TOPICS IN MEASUREMENT ERROR AND MISSING DATA PROBLEMS

A Dissertation

by

LIAN LIU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2007

Major Subject: Statistics

TOPICS IN MEASUREMENT ERROR AND MISSING DATA PROBLEMS

A Dissertation

by

LIAN LIU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Raymond J. Carroll
Committee Members,	Ruzong Fan
	Jeffrey D. Hart
	Faming Liang
	Philip E. Mirkes
Head of Department,	Simon J. Sheather

August 2007

Major Subject: Statistics

ABSTRACT

Topics in Measurement Error and Missing Data Problems. (August 2007)

Lian Liu, B.S., Peking University, P. R. China;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. Raymond J. Carroll

This dissertation research consists of two problems, which cut across the fields of measurement error methods, semiparametric methods, missing data problems and statistical genetics. The following two paragraphs give brief introductions to each of the two problems, respectively.

We study the partially linear model in logistic and other types of canonical exponential family regression when the explanatory variable is measured with independent normal error. We develop a backfitting estimation procedure to this model based upon the parametric idea of sufficiency scores so that no assumptions are made about the latent variable measured with error. By a numerical example and a simulation study, we show that the proposed method gives better results than the naive method.

In genetics study, the genotypes or phenotypes can be missing due to various reasons. In this research, the impact of missing genotypes is investigated for high resolution combined linkage and association mapping of quantitative trait loci (QTL). We assume that the genotype data are missing completely at random (MCAR). Two regression models are proposed to model the association between the markers and the trait locus, and account for the missing genotypes. By simulation study we show that the proposed method can help to get correct type I error rates for a moderate

size data, although it does not improve power.

In this dissertation, the sufficiency score method has improved the functional approach of measurement error study. For a canonical exponential family, this semi-parametric method have provided better estimation and asymptotic properties. In the genetics study, a new method is proposed to account for the missing genotype in a combined linkage and association study. We have concluded that this method does not improve power but it will provide better type I error rates for a moderate size data.

To Yingxue and Shu.

ACKNOWLEDGEMENTS

It is a big fortune to be a Dr. Raymond J. Carroll's student in the most important stage of my life. I would like to express my sincerest appreciation for his direction, suggestion, encouragement, patience and continuous support for my professional development. Having Dr. Carroll as my advisor is something that I will always be proud of and cherish for the rest of my life.

A special thanks is owed to Dr. Ruzong Fan, who served on my committee, although the help he offered was so much more than from a committee member, I consider him as my second advisor. I would like to express my deep gratitude for all he has done to help me during the past few years.

My thanks are extended to Dr. Jeff Hart, Dr. Faming Liang and Dr. Philip Mirkes for their willingness to serve on my committee and for providing me with their valuable comments.

I also want to thank Dr. Jianhua Huang for his helpful comments and suggestions, Dr. Daren Cline for being my master's advisor, Dr. Michael Longnecker for his consistent support over the years, and Dr. Fred Dahm for his effort recruiting me to this department five years ago. Thanks to all my friends who have given me help.

My greatest appreciation is due to my lovely wife Yingxue. I appreciate the love, joy, hope, support, encouragement, trust, comfort and understanding that she has brought to my life. Being with her is the most wonderful thing that has ever happened in my life. Thank you, Yingxue!

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

Measurement error and missing data problems have been widely studied in literature. These topics have many biological applications, especially in nutrition, genetics and epidemiology areas. The interest in this dissertation research mainly focuses on two problems. In Chapter II, a semiparametric measurement error problem is studied, and we develop a sufficiency score method to solve this problem. In Chapter III, a high resolution combined linkage and association mapping analysis is carried out in a statistical genetics study, and the impact of missing genotype is investigated for this method. Chapter IV gives a summary of the dissertation and some future research topics. The rest of this chapter gives the background behind Chapter II and III, respectively. Section 1.1 introduces some basics about measurement error model. Section 1.2 turns into the missing data problems in statistical genetics study. Section 1.3 describes the structure of this dissertation.

1.1 Measurement Error

Measurement error study is usually about the regression problems in which some predictors are measured with error. A simple measurement error problem consists of a response variable Y, predictors X and Z and a model relating Y and (X, Z), e.g.,

$$Y = \beta_0 + X\beta_1 + Z\beta_2 + \varepsilon. \tag{1.1}$$

The format and style follow that of *Biometrics*.

However, X is not observable. This is the essential point of a measurement error model. From one perspective, measurement error models can be viewed as a special case of missing data problems. The predictor X is entirely missing. The difference from the missing data problem is that we observe a variable W related to X. And we usually assume W satisfies the following relationships.

$$\begin{aligned} W &= X + U; \\ U &= \text{Normal}(0, \Sigma_{uu}); \\ U &= \text{independent of } (Y, X, Z). \end{aligned} \tag{1.2}$$

Note that the observed data are (Y, W, Z), instead of (Y, X, Z).

The effects of measurement error are well-known. The measurement error in predictors causes biases in estimated regression coefficients. Let us consider the simple measurement error example defined in (1.2) and (1.2). Assume that all variables are scalar. If we ignore the measurement error and estimate the regression coefficients simply using the observed W, then we would not estimate β_1, instead we would estimate

$$\frac{\text{var}(X|Z)}{\text{var}(X|Z) + \Sigma_{uu}}\beta_1.$$

So, the goal of the measurement error study is to correct for such effects.

There are two basic approaches in the literature. See Carroll et al. (2006) for a review. **Structural** methods are likelihood-based approaches which require a distribution of the missing predictor X. These methods also include Bayesian modeling. **Functional** methods, on the contrary, make no assumptions about the distribution of the missing predictor. Structural methods require parametric models for the distribution of X, sometimes conditional on the observed covariates Z. When structural models are used, there are always concerns about the possible non-robustness of estimation and inference due to model misspecification of the unobserved X. Fuller

(1987, page 263) and Carroll et al. (1984) discuss this issue in the classic nonlinear regression and probit regression problems, respectively. There is no general agreement in the statistical literature about whether structural or functional methods are more appropriate. This dissertation research will focus on functional methods.

There is an enormous number of publications on this topic in linear regression, as summarized by Fuller (1987). In many cases, instead of a linear model, a flexible model might be allowed for one of the covariates of interest, e.g., age or body mass index (BMI). The partially linear model was built for this as given below,

$$Y = X\beta + \theta(Z) + \varepsilon.$$

Here, the function $\theta(Z)$ is unknown and the purpose of the study is to estimate the unknown parameter and function without assumptions about X. One may further extend this model to a generalized partially linear model. We present a sufficiency score method in Chapter II to study the measurement error problem under this framework.

1.2 Combined Linkage and Association Mapping

For many complex traits, such as diabetes, depression, alcoholism and hypertension, quantitative phenotypes can be very informative. Hence, it is of importance to develop statistical methods for mapping of quantitative trait loci/locus (QTL). There has been a long history in the research of linkage mapping of QTL (Almasy and Blangero, 1998; Feingold, 2002; Fulker, Cherny, and Cardon, 1995; Goldgar, 1990; Haseman and Elston, 1972; Pratt, Daly, and Kruglyak, 2000). Moreover, variance component models have been proposed for combined linkage and association mapping of QTL (Abecasis, Cardon, and Cookson, 2000a; Abecasis, Cookson, and Cardon, 2000b; Allison, 2001; Almasy et al., 1999; Boerwinkle, Chakraborty, and Sing, 1986; George et al., 1999; Fulker et al., 1999; Sham et al., 2000). Based on combinations

of population and pedigree data, we have proposed variance component models for combined linkage and association mapping of QTL for complex diseases (Fan and Jung, 2003; Fan, Jung, and Jin, 2006; Fan et al., 2005; Fan and Xiong, 2002; Fan and Xiong, 2003; Jung, Fan, and Jin, 2005).

