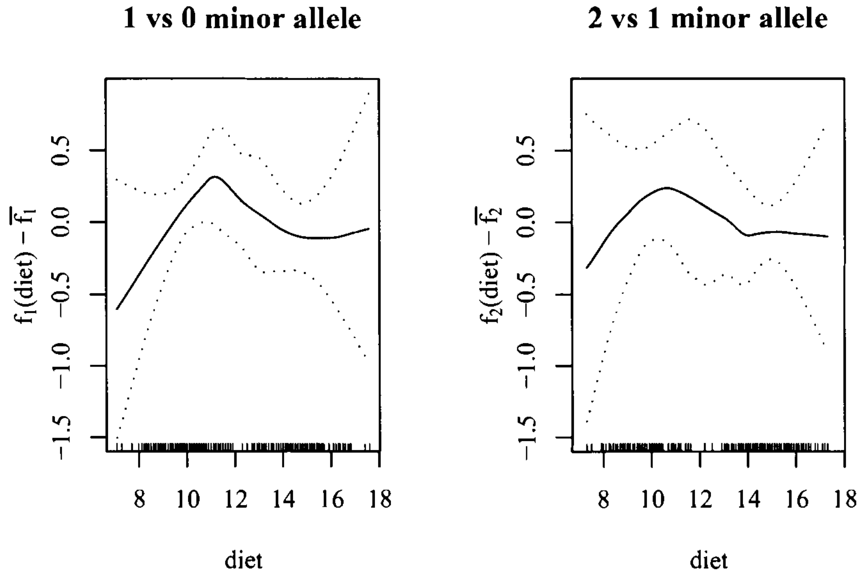


Figure 3.13: Smooths of dietary factor specific log relative risks from the simulated dataset.

covariate, such as gender. This situation is the simplest example of a smooth (Hastie and Tibshirani 1990). Figure 3.14 presents the plots generated by using the **trioplot()** function to investigate SNP-by-gender interaction. As in the earlier plots, the first panel presents the results for f_1, and the second panel presents the results for f_2. SNP-by-gender interaction seems unlikely given the overlapping confidence intervals.

3.6 Likelihood-Based Tests

The exploratory phase discussed in the previous section indicates that a term representing SNP-by-age interaction should be included in the model fit to the **diabdat** dataset. To fit our risk model, we can exploit the fact that the likelihood has the same form as the likelihood from a conditional logistic regression for a matched case-control study (Hastie and Tibshirani 1990). In our case, the match-sets are the four possible "pseudo-sibs" in * that could have resulted from the parents' transmissions. Our likelihood also has the same form as a Cox proportional-hazards likelihood for survival data, with the risk-sets being the

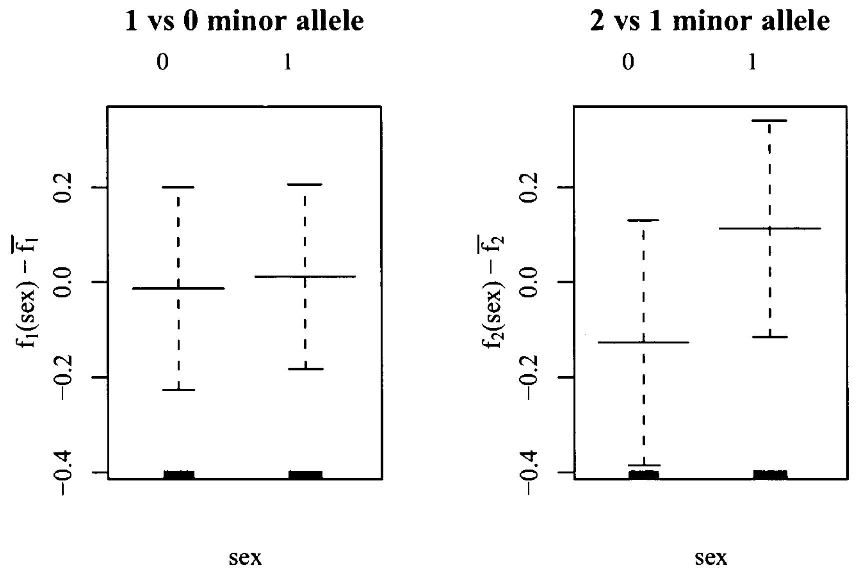


Figure 3.14: Smooths of gender specific log relative risks from the simulated dataset.

four pseudo-sibs (Collett 2003).