However, there is limited research to investigate the impact of missing data on our models. In genetics study, the genotypes or phenotypes can be missing due to various reasons. In Chapter IV, the impact of missing genotypes is investigated for high resolution combined linkage and association mapping of quantitative trait loci (QTL). We assume that the genotype data are missing completely at random (MCAR). Two regression models, "genotype effect model" and "additive effect model", are proposed to model the association between the markers and the trait locus. If the marker genotype is not missing, the model is exactly the same as those in previous study. If the marker genotypes are missing, the expected number of genotypes or alleles is used as weight to model the effect of the genotypes or alleles. By analytical formulae, we show that the "genotype effect model" can be used to model the additive and dominance effects simultaneously; the "additive effect model" only takes care of additive effect. Based on the two models, F-test statistics are proposed to test association between the QTL and markers. The noncentrality parameter approximations of F-test statistics are derived to make power calculation and comparison, which show that the power of the F-tests is reduced due to the missingness. By simulation study, we show that the two models have reasonable type I error rates for a dataset of moderate sample size. However, the type I error rates can be inflated if all individuals with missing genotypes are removed from analysis. Hence, the proposed method can help to get correct type I error rates although it does not improve power. As a practical example, the method are applied to analyze the angiotensin-1 converting enzyme (ACE) data.

1.3 Dissertation Structure

Chapter II develops a sufficiency score method to a measurement error problem in a generalized partially linear model framework. This chapter mainly contains a methodological proposal, a data analysis employing this method, and a complementary simulation experiment evaluating the methodology. Chapter III investigates the impact of missing genotypes in a combined linkage and association mapping study. This chapter consists of regression models accounting for missingness, hypothesis F-tests of associations, a simulation study and some data analyses employing the proposed models. Chapter IV gives a summary of the dissertation. Regularity conditions and proofs of the theorems are detailed in the appendices.

CHAPTER II

A SUFFICIENCY SCORE METHOD IN GENERALIZED PARTIALLY LINEAR MODELS WITH MEASUREMENT ERROR*

2.1 Introduction

In this chapter, we study a measurement error problem in a generalized partially linear model framework. More specifically, we consider a partially linear canonical exponential family models with covariate measurement errors. In the parametric problem, Stefanski and Carroll (1987) constructed unbiased score functions by conditioning on certain parameter-dependent sufficient statistics, a technique they called sufficiency scores. The purpose of the research in this chapter is to generalize their method to semiparametric generalized partially linear models. The resulting methods are straightforward to compute and have relatively clean asymptotic properties.

We start with a canonical exponential family for a response Y. Given a covariate vector $(X^{\mathrm{T}}, Z)^{\mathrm{T}} = (x^{\mathrm{T}}, z)^{\mathrm{T}}$, assume that the response Y has the density/mass function

$$h(y|x,z) = \exp\left[\frac{y\{x^{\mathrm{T}}\beta + \theta(z)\} - b\{x^{\mathrm{T}}\beta + \theta(z)\}}{a(\phi)} + c(y,\phi)\right], \tag{2.1}$$

where $a(\cdot), b(\cdot), c(\cdot,\cdot)$ are known functions, $\theta(\cdot)$ is an unknown nuisance nonparametric function and $\kappa = (\beta^{\mathrm{T}}, \phi)$ is the parameter of interest. Then the conditional mean and variance functions of Y given (X, Z) can be defined by

$$\begin{aligned} E(Y|X,Z) &= \mu\{X^{\mathrm{T}}\beta + \theta(Z)\}; \\ \mathrm{var}(Y|X,Z) &= \phi V[\mu\{X^{\mathrm{T}}\beta + \theta(Z)\}], \end{aligned}$$

* This article was in press in Statistics and Probability Letters, Liu, L., "Estimation of generalized partially linear models with measurement error using sufficiency scores", Elsevier (2007).

where $\mu(\cdot)$ and $V(\cdot)$ are known functions. Suppose, however, that the covariate X cannot be observed but that $W = X + U$ is available to the study. And U is assumed to be normally distributed with mean zero and covariance matrix Σ_{uu}, independent of (Y, X, Z).

So, summarizing things together, we have

$$\begin{aligned} Y|X, Z &\sim h(y|x, z); \\ W &= X + U; \\ U &= \text{Normal}(0, \Sigma_{\text{uu}}), \end{aligned}$$

and (Y, W, Z) are observable.

This defines a generalized partially linear measurement error model, or semiparametric measurement error model, in a canonical exponential family framework. The density of canonical exponential family as in (2.1) includes normal, Poisson, logistic and gamma regression models. These models have a common property that there exists a natural sufficient statistic for the unobserved covariate X when other parameters are fixed. This is crucial because the goal of the study is to develop a functional method to estimate the unknown parameters with no distribution assumption of the unobserved covariate X. By the property of sufficiency we are able to achieve this goal.

In the reminder of this chapter, a brief literature review is presented in Section 2.2 and the proposed method is developed in Section 2.3. Asymptotic properties of our methodology are presented in Section 2.4. Section 2.5 describes a data analysis of the Framingham Heart Study, with a complementary simulation study performed in Section 2.6. Section 2.7 gives concluding remarks. All the technical details are listed in Appendix A.

2.2 Literature Review

Severini and Staniswalis (1994) studied quasi-likelihood estimation in semiparametric models without measurement error. Stefanski and Carroll (1987) studied the measurement error in a parametric model.

Let us review the results in Stefanski and Carroll (1987). In parametric model, $\theta(Z)$ is reduced to a fixed parameter and Σ_{uu} is assumed as known. Let $Z = z_0$ and $\theta(z_0) = \alpha_0$. If X is viewed as a parameter and α_0, κ as fixed, then the statistic

$$\Delta = \Delta(Y, W, \Sigma_{uu}, \beta) = W + Y\Sigma_{uu}\beta \tag{2.2}$$

is complete and sufficient for X. Consequently, by the properties of sufficient statistics, the conditional distribution of Y given Δ does not depend on X and can be computed analytically.

Let $h_{Y|\Delta}(y|\delta; \alpha_0, \kappa)$ denote the conditional distribution of Y given $\Delta = \delta$. In the calculations, Δ is treated as a fixed conditioning argument until the final step of the analysis, equation (2.4), wherein Δ is evaluated as $\Delta = W + Y\Sigma_{uu}\beta$, as in equation (2.2). It is easy to show that

$$h_{Y|\Delta}(y|\delta; \alpha_0, \kappa) = \exp[y\eta - \frac{1}{2}y^2\beta^{\mathrm{T}}\Sigma_{uu}\beta/a(\phi) + c(y, \phi) - \log\{S(\eta, \beta, \phi)\}], \tag{2.3}$$

where $\eta = (\alpha_0 + \delta^{\mathrm{T}}\beta)/a(\phi)$ and $S(\eta, \beta, \phi)$ is the normalizing constant.

By defining $\Psi(y, \delta, \alpha_0, \kappa) = \dfrac{\partial}{\partial(\alpha_0, \kappa)}\log\{h_{Y|\Delta}(y|\delta; \alpha_0, \kappa)\}$ evaluated at $\delta = w + y\Sigma_{uu}\beta$, Stefanski and Carroll (1987) define the sufficiency score by

$$\Psi(Y, \Delta, \alpha_0, \kappa) = \begin{bmatrix} \{Y - E_\delta(Y)\}/a(\phi) \\ \{Y - E_\delta(Y)\}\delta/a(\phi) - \{Y^2 - E_\delta(Y^2)\}\Sigma_{uu}\beta/a(\phi) \\ r(Y, \Delta, \alpha_0, \kappa) - E_\delta\{r(Y, \Delta, \alpha_0, \kappa)\} \end{bmatrix}, \tag{2.4}$$

evaluated at $\delta = W + Y\Sigma_{uu}\beta$, where $E_\delta(\cdot) = E(\cdot|\Delta = \delta)$ and

$$r(y, w, \alpha_0, \kappa) = \frac{\partial c(y, \phi)}{\partial \phi} - y\frac{\alpha_0 + \delta^{\mathrm{T}}\beta}{a^2(\phi)}a'(\phi) + y^2\frac{\beta^{\mathrm{T}}\Sigma_{uu}\beta}{2a^2(\phi)}a'(\phi).$$

Also $\Psi(\cdot)$ is unbiased for $(\alpha_0, \kappa) = \{\theta(z_0), \kappa\}$, that is

$$E\{\Psi(Y, \Delta, \alpha_0, \kappa)\} = E[E\{\Psi(Y, \Delta, \alpha_0, \kappa)|\Delta\}] = 0.$$

Equation (2.4) defines an unbiased sufficiency score when $\theta(\cdot)$ is treated as a parameter α_0. Therefore, sufficiency score estimators $(\widehat{\alpha}_0, \widehat{\kappa})$ are solved by

$$0 = \sum_{i=1}^{n} \Psi(Y_i, \Delta_i, \widehat{\alpha}_0, \widehat{\kappa}).$$

2.3 Backfitting Method

In this section, we generalize the sufficiency score method to handle the semiparametric situation when $\alpha_0 = \theta(Z)$ is a function of Z. For simplicity we assume that Σ_{uu} is known throughout this chapter. This is of course an ideal assumption to make the study easier. In the data analysis we simply use the estimate of Σ_{uu} as the true value. We will revisit this issue at the end of this chapter.

Since the parameter of interest β is involved in the expression of Δ, we denote the sufficiency score in the previous section as $\Psi\{Y, \Delta(\beta), \theta(Z), \kappa\}$. Also, write

$$\Psi\{Y, \Delta(\beta), \theta(Z), \kappa\} \quad = \quad \begin{bmatrix} \Psi_\theta\{Y, \Delta(\beta), \theta(Z), \kappa\} \\ \Psi_\kappa\{Y, \Delta(\beta), \theta(Z), \kappa\} \end{bmatrix},$$

where $\Psi_\theta(\cdot)$ is the estimating function corresponding to α_0 or $\theta(\cdot)$ and $\Psi_\kappa(\cdot)$ is the estimating function corresponding to κ. Suppose that $K_h(x) = K(x/h)/h$, $K(\cdot)$ is a symmetric (kernel) density function with compact support, h is the bandwidth, $h \to 0$ as $n \to \infty$, and $Z_1, \cdots, Z_n$ have marginal density $f_z(\cdot)$.

We propose a two-step iterative estimation procedure as below.

1. First we estimate $\theta(\cdot)$ by $\widehat{\theta}(\cdot, \kappa_c)$ with a current value $\kappa_c = (\beta_c^{\mathrm{T}}, \phi_c)^{\mathrm{T}}$ in the iteration. Adapting the idea of local estimating equation of Carroll, Ruppert, and

Welsh (1998), $\widehat{\theta}(z_0, \kappa_c)$ is the intercept α_0 in solving the local linear likelihood equation

$$\begin{aligned} 0 &= \sum_{i=1}^{n} K_h(Z_i - z_0) \begin{bmatrix} 1 \\ (Z_i - z_0)/h \end{bmatrix} \\ &\quad \times \Psi_\theta\{Y_i, \Delta_i(\beta), \alpha_0 + \alpha_1(Z_i - z_0)/h, \kappa_c\}. \end{aligned} \tag{2.5}$$

Here α_1 is an estimate of $h\theta'(z_0)$. At the solution, for any κ we have that

$$0 = E[\Psi_\theta\{Y, \Delta(\beta), \theta(z_0, \kappa), \kappa\}]. \tag{2.6}$$

2. Next we update κ with given $\widehat{\theta}(\cdot, \kappa_c)$ by solving

$$0 = \sum_{i=1}^{n} \Psi_\kappa\{Y_i, \Delta_i(\beta_c), \widehat{\theta}(Z_i, \kappa_c), \kappa\}, \tag{2.7}$$

in κ.

We solve equations (2.5) and (2.7) iteratively until convergence to obtain the back-fitting estimator $\widehat{\kappa}$ and the fitted function $\widehat{\theta}(\cdot)$.

To get the starting value of κ in the iteration, we suggest to solve some naive version of the problem. For example, we may pretend that there is no measurement error and the nonparametric function is quadratic, that is, setting $\Sigma_{uu} = 0$ and $\theta(Z) = \theta_0 + \theta_1 Z + \theta_2 Z^2$. Then

$$E(Y|X, Z) = \mu\{X^{\mathrm{T}}\beta + \theta_0 + \theta_1 Z + \theta_2 Z^2\}.$$

Therefore, this naive version of the problem is simply a generalized linear model. So it will be easy and quick to get a reasonable starting value.

2.3.1 Logistic Regression

When the response variable in the study is binary, logistic regression is a commonly used model. We illustrate this sufficiency score method with logistic regression. Sup-

pose that

$$\Pr(Y = 1|X, Z) = H\{X^{\mathrm{T}}\beta + \theta(Z)\},$$

where $H(x) = e^x/(1+e^x)$ is the logistic distribution function. For this model, $a(\phi) = 1$, and the parameter of interest κ is reduced to β. Then, derived from (2.3) we have $\Pr(Y = 1|\Delta = \delta) = H\{\theta(Z) + (\delta - \Sigma_{uu}\beta/2)^{\mathrm{T}}\beta\}$, and corresponding to (2.4) is the logistic sufficiency score

$$\begin{aligned}\Psi\{Y, \Delta(\beta), \theta(Z), \beta\} &= \begin{bmatrix}\Psi_\theta(\cdot)\\ \Psi_\beta(\cdot)\end{bmatrix}\\ &= [Y - H\{\theta(Z) + (\Delta(\beta) - \Sigma_{uu}\beta/2)^{\mathrm{T}}\beta\}]\begin{bmatrix}1\\ \Delta(\beta) - \Sigma_{uu}\beta\end{bmatrix},\end{aligned}$$

evaluated at $\Delta(\beta) = W + Y\Sigma_{uu}\beta$. It has an equivalent estimating equation given as

$$\sum_{i=1}^{n}[Y_i - H\{\theta(Z) + \beta^{\mathrm{T}}\Delta_i^*(\beta)\}]\begin{bmatrix}1\\ \Delta_i^*(\beta)\end{bmatrix} = 0, \tag{2.8}$$

where $\Delta_i^*(\beta) = \Delta_i(\beta) - \Sigma_{uu}\beta/2$. Conditioned on $\Delta_i^*(\beta)$, Y_i is Bernoulli distributed with mean $H\{\theta(Z) + \beta^{\mathrm{T}}\Delta_i^*(\beta)\}$.