R's **clogit()** function for conditional logistic regression was used to fit various risk models to our simulated dataset. In order to use the **clogit()** function, the **diabdat** dataset had to be "expanded" into the required match-sets. The first three match-sets for the expanded **diabdat** dataset are shown in Table 3.6. The first column of Table 3.6 identifies the family, and the last column of the table gives the affection status of the child. The row with an affection status of one represents the original case. The remaining three rows within each family represent the pseudo-sibs. The age, gender and dietary factor for each of the pseudo-sibs remains the same as that for the original case. The column titled "ncop" indicates the number of copies of the minor allele that the original case and the three pseudo-sibs possess. The expanded dataset was first used to fit the following risk model:

$$\log\left(P(D=1 \mid G_c = g)\right) = \beta_0 + \beta_1 Z_1(g) + \beta_2 Z_2(g),$$

where β_0, β_1 and β_2 are regression parameters, $Z_1(g) = 1$ if $g = (+-)$ or $(--)$ and zero otherwise, and $Z_2(g) = 1$ if $g = (--)$ and zero otherwise. This risk model considered only

famid	age	sex	diet	ncop	aff
1	11.9	1	15.6	0	0
1	11.9	1	15.6	1	1
1	11.9	1	15.6	1	0
1	11.9	1	15.6	2	0
2	20.2	1	14.3	0	0
2	20.2	1	14.3	0	0
2	20.2	1	14.3	1	1
2	20.2	1	14.3	1	0
3	13.7	0	10.9	0	0
3	13.7	0	10.9	0	0
3	13.7	0	10.9	1	1
3	13.7	0	10.9	1	0

Table 3.6: The first three match-sets of the expanded **diabdat** dataset.

a genetic main effect and was specified by the following model formula in **clogit()**

$$aff \sim I(ncop > 0) + I(ncop = 2) + strata(famid).$$

In the above formula, the term $I(ncop > 0)$ is an indicator function that represents Z_1 and $I(ncop = 2)$ is an indicator function that represents Z_2. For example, consider the fourth row of Table 3.6. This pseudo-sib carries two copies of the minor allele; therefore, the term $I(ncop > 0)$ equals one and the term $I(ncop = 2)$ equals one. The term $strata(famid)$ tells the **clogit()** function that match-sets are indicated by the family identification code. In this way, the **clogit()** function allows us to condition on parental genotypes. Under this risk model, the score test for the genetic effect is not significant (*Score test* $= 1.93$, *df* $= 2$, *p-value* $= 0.38$).

Since the smooth for age (presented in Figure 3.12) suggests a linear statistical interaction between SNP and age we next considered the risk model:

$$\begin{aligned} &\log\left(P(D = 1 \mid G_c = g, A = a)\right) \\ &= \beta_0 + \beta_1 Z_1(g) + \beta_2 Z_2(g) + \beta_3 a + \beta_4 Z_1(g) a + \beta_5 Z_2(g) a, \end{aligned} \tag{3.2}$$

where β_0 - β_5 are regression parameters, $Z_1(g)$ and $Z_2(g)$ are defined as in equation 3.2 and a is the age of an individual. This risk model considers main effects for the genotypes and age, as well as their interaction, and was specified by the following model formula in

clogit():

$$aff \sim I(ncop > 0) + I(ncop = 2) + I\left((ncop > 0) * age\right) + \\ I\left((ncop = 2) * age\right) + strata(famid),$$

Since we are conditioning on the parental genotypes and the age of an affected child, the main effect for age cancels out in the numerator and denominator of the conditional likelihood (because all pseudo-sibs within the same family have the same age). Therefore, β_3 is not estimable under this model. As a result, we do not need to specify the main effect for age in the model formula for **clogit()**. Under this interaction model, the score test for the overall genetic effect is significant (*Score test* = 33.7, *df* = 4, *p-value* = 8.7×10^{-7}). Therefore, adjusting for the interaction uncovers a genetic effect that would otherwise have been missed.