Therefore, given the current value β_c the two-step iterative estimating equations (2.5) and (2.7) are simplified as

$$\begin{aligned}0 &= \sum_{i=1}^{n} K_h(Z_i - z_0)\begin{bmatrix}1\\ (Z_i - z_0)/h\end{bmatrix}\\ &\qquad \times [Y_i - H\{\alpha_0 + \alpha_1(Z_i - z_0)/h + \beta_c^{\mathrm{T}}\Delta_i^*(\beta_c)\}]; \qquad (2.9)\\ 0 &= \sum_{i=1}^{n}[Y_i - H\{\widehat{\theta}(Z_i, \beta_c) + \beta^{\mathrm{T}}\Delta_i^*(\beta_c)\}]\Delta_i^*(\beta_c). \qquad (2.10)\end{aligned}$$

It is easy to see that solving equation (2.9) in $\widehat{\theta}(z_0, \beta_c) = \alpha_0$ is simply a logistic regression problem with weights $K_h(Z_i - z_0)$ and offsets $\beta_c^{\mathrm{T}}\Delta_i^*(\beta_c)$. Similarly, solving

equation (2.10) in β is a logistic regression problem with offsets $\widehat{\theta}(Z_i, \beta_c)$ but without intercept, thus simplifying computation.

2.4 Asymptotic Results

In this section we describe the limiting distribution of $\widehat{\kappa}$ and $\widehat{\theta}(\cdot)$. Let κ_0 and $\theta_0(\cdot)$ be the true parameter and function. Define argument $(\bullet)$ as $\{Y, \Delta(\beta_0), \theta_0(Z), \kappa_0\}$ and $(\bullet_i)$ as $\{Y_i, \Delta_i(\beta_0), \theta_0(Z_i), \kappa_0\}$. Also let $\Psi_{\kappa\theta}(\cdot)$ be the partial derivative of $\Psi_\kappa(\cdot)$ with respect to the term $\theta(\cdot)$ and similarly for $\Psi_{\theta\theta}(\cdot)$, $\Psi_{\theta\kappa}(\cdot)$, etc. Also define

$$\begin{aligned} D(z) &= E[\{\Psi_{\theta\Delta}(\bullet)\frac{\partial\Delta}{\partial\kappa} + \Psi_{\theta\kappa}(\bullet)\}|Z=z]/E\{\Psi_{\theta\theta}(\bullet)|Z=z\}; \\ \mathcal{U}(z) &= E\{\Psi_{\kappa\theta}(\bullet)|Z=z\}/E\{\Psi_{\theta\theta}(\bullet)|Z=z\}; \\ \mathcal{F} &= E\{\frac{d}{d\kappa}\Psi_\kappa(\bullet)\} = E\{\Psi_{\kappa\Delta}(\bullet)\frac{\partial\Delta}{\partial\kappa} + \Psi_{\kappa\kappa}(\bullet) - \Psi_{\kappa\theta}(\bullet)D(Z)\}, \end{aligned}$$

where $\partial\Delta/\partial\kappa = (Y\Sigma_{uu}, 0)$ is independent of κ.

Lemma 2.4.1. *The derivative of the curve $\theta(z,\kappa)$ satisfies*

$$\frac{\partial}{\partial\kappa}\theta(z,\kappa) = -D(z).$$

Proof. The lemma follows by differentiating equation (2.6) with respect to κ and solving the resulting equation. □

Theorem 2.4.2. *Assume that the bandwidth h satisfies $nh^4 \to 0$ and $nh^2/log^2(n) \to \infty$. Then under the regularity conditions outlined in the Appendix A.1, the backfitting estimator $\widehat{\kappa}$ has the asymptotic expansion*

$$-\mathcal{F}n^{1/2}(\widehat{\kappa} - \kappa_0) = n^{-1/2}\sum_{i=1}^{n}\{\Psi_\kappa(\bullet_i) - \Psi_\theta(\bullet_i)\mathcal{U}(Z_i)\} + o_p(1). \qquad (2.11)$$

Hence $n^{1/2}(\widehat{\kappa} - \kappa_0)$ is asymptotically normally distributed with mean zero and covariance matrix $\mathcal{F}^{-1}\Sigma\mathcal{F}^{-\mathrm{T}}$, where $\Sigma = cov\{\Psi_\kappa(\bullet) - \Psi_\theta(\bullet)\mathcal{U}(Z)\}$.

Remark. The condition $nh^4 \to 0$ is typically necessary for the backfitting method. From the proof in the Appendix A we can see this undersmoothing of $\theta(\cdot)$ is a direct result of the bias of the nonparametric regression estimator, which is of order $O(h^2)$. In order for (2.11) to hold, we used $n^{1/2}h^2 \to 0$, i.e., $nh^4 \to 0$. The proof of Theorem 2.4.2 is given in Appendix A.3.

Consistent estimators of $\mathcal{F}$ and Σ can be constructed as follows. Define argument $(\widehat{\bullet}_i)$ as $\{Y_i, \Delta_i(\widehat{\beta}), \widehat{\theta}(Z_i, \widehat{\kappa}), \widehat{\kappa}\}$. First we estimate the conditional expectations in the definitions by fitting smooth functions of Z, e.g., $E\{\Psi_{\theta\theta}(\bullet)|Z\}$ is estimated by fitting a smooth function with responses $\Psi_{\theta\theta}(\widehat{\bullet}_i)$ and predictors Z_i using kernel regression. Then we obtain $\widehat{D}(z)$ and $\widehat{\mathcal{U}}(z)$ for $z = Z_1, \cdots, Z_n$ by plugging in the ratio of estimated conditional expectations. Because all the kernel regressions result in consistent estimation, a consistent estimator of $\mathcal{F}$ is

$$\widehat{\mathcal{F}} = n^{-1} \sum_{i=1}^{n} \{\Psi_{\kappa\Delta}(\widehat{\bullet}_i) \frac{\partial \Delta}{\partial \kappa} + \Psi_{\kappa\kappa}(\widehat{\bullet}_i) - \Psi_{\kappa\theta}(\widehat{\bullet}_i) \widehat{D}(Z_i)\}.$$

Further a consistent estimator of Σ is the sample covariance matrix of the terms

$$\Psi_{\kappa}(\widehat{\bullet}_i) - \Psi_{\theta}(\widehat{\bullet}_i) \widehat{\mathcal{U}}(Z_i).$$

The consistency of the estimators of $\mathcal{F}$ and Σ follows because of the uniformly consistency of the nonparametric function estimators.

Theorem 2.4.3. *Under the regularity conditions outlined in the Appendix A.1, the fitted function $\widehat{\theta}(z, \kappa_0)$ has the asymptotic expansion*

$$\widehat{\theta}(z, \kappa_0) - \theta_0(z) = \frac{h^2}{2} \theta_0''(z) - \frac{n^{-1} \sum_{i=1}^{n} K_h(Z_i - z) \Psi_{\theta}(\bullet_i)}{f_z(z) E[\Psi_{\theta\theta}(\bullet)|Z = z]} + o_p(n^{-1/2}).$$

Thus, the asymptotic bias and variance of $\widehat{\theta}(z, \kappa_0)$ are

$$\begin{aligned} E\{\widehat{\theta}(z, \kappa_0)\} - \theta_0(z) &= \frac{h^2}{2} \theta_0''(z) + o(h^2), \\ var\{\widehat{\theta}(z, \kappa_0)\} &= \frac{g}{nh f_z(z)} \frac{var\{\Psi_{\theta}(\bullet)|Z = z\}}{E^2\{\Psi_{\theta\theta}(\bullet)|Z = z\}} + o\{(nh)^{-1}\}, \end{aligned}$$

where $g = \int K^2(s)ds$.

The proof of Theorem 2.4.3 is given in Appendix A.2.