Our smooths for SNP-by-diet interaction suggested no genotype-by-diet interaction (Figure 3.13). To check this result, we considered the risk model in equation 3.2, with a now being the level of an individual's dietary factor. The risk model was specified by the following model formula in **clogit()**:

$$aff \sim I(ncop > 0) + I(ncop = 2) + I((ncop > 0) * diet) + I((ncop = 2) * diet) + strata(famid).$$

The resulting score test confirmed that there is no SNP-by-diet interaction (*Score test* = 3.16, *df* = 4, *p-value* = 0.53). Likewise, the score test for SNP-by-gender interaction was also insignificant (*Score test* = 3.39, *df* = 4, *p-value* = 0.50).

In summary, is it worthwhile emphasizing that without accounting for SNP-by-age interaction, the genetic effect would have been missed. However, when the possibility of interaction is taken into account, the genetic effect is detected.

Chapter 4

Conclusions and Future Work

We proposed a data-smoothing method for exploring statistical interaction between a single nucleotide polymorphism (SNP) and a non-genetic risk factor, such as age, in case-parent trios. Our smoother can be used as a diagnostic tool for checking for the presence of interaction after conducting a genetic association test, such as the transmission/disequilibrium test, that does not account for interaction. Alternatively, if an interaction model is fit to the data, the smoother can be used to check the adequacy of the chosen model.

As illustrated in Chapter 3, our smoother uncovered important genotype-by-age interaction in the simulated dataset. The subsequent incorporation of a genotype-by-age interaction term into our risk model allowed us to detect an overall genetic effect that would otherwise have been overlooked. Further, our investigation of genotype-by-diet interaction illustrated that our smoothing method was not misled by the population substructure present in our simulated dataset. Transmission based approaches can be misled by such substructure (Umbach and Weinberg 2000) and in fact suggested possible genotype-by-diet interaction. Therefore, our smoothing method provides a more robust tool for exploring statistical interaction than a transmission-based approach.

We used simulated data to illustrate the properties of our smoothing method. Therefore, one obvious direction for future work would be to apply the smoother to a real dataset. Some other possible extensions of this work include the development of more efficient smoothers that enable us to use all of the data available. As discussed in Chapter 2, conditioning on $\{G_c = ++ \text{ or } +-\} \equiv \{Z_2 = 0\}$ to extract information about (β_1, f_1) and then conditioning on $\{ G_c = +- \text{ or } -- \} \equiv \{ Z_1 = 1 \}$ to extract information about (β_2, f_2) resulted in a loss of information. For example, by conditioning on the event $Z_2 = 0$, case-parent trios with $G_c = (--)$ were excluded from the estimation of f_1. Similarly, by conditioning on the event $Z_1 = 1$, case-parent trios with $G_c = (++)$ were excluded from the estimation of f_2. As a result, not all of the available data is incorporated into the smooths for f_1 and f_2. Since data collection is an expensive and time-consuming procedure, we would like to extend our current method so that all of the available data is used.

Family-based association studies that require the affected child's genotype and the genotypes of the child's parents are limited to diseases with a relatively early age-of-onset because parental genotype information for late age-of-onset diseases (e.g. Alzheimer's disease) may not be obtainable; the parents of the affected child may no longer be alive to sample. As a result, our approach would not be applicable to diseases with a late age-of-onset.

In this project, we assumed that all parental genotypes were available. However, missing parental information is a common problem in family-based association studies using real data, and according to Curtis and Sham (1995), incorrectly applying the standard TDT to trios in which missing parental genotypes can be inferred from the genotype of the affected child can produce false-positive results, particularly when dealing with a diallelic locus. In their paper, Curtis and Sham (1995) demonstrate how bias in the standard TDT can result when genotype information on one parent is missing and the other parent is heterozygous (i.e. informative). They consider a diallelic locus with one common allele and one extremely rare allele and show that, even if transmissions from the heterozygous parent can be inferred given the genotype of the affected child, including such trios leads to an apparent (but false) preferential transmission of the more common allele. Thus, such incomplete trios should not be included in a standard TDT analysis. With this issue in mind, it would be interesting to investigate whether our method can be extended to correctly incorporate trios with one heterozygous parent and another parent with missing genotype information.

Parental genotypes may be missing not only because of a later-onset disease but also because of genotyping error. For example, some parents who are sampled may not be able to be reliably genotyped and so their genotypes will not be called. In SNP genotyping, heterozygotes can be more difficult to call than homozygotes (Mitchell et al. 2003), leading to *differential dropout* (Hao and Cawley 2007), in which the genotypes that fail to be called are more likely to be heterozygous. Allen et al. (2003) define *informative missingness* as any situation where the parent's genotype at the locus under investigation is associated with the reason why the parent is missing. Differential dropout can therefore lead to informative missingness. Failing to account for informative missingness may result in biased inference (Little and Rubin 2002). In their paper, Hao and Cawley (2007) show that the presence of even moderate differential dropout leads to severe bias in the standard TDT.

Undetected genotyping error may also exist in the complete case-parent trio data that we do have. For example, according to Cutler et al. (2001), heterozygotes also tend to have higher miscall rates. This phenomenon is known as *allelic dropout.* Gordon et al. (2001) showed that the TDT can have inflated type 1 error rates under simple models of genotyping error. Further, Mitchell et al. (2003) demonstrated that undetected genotype errors can cause apparent over transmission of common alleles in the TDT and concluded that such errors may contribute to an inflated false-positive rate among reported TDT associations.

In our approach, we have assumed that the genotypes are uninformatively missing and are observed without error. Allowing for both differential and allelic dropout would thus be a very important direction for future extensions of our approach. A possible starting point might be the work of Chen (2004), which is conceptually similar to our approach, but allows for informatively missing parents. It would be interesting to see if our results could be extended to accommodate not only informatively missing parents, but also genotyping error.

An additional direction for future research would be to try settings for the smoothing parameter of the loess smoother (Cleveland et al. 1993) other than the value of 2/3 that we used in our approach. It would be interesting to explore how the appearance of the smooth changes as the smoothing parameter is altered. We are interested in investigating the sensitivity of the suggested interaction (or lack of interaction) to adjustments of the smoothing parameter. *Cross-validation (CV)* can be used to help select the appropriate span size or appropriate target equivalent degrees of freedom. Suppose we have a set of n points, represented as (x_i, y_i), where i represents the ith point. In leave-out-one cross-validation, one of the n points is temporarily removed and the smooth is estimated using the remaining $n - 1$ points. This process is repeated, one point at a time, for all n points. The *cross-validation sum of squares* is then calculated. Using the notation developed by Hastie and Tibshirani (1990) the cross-validation sum of squares is:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_{\lambda}^{-i}(x_i)\right)^2, \tag{4.1}$$

where λ represents the smoothing parameter or span, the superscript $-i$ indicates that the ith point has been removed from the computation, and $\hat{f}_{\lambda}^{-i}(x_i)$ indicates the fit at x_i for smoothing parameter is λ. When the cross-validation sum of squares is computed for a number of different λ then the value of λ, referred to as $\hat{\lambda}$, that minimizes the cross-validation sum of squares is selected as the most appropriate value for the smoothing parameter or span (Hastie and Tibshirani 1990). However, this approach has a few limitations, including the fact that it is computationally intensive for large datasets, and that it requires preselection of an appropriate subset of λ values, which may be hard to determine from the outset of the analysis. In our context, a set of possible span values (e.g. $1/4, 1/3, 2/3$, and $3/4$) could be evaluated using cross-validation, thereby providing a more rigorous span selection procedure for use in our smoothing method. For a more in-depth discussion of the benefits and limitations of using cross-validation for selection of the smoothing parameter, see Hastie and Tibshirani (1990).

We chose a loess smoother because we felt it was more intuitive with respect to how the smoothing parameter is set. In loess, the smoothing parameter, or span, is set by specifying the proportion of the data to use in the calculation of each plotted point in the estimated curve. Therefore, the larger the span (i.e. the closer the span is to one), the larger the window and smoother the estimated curve. Further, loess smoothers have several attractive features, including: (i) they can adapt their window size to reflect the density of the points in a given neighbourhood, and (ii) since they use a smooth weighting function called a tri-cube function, they are not as "wiggly" as smoothers based on running means (Hastie and Tibshirani 1990). Other smoothers are available in R's **gam** package; for example, the function s() fits a smoothing spline to the data (Hastie 1993). The smoothing spline option may be a viable alternative, but we have not considered it here. However, the choice of smoother may be largely subjective. As long as the smoothing parameter is appropriately specified there do not seem to be large differences between the various types of smoothers (Silverman 1984; Müller 1987). Therefore, it is probably more important to investigate various smoothing parameters rather than different smoothers.