2.5 Data Analysis

In this section, we applied our methods to the Framingham Heart Study data, which has a long history in measurement error modeling, see Carroll et al. (2006) for a review. In this study, the response variable Y is the indicator of first evidence of coronary heart disease (CHD) occurring at Exam 3 through Exam 6. We use the age as our covariate Z under a linear transformation to range $(0, 1)$. Another good covariate in this study is the systolic blood pressure. Since it is impossible to measure the long-term systolic blood pressure X, we treat the systolic blood pressure measured at Exam 3 as the observed covariate W. The difference between the long-term systolic blood pressure X and the single-visit W is due to the daily and seasonal variation of the blood pressure. We also apply a standard transformation to the raw measurements as

$$W = \log(W_{\text{raw}} - 50) - \text{mean}\{\log(W_{\text{raw}} - 50)\},$$

such that the measurement error U is normally distributed with mean zero and variance Σ_{uu}. Carroll et al. (2006) estimate $\Sigma_{uu} = 0.0126$ based on 1,615 degrees of freedom, so we consider Σ_{uu} as known in this analysis. We choose the Epanechnikov kernel function, apply equation (2.8) and solve $\widehat{\beta}$ and $\widehat{\theta}(\cdot)$ iteratively by equations (2.5) and (2.7). Theorem 2.4.2 is used to estimate the standard error of β and to determine the bandwidth level. Framingham study has 1,615 subjects. Since $1615^{-1/4} \approx 0.2$, We choose several bandwidths around $h = 0.2$ for the analysis.

The estimates of β's and the standard errors for different bandwidths are summarized in Table 1. We can see for all these bandwidths the estimated $\widehat{\beta}$ is close to 2.00. Also, the estimated standard error is around 0.45 and it is roughly independent

Table 1: Framingham data parameter estimation. Estimates $\widehat{\beta}$ and the standard errors for different bandwidth h.

Bandwidth h	Estimate $\widehat{\beta}$	Standard Error
0.1	2.06	0.45
0.2	2.00	0.45
0.3	2.00	0.45
0.4	2.08	0.46
0.5	2.14	0.47

of the bandwidth. The estimated function $\widehat{\theta}(\cdot)$ is somewhat curved. See Figure 1 for the curve for bandwidth $h = 0.2$. Note that we transform variable age back to the original value.

2.6 Simulation Study

In our simulation study, we construct a framework which is similar to the Framingham data analyzed in Section 2.5 above. We assume that the true disease status Y follows a logistic relationship by $\Pr(Y = 1|X, Z) = H\{X\beta+\theta(Z)\}$. We set the true parameter $\beta = 2.0$ and the true function $\theta(z) = -4.15 + 4.60z - 2.35z^2$. The reason we use this quadratic function is because the curve of this function in the range $(0, 1)$ is close to the $\widehat{\theta}(\cdot)$ curve in the Framingham data analysis. Also, the transformed age Z and the measurement error variance Σ_{uu} are set exactly the same as those in Framingham data. The true transformed blood pressure X is set to the average of the measurements of Exam 2 and Exam 3. In a single simulation, we randomly generate disease status Y using the Bernoulli distribution with success probability $H\{X\beta + \theta(Z)\}$, and the measured transformed blood pressure W using the normal distribution with mean X and variance Σ_{uu}. We obtain the estimate $\widehat{\beta}$ and the standard error SE using the proposed method. Then we construct a 95% confidence interval of β as $\widehat{\beta} \pm (1.96)$SE and record whether this interval covers the true β or not. We ran 1,000 simulations and calculated the coverage probability and the

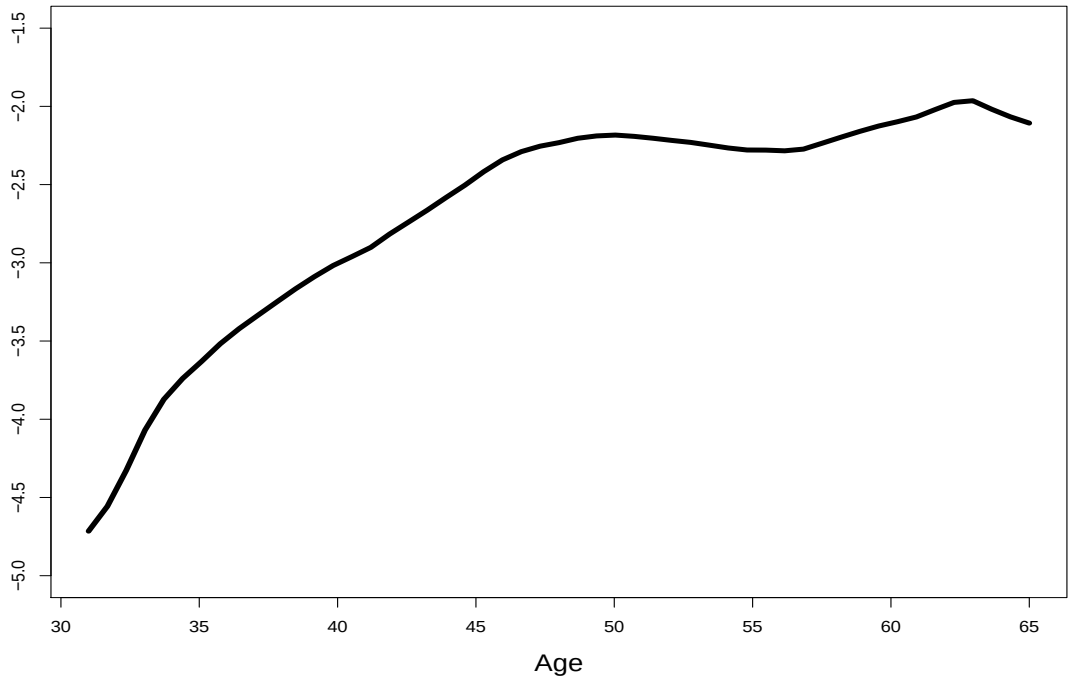

Figure 1: Framingham data function fits. Estimated function $\widehat{\theta}(\text{age})$ versus age for bandwidth $h = 0.2$.

standard errors. Similar to the real data analysis, we selected several bandwidths around $h = 0.2$. We also compared the results with those using a naive method. The naive method simply assumed that there is no measurement error, i.e., $\Sigma_{uu} = 0$. The results were summarized in Table 2.

From Table 2 we can see that the coverage probability for β using proposed method is very close to the nominal level 0.95, much better than those using the naive method. Also, the Monte Carlo standard error of the $\widehat{\beta}$ is very close to the sample mean of the estimated standard error of $\widehat{\beta}$ using Theorem 2.4.2 for all the bandwidths, indicating the effectiveness of our method for standard error estimation. The results of the nonparametric function estimation $\widehat{\theta}(\cdot)$ are as expected. When the bandwidth is too small, the curves become rougher ($h \leq 0.1$), while there is little

Table 2: Simulation results. Coverage probability of 95% confidence intervals of β using naive method and proposed method, Monte Carlo SE of all $\widehat{\beta}$'s and the average of the estimated SE of each $\widehat{\beta}$ for different bandwidth h.

Bandwidth h	Naive %	Proposed %	SE of $\widehat{\beta}$'s	$\overline{\text{SE}}(\widehat{\beta})$
0.05	0.88	0.94	0.50	0.47
0.1	0.84	0.95	0.48	0.46
0.2	0.87	0.95	0.48	0.46
0.3	0.88	0.95	0.48	0.47
0.4	0.89	0.95	0.48	0.47
0.5	0.89	0.94	0.49	0.47

effect for $h \geq 0.2$. We use $h = 0.2$ for illustration. Note, however, that as explained above, the bandwidth has little effect on the estimate of β. A plot of the mean $\widehat{\theta}(\cdot)$ over the 1,000 simulations with the true function $\theta(\cdot)$ is in Figure 2. As we expect, they are very similar.

2.7 Discussion

In this chapter, we have developed a backfitting method for generalized partially linear models with independent normal measurement error. The estimating equations are constructed in (2.5) and (2.7). The main asymptotic results about the parameters are summarized in Theorem 2.4.2 and 2.4.3. In Appendix A we showed in detail that the undersmoothing is needed for backfitting. We applied our method to Framingham Heart Study and a simulation study and the results verified our methodology very well. The main advantage of this method is the simple therefore fast computation and the relatively clean asymptotic theorem.

In many cases, we will have replicate measurements. Suppose that there are $m_i \geq 1$ replications for subject i, and denote $\delta_i = I(m_i \geq 2)$. Let s_{ui} be the sample covariance matrix for subject i. Then an unbiased estimating equation for Σ_{uu} is

$$0 = \sum_{i=1}^{n} \delta_i (s_{ui} - \Sigma_{uu}). \tag{2.12}$$

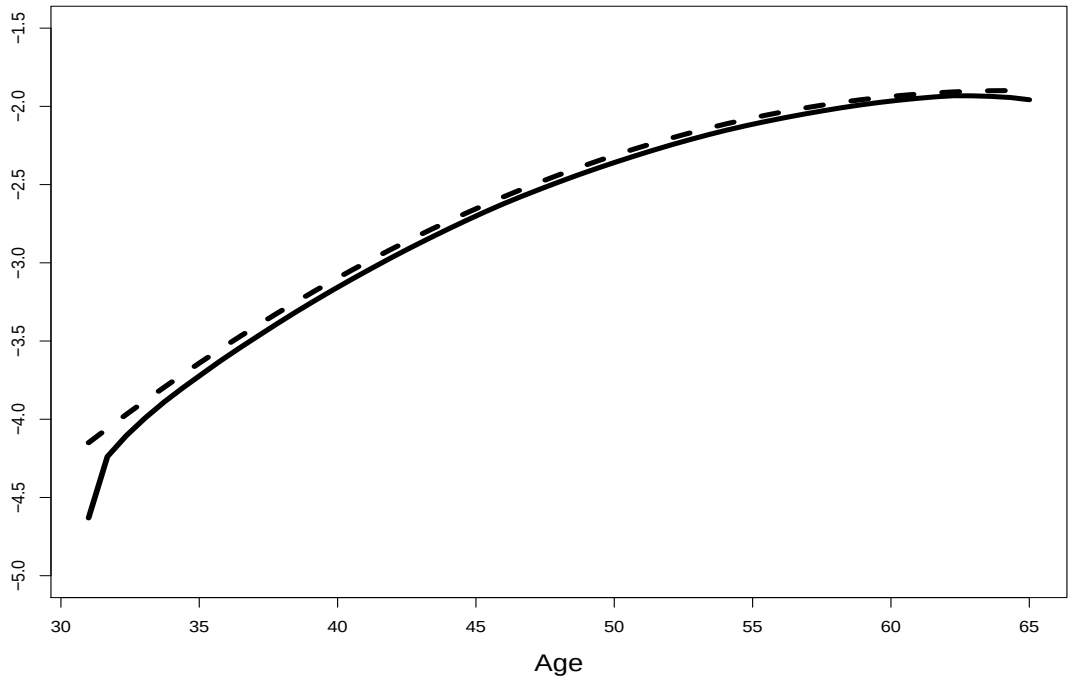

Figure 2: Function fits for simulation. The solid line is the average of the estimated function $\widehat{\theta}$(age) over 1000 simulations, while the dashed line is the true function.

Thus, to allow for estimation of the measurement error covariance matrix, one merely appeals (2.12) to equations (2.5) and (2.7). In the future more research is needed to address this issue.

CHAPTER III

COMBINED LINKAGE AND ASSOCIATION MAPPING OF QUANTITATIVE TRAIT LOCI WITH MISSING GENOTYPE DATA

3.1 Introduction

In disease gene mapping, linkage analysis and linkage disequilibrium mapping (or association study) can be carried out. Linkage analysis is based on pedigree data, and association study can be based on either population data or pedigree data or combinations of population and pedigree data. Linkage analysis is robust to population structure, and is appropriate for low resolution genetic mapping to localize trait loci to broad chromosome regions within a few centiMorgan (cM). In contrast to linkage analysis, association study for genetic traits is useful in high resolution of gene mapping, i.e., fine disease gene mapping; however, association study is prone to population structure and the false positives can be high. In recent years, there has been great interest in carrying out combined linkage and association mapping of complex genetic traits (Li, Boehnke, and Abecasis, 2005; Xiong and Jin, 2000). The combined analysis of linkage and association can take the advantage of the robustness of linkage analysis, and the high resolution of association study. In addition, it may minimize the limits of each.

However, there is limited research to investigate the impact of missing data on our models. In genetics study, the genotypes or phenotypes can be missing due to various reasons. It is important to develop models which account for missing data. In this chapter, we are going to develop models which account for missing data, and to investigate the impact of missing genotypes on combined linkage and association mapping of QTL. Two regression models, "genotype effect model" and "additive effect

model", are proposed to model the association between the markers and the trait locus when there are missing genotypes. Based on the two models, F-test statistics or likelihood ratio test statistics can be used to test association between the QTL and markers. We will investigate the impact of missing data on the models, under an assumption that the genotype data are missing completely at random (MCAR). Simulation study will be performed to evaluate the robustness of the proposed models, and to make comparison with models which exclude the individuals with missing genotypes from analysis. In addition, the method will be applied to analyze the angiotensin-1 converting enzyme data (Farrall et al., 1999; Keavney et al., 1998).

3.2 Models

Consider a quantitative trait locus Q, which is located at an autosome. Suppose that there are two alleles Q_1 and Q_2 at the trait locus with frequencies q_1 and q_2, respectively. In a region of the QTL Q, suppose that one marker or multiple markers are typed for a sample; and the sample may include multi-generation pedigrees of any sizes and any types of relatives, nuclear families, sibships and unrelated individuals. However, the marker information may be missing for some individuals of the sample at some markers. That is to say, some genotype information may not be available for some individuals. In multiple marker case, the genotypes of an individual may be missing at some markers and may be available at the other markers. In this chapter, we assume that the genotype data are missing completely at random (MCAR) (Little and Rubin, 2002), i.e., the missingness does not depend on the genotype and phenotype data. In the following, we first present the models by one marker, and extend to use two/multiple markers in analysis.

3.2.1 Log-likelihoods and Mapping Strategy

Suppose that the data are composed of a combination of N unrelated individuals and I independent families. The I families can be multi-generation pedigrees, nuclear families, sibships, or their combinations. Let us list the log-likelihoods of the N individuals by $L_1, \cdots, L_N$, and the log-likelihoods of the I families by $L_{N+1}, \cdots, L_{N+I}$. The overall log-likelihood is $L = \sum_{i=1}^{N+I} L_i$. In the i-th family, let t_i be the total number of individuals who are listed as $j = 1, 2, \cdots, t_i$; each individual j is preceded by all his/her ancestors. Let us denote the quantitative traits of i-th family by a vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{it_i})^{\mathrm{T}}$. In addition, assume that marker genotypes are either available or missing for a family member. The log-likelihood is defined by $L_i = -\frac{t_i}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{y}_i - X_i\phi)^{\mathrm{T}}\Sigma_i^{-1}(\mathbf{y}_i - X_i\phi)$, under the assumption of multivariate normality. In the log-likelihood, Σ_i is the variance-covariance matrix which is defined in the paragraph below; X_i is a model matrix defined in Subsections 3.2.2 and 3.2.3, and ϕ is a column vector of regression coefficients related to the model matrix.