Finally, it would be exciting to explore the two major underlying assumptions of this method: (i) conditional independence of the case's genotype and the non-genetic covariate, given the parental genotypes, and (ii) Mendelian transmission probabilities and then adapt our method to handle those situations where these two assumptions are not met. In case-control or case-only association studies, incorrectly assuming independence of the genetic and non-genetic risk factors in the population leads to false-positive statistical interactions (e.g. Shin et al. 2005). By analogy, one would expect that incorrectly assuming independence within a family in a family-based association study would also lead to a false impression of interaction. It is more difficult to speculate on the impact that non-Mendelian transmission will have on our smoother. However, adjusting for known non-Mendelian transmission appears to be relatively straightforward. The derivation of the likelihood contributions suggests that the disease risks $P(D = 1 \mid T_c = t, G_p = g_p, X = x)$ in equation 2.5 should be reweighted by the true non-Mendelian offspring probabilities $P(T_c = t \mid G_p = g_p)$, given the parental mating type. When these non-Mendelian offspring probabilities are known, they would lead to different offset terms in the smoother that are specific to the parental mating type.

Appendix A

The trioplot() Function

The R function **trioplot()** produced the smooths presented in Figures 3.12, 3.13 and 3.14. The **trioplot** function takes the following arguments:

i. a dataframe with columns for parental genotypes, child genotypes and the non-genetic covariates of the child,

ii. the name of the two columns in the dataframe that hold parental genotypes, as genotype objects,

iii. the names of the column in the dataframe that holds the child genotypes, as genotype objects,

iv. the name of the non-genetic covariate being examined for interaction with genotype, and

v. the character string or number that represents the minor allele.

In order to apply the trioplot function several R packages must be loaded. The first necessary package is the **genetics** package. This package converts the parental genotypes and child genotypes into genotype objects. The smooths themselves are generated via a generalized additive model, and as a result, the **gam** package must be loaded as well. At the time this project was written, R could be downloaded from http://cran.r-project.org/.

After basic error checking to ensure that only trios with complete genotype information are included in the generation of the smooths, the function determines which trios contain informative parental genotypes. This step is accomplished by counting the number of copies of the minor allele that are present in each parent. Only those trios where at least one parent is heterozygous are informative (i.e. can have offspring that are genetically different). Next, the complete and informative trios are checked to ensure that no Mendelian inconsistencies are present. Further, since we are interested in exploring statistical interaction between the SNP and a non-genetic covariate, the selected trios are also checked to ensure that complete data on the non-genetic covariate is present. Only those trios that are complete, informative, contain no Mendelian inconsistencies, and contain valid values for the non-genetic covariate are retained to produce the smooths.

As discussed in Chapter 2, the smooths are generated by forming two groups. The first group is created by pooling the cases that contain zero or one copy of the minor allele, and the second group is formed by pooling the cases that contain one or two copies of the minor allele. After each grouping, we can apply the **gam()** function. In order to use the **gam()** function we must specify the *formula*, which as with other regression models is of the form *response* $\sim$ *predictors*, and the *family*. Specifying the *family* ensures that the appropriate error distribution and link function is used in the generalized additive model. Since we are working with a binary response for each group, we specified the family as binomial with the canonical logit link (McCullagh and Nelder 1989).

The response variable for the first generalized additive model is an indicator variable for cases that have one copy *versus* zero copies of the minor allele. The second gam was fitted using an indicator of two *versus* one copy of the minor allele as the response variable. Both of these generalized additive models were generated using the function **lo()** as the smoother. In order to use the **lo()** function we specified both the *span* and the degree of smoothing. In **lo()**, the argument *span* is the smoothing parameter, and it essentially indicates the size of the neighbourhood to use for taking the moving average. We specified our *span* to be 2/3 of the data; however, the *span* in the function **lo()** can be set to any value between zero and one. Setting the *span* to zero implies no smoothing, whereas setting the span to one means that all of the data passed to the function is used to generate the smooth. The argument *degree* specifies the degree of the local polynomial to be fit. In **lo()** the degree is currently restricted to one or two; for our purposes we set the degree to the default (one).

Bibliography

Allen, A., P. Rathouz, and G. Satten (2003). Informative missingness in genetic association studies: case-parent designs. *American Journal of Human Genetics. 72*, 671–680.

Cardon, L. and J. Bell (2001). Association study designs for complex diseases. *Nature Reviews: Genetics. 2*, 91–99.