Σ_i is a $t_i \times t_i$ matrix defined as

$$\Sigma_i = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1t_i} \\ \rho_{12} & 1 & \cdots & \rho_{2t_i} \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{1t_i} & \rho_{2t_i} & \cdots & 1 \end{bmatrix} \sigma^2,$$

where $\sigma^2 = \sigma_g^2 + \sigma_{Ga}^2 + \sigma_e^2$, σ_g^2 is variance explained by the putative QTL Q, σ_{Ga}^2 is polygenic additive variance, and σ_e^2 is error variance. The genetic variance $\sigma_g^2 = \sigma_{ga}^2 + \sigma_{gd}^2$ is decomposed into additive and dominance components. As in the traditional quantitative genetics, let a be the effect of genotype Q_1Q_1, d be the effect of genotype Q_1Q_2, and $-a$ be the effect of genotype Q_2Q_2 (Falconer and Mackay, 1996). Let $\alpha_Q =$

$a+(q_2-q_1)d$ be the average effect of gene substitution, and $\delta_Q = 2d$ be the dominance deviation. In addition, let $\mu = a(q_1 - q_2) + 2dq_1q_2$ be the aggregate effect of the QTL on the trait mean in the population. It is well known that the additive variance $\sigma^2_{ga} = 2q_1q_2\alpha^2_Q$ and the dominance variance $\sigma^2_{gd} = (q_1q_2)^2\delta^2_Q$. $\rho_{jk} = (\pi_{jkQ}\sigma^2_{ga} + \Delta_{jkQ}\sigma^2_{gd} + 2\Phi_{jk}\sigma^2_{Ga})/\sigma^2$ is correlation between the j-th individual and the k-th individual of the family, where π_{jkQ} is the proportion of alleles shared identically by descent (IBD) at QTL Q by the j-th and the k-th individuals, Δ_{jkQ} is the probability that both alleles at QTL Q shared by the j-th and the k-th individuals are IBD, and Φ_{jk} is the kinship coefficient of individuals j and k. π_{jkQ} and Δ_{jkQ} are usually estimated by marker information (Amos, 1994; Amos and Elston, 1989). The recombination fractions between the genotyped markers and the unobserved QTL are contained in the estimations of π_{jkQ} and Δ_{jkQ}. Hence, linkage information is modeled in variance-covariance matrix.

For the N unrelated individuals, the log-likelihoods are $L_i = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_{i1} - X_i\phi)^{\mathrm{T}}(y_{i1} - X_i\phi), i = 1, \cdots, N$. Here, y_{i1} is the trait value of the i-th individual. It can be seen that no linkage information is contained in the log-likelihoods of the N unrelated individuals. The linkage is modeled solely in variance-covariance matrices of the I families. Therefore, family data can be used for linkage analysis. In Subsections 3.2.2 and 3.2.3, we will show that linkage disequilibrium information is contained in the regression coefficients ϕ. Thus, linkage disequilibrium information is contained in both population data and family data. The linkage analysis can usually locate the trait locus in a broad chromosome region within a few cM or even around 15cM. Linkage analysis is less sensitive to population structures of subdivisions and admixtures, although its resolution can be low. In contrast, the linkage disequilibrium analysis has an advantage for high resolution mapping of trait locus, but can be prone to false positives. In practice, linkage analysis can be performed as the first

step of analysis to obtain suggestive linkage information. With evidence of suggestive linkage from linkage study, population data and family data can be combined together for linkage disequilibrium analysis for fine mapping of QTL. Using this strategy to map the trait locus, one may take advantage of both linkage analysis and linkage disequilibrium mapping, and be more likely to avoid the spurious association.

3.2.2 Mixed Effect Models by One Marker

In a region of the QTL Q, suppose that one marker A is typed, which may be di-allelic or multi-allelic. Let us denote the alleles of marker A by $A_1, \cdots, A_m$, where m is the number of alleles. Suppose that the marker A is in Hardy-Weinberg equilibrium (HWE). Let the frequency of A_g be $P_{A_g}, g = 1, 2, \cdots, m$. Consider the j-th pedigree member of the i-th family with trait value y_{ij} and genotype G_{Aij}. If the genotype G_{Aij} is not missing, there are $J_A = m(m+1)/2$ possibilities for G_{Aij}, which can be listed as $A_1A_1, \cdots, A_mA_m, A_1A_2, \cdots, A_1A_m, \cdots, A_{m-1}A_m$. In practice, the genotype G_{Aij} can be missing. Therefore, the genotype G_{Aij} can be one of the J_A genotypes if it is not missing and can be missing. If G_{Aij} is missing, we denote it by $G_{Aij} =?$; and if G_{Aij} is not missing, we denote it by $G_{Aij} \neq ?$, i.e., the complementary set of $G_{Aij} =?$. Let us denote the probability that the genotype G_{Aij} is missing by ε_A, i.e., $P(G_{Aij} =?) = \varepsilon_A$. Notice that $P(G_{Aij} \neq ?) = 1 - \varepsilon_A$. In addition, let $P(G_{Aij} = A_gA_h | G_{Aij} =?)$ or $P(G_{Aij} = A_gA_h | G_{Aij} \neq ?)$ be the conditional probability of genotype A_gA_h given $G_{Aij} =?$ or $G_{Aij} \neq ?$. Since the missing mechanism is MCAR, the probability

$$\begin{aligned} P(G_{Aij} = A_gA_h | G_{Aij} =?) &= P(G_{Aij} = A_gA_h | G_{Aij} \neq ?) \\ &= P(A_gA_h) = \begin{cases} P_{A_g}^2 & \text{if } g = h \\ 2P_{A_g}P_{A_h} & \text{if } g \neq h \end{cases}. \end{aligned}$$

Genotype Effect Model. For the listed J_A genotypes, let $\beta_{11}, \cdots, \beta_{mm}, \beta_{12}, \cdots, \beta_{1m}, \cdots, \beta_{m-1,m}$ be the corresponding effects on quantitative trait. The "genotype effect model" can be written as

$$\begin{aligned} y_{ij} &= w_{ij}\gamma + \sum_{1\le g\le h\le m} [1_{(G_{Aij}=A_gA_h)} + P(G_{Aij}=A_gA_h|G_{Aij}=?)1_{(G_{Aij}=?)}]\beta_{gh} \\ &\quad + H_{ij} + e_{ij} \\ &= w_{ij}\gamma + \sum_{1\le g\le h\le m} [1_{(G_{Aij}=A_gA_h)} + P(A_gA_h)1_{(G_{Aij}=?)}]\beta_{gh} + H_{ij} + e_{ij}, \end{aligned} \tag{3.1}$$

where $1_E = \begin{cases} 1 & \text{if E is true} \\ 0 & \text{else} \end{cases}$ is indicator function, w_{ij} is a row vector of co-variates such as sex and age, γ is a column vector of regression coefficients of w_{ij}, H_{ij} is polygenic additive effect, and e_{ij} is the error term. Assume that H_{ij} is random normal $N(0, \sigma^2_{Ga})$, and e_{ij} is normal $N(0, \sigma^2_e)$. In addition, γ and β_{gh} are fixed effect. Hence, model (3.1) is a mixed effect model (Pinherio and Bates, 2000). The contribution of polygenic additive effect to the variance-covariance matrix Σ_i is from the terms which contain σ^2_{Ga}.