Chen, Y.-H. (2004). New approach to association testing in case-parent designs under informative parental missingness. *Genetic Epidemiology. 27*, 131–140.

Clayton, D. (2003). Chapter 32: Population association. In D.J. Balding, M. Bishop and C. Cannings (Ed.), *Handbook of Statistical Genetics: Volume 2, 2nd Edition*, pp. 939–960. Chichester, UK: John Wiley and Sons, Ltd.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association. 74*, 829–836.

Cleveland, W., E. Grosse, and W. Shyu (1993). Chapter 8: Local regression models. In J.M. Chambers and T.J. Hastie (Ed.), *Statistical Models in S*, pp. 309–376. London, UK: Chapman and Hall, Inc.

Collett, D. (2003). *Modelling Survival Data in Medical Research, 2nd Edition.* Boca Raton, FL: Chapman and Hall/CRC Press LLC.

Curtis, D. and P. Sham (1995). Letters to the editor: a note on the application of the transmission disequilibrium test when a parent is missing. *American Journal of Human Genetics. 56*, 811–812.

Cutler, D. J., M. Zwick, M. Carrasquillo, C. Yohn, K. Tobin, C. Kashuk, D. Mathews, N. Shah, E. Eichler, J. Warrington, and A. Chakravarti (2001). High-throughput variation detection and genotyping using microarrays. *Genome Research. 11*, 1913–1925.

Ewens, W. and R. Spielman (2003). Chapter 33: The transmission/disequilibrium test. In D.J. Balding, M. Bishop and C. Cannings (Ed.), *Handbook of Statistical Genetics: Volume 2, 2nd Edition*, pp. 961–972. Chichester, UK: John Wiley and Sons, Ltd.

Freeman, S. and J. Herron (2007). *Evolutionary Analysis, 4th Edition.* Upper Saddle River, NJ: Pearson Prentice Hall.

Gordon, D., S. Heath, X. Liu, and J. Ott (2001). A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *American Journal of Human Genetics. 69*, 371–380.

Hao, K. and S. Cawley (2007). Differential dropout among SNP genotypes and impacts on association tests. *Human Heredity. 63*, 219–228.

Hastie, T. (1993). Chapter 7: Generalized additive models. In J.M. Chambers and T.J. Hastie (Ed.), *Statistical Models in S*, pp. 249–307. London, UK: Chapman and Hall, Inc.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models.* Boca Raton, FL: Chapman and Hall/CRC Press LLC.

Lazzeroni, L. and K. Lange (1998). A conditional inference framework for extending the transmission/disequilibrium test. *Human Heredity. 48*, 67–81.

Li, H. and J. Fan (2000). A general test of association for complex diseases with variable age of onset. *Genetic Epidemiology. 19 (Suppl 1)*, S43–S49.

Little, R. and D. Rubin (2002). *Statistical Analysis with Missing Data, 2nd Edition.* Chichester, NY: John Wiley and Sons, Ltd.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models, 2nd Edition.* London, UK: Chapman and Hall.

Mitchell, A., D. Cutler, and A. Chakravarti (2003). Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *American Journal of Human Genetics. 72*, 598–610.

Müller, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association. 82*, 231–238.

Piegorsch, W., C. Weinberg, and J. Taylor (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine. 13*, 153–162.

Self, S., G. Longton, K. Kopecky, and K.-Y. Liang (1991). On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics. 47*, 53–61.

Shin, J., B. McNeney, and J. Graham (2005, December). Likelihood inference in case-control studies of a rare disease under independence of genetic and continuous non-genetic covariates. *COBRA Preprint Series. Article 8.* http://biostats.bepress.com/cobra/ps/art8.

Silverman, B. (1984). Spline smoothing: the equivalent variable kernel method. *The Annals of Statistics. 12*, 898–916.

Spielman, R., R. McGinnis, and W. Ewens (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics. 52*, 506–516.

Thomas, D. (2004). *Statistical Methods in Genetic Epidemiology.* New York, NY: Oxford University Press, Inc.

Umbach, D. and C. Weinberg (2000). The use of case-parent triads to study joint effects of genotype and exposure. *American Journal of Human Genetics. 66*, 251–261.

Vittinghoff, E., D. Glidden, S. Shiboski, and C. McCulloch (2005). *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models.* New York, NY: Springer.