Now let us show that model (3.1) extends the "genotype effect model" in Fan et al. (2006). If the genotype is not missing and $G_{Aij} = A_gA_h$, the model (3.1) becomes $y_{ij} = w_{ij}\gamma + \beta_{gh} + H_{ij} + e_{ij}$, which is similar to "genotype effect model" (1), Fan et al. (2006). Note that the polygenic additive effect H_{ij} is not modeled in Fan et al. (2006). Since only population data are used in Fan et al. (2006), polygenic effect is not modeled to avoid redundancy and the models therein are fixed effect models. In this chapter, the polygenic effect is modeled as random effect and so the models are mixed effect models. Since we use both population data and family data, the polygenic additive effect is assumed to be estimable.

If G_{Aij} =? is missing, the model (3.1) is $y_{ij} = w_{ij}\gamma + \sum_{1\le g\le h\le m} P(A_gA_h)\beta_{gh} +$

$H_{ij}+e_{ij}$, which uses the conditional probability $P(G_{Aij}=A_gA_h|G_{Aij}=?)=P(A_gA_h)$ as the weight to model the effect β_{gh} of genotype A_gA_h. Let us denote

$$x_{Aij}^{(gh)} = 1_{(G_{Aij}=A_gA_h)}+P(A_gA_h)1_{(G_{Aij}=?)}, \tag{3.2}$$

which can be thought as the expected number of genotype A_gA_h given observed genotype G_{Aij} at marker A. The "genotype effect model" (3.1) can be re-written as $y_{ij}=w_{ij}\gamma+\sum_{1\le g\le h\le m}x_{Aij}^{(gh)}\beta_{gh}+H_{ij}+e_{ij}$. Here, we add the polygenic effect to the model proposed in Fan et al. (2006). Based on "genotype effect model" (3.1), we may get the model matrix X_i and regression coefficient vector ϕ as follows: $\phi=(\gamma^{\mathrm{T}},\beta_{11},\cdots,\beta_{mm},\beta_{12},\cdots,\beta_{1m},\cdots,\beta_{m-1,m})^{\mathrm{T}}$ and $X_i=(X_{Ai1},\cdots,X_{Ait_i})^{\mathrm{T}}$, where $X_{Aij}=(w_{ij},x_{Aij}^{(11)},\cdots,x_{Aij}^{(mm)},x_{Aij}^{(12)},\cdots,x_{Aij}^{(1m)},\cdots,x_{Aij}^{(m-1,m)})^{\mathrm{T}}, j=1,2,\cdots,t_i$.

Additive Effect Model. Assume that the genetic effect is additive, i.e., $\beta_{gh}=\alpha_g+\alpha_h$, where α_g is effect of allele A_g. The "additive effect model" can be written as

$$y_{ij}=w_{ij}\gamma+\sum_{1\le g\le h\le m}[1_{(G_{Aij}=A_gA_h)}+P(A_gA_h)1_{(G_{Aij}=?)}](\alpha_g+\alpha_h)+H_{ij}+e_{ij}. \tag{3.3}$$

If the genotype is not missing and $G_{Aij}=A_gA_h$, the model (3.3) becomes $y_{ij}=w_{ij}\gamma+\alpha_g+\alpha_h+H_{ij}+e_{ij}$, which is similar to "additive effect model", Fan et al. (2006). Therefore, model (3.3) extends the "additive effect model", Fan et al. (2006). If $G_{Aij}=?$ is missing, the model (3.3) is

$$\begin{aligned} y_{ij} &= w_{ij}\gamma+\sum_{1\le g\le h\le m}P(A_gA_h)(\alpha_g+\alpha_h)+H_{ij}+e_{ij} \\ &= w_{ij}\gamma+\sum_{g=1}^{m}P_{A_g}\alpha_g+\sum_{h=1}^{m}P_{A_h}\alpha_h+H_{ij}+e_{ij} \\ &= w_{ij}\gamma+\sum_{g=1}^{m}2P_{A_g}\alpha_g+H_{ij}+e_{ij}. \end{aligned}$$

Note that $2P_{A_g}=2P(G_{Aij}=A_gA_g|G_{Aij}=?)+\sum_{h\neq g}P(G_{Aij}=A_gA_h|G_{Aij}=?)$ is the expected number of alleles A_g given $G_{Aij}=?$, which is the weight to model the effect

α_g of allele A_g. Let us denote

$$x_{Aij}^{(g)} = 2 \cdot 1_{(G_{Aij}=A_gA_g)} + \sum_{h \neq g} 1_{(G_{Aij}=A_gA_h)} + 2P_{A_g} 1_{(G_{Aij}=?)}, \tag{3.4}$$

which is the expected number of alleles A_g given observed genotype G_{Aij} at marker A. From the discussion above, we may re-write the "additive effect model" (3.3) as $y_{ij} = w_{ij}\gamma + \sum_{g=1}^{m} x_{Aij}^{(g)} \alpha_g + H_{ij} + e_{ij}$. Again, we add the polygenic effect to the model proposed in Fan et al. (2006). Based on "additive effect model" (3.3), we may get the model matrix X_i and regression coefficient vector ϕ as follows: $\phi = (\gamma^{\mathrm{T}}, \alpha_1, \cdots, \alpha_m)^{\mathrm{T}}$ and $X_i = (Z_{Ai1}, \cdots, Z_{Ait_i})^{\mathrm{T}}$, where $Z_{Aij} = (w_{ij}, x_{Aij}^{(1)}, \cdots, x_{Aij}^{(m)})^{\mathrm{T}}, j = 1, 2, \cdots, t_i$.

Property of Model Coefficients. For $g = 1, 2, \cdots, m$, let us denote $D_{A_gQ} = P(Q_1A_g) - q_1P_{A_g}$, which are measures of LD between QTL Q and marker A. Here, $P(Q_1A_g)$ is the frequency of haplotype Q_1A_g. In Appendix B.1, the regression coefficients of "genotype effect model" (3.1) are calculated as

$$\beta_{gh} = \mu + \alpha_Q[D_{A_gQ}/P_{A_g} + D_{A_hQ}/P_{A_h}] - \delta_Q D_{A_gQ} D_{A_hQ}/[P_{A_g}P_{A_h}]. \tag{3.5}$$

In Appendix B.2, we will show that the regression coefficients of "additive effect model" (3.3) are given by

$$\alpha_g = \mu/2 + \alpha_Q D_{A_gQ}/P_{A_g}. \tag{3.6}$$

Notice that relations (3.5) and (3.6) are exactly the same as those of Fan et al. (2006). Assume that the additive effect is significantly present, but the dominance effect is not significantly present, i.e., $\alpha_Q \neq 0$ but $\delta_Q = 0$. To test association between the marker A and the QTL Q, one may test hypotheses $H_{a0} : \alpha_1 = \cdots = \alpha_m$ vs. H_{a1}: at least two α_g's are not equal. On the other hand, assume that both additive and dominance effects are significantly present at the putative QTL Q, i.e., $\alpha_Q \neq 0$ and $\delta_Q \neq 0$. To test association between the marker A and the QTL Q, one may test

hypotheses $H_{ad0}: \beta_{11} = \cdots = \beta_{mm} = \beta_{12} = \cdots = \beta_{1m} = \cdots = \beta_{m-1,m}$ vs. H_{ad1}: at least two β_{gh} are not equal.

***F*-tests and Noncentrality Parameter Approximations**. Assume that there are no covariates. Let us denote $X = (X_1^{\mathrm{T}}, \cdots, X_N^{\mathrm{T}}, X_{N+1}^{\mathrm{T}}, \cdots, X_{N+I}^{\mathrm{T}})^{\mathrm{T}}$, $Y = (y_{11}, \cdots, y_{N1}, \mathbf{y}_{N+1}^{\mathrm{T}}, \cdots, \mathbf{y}_{N+I}^{\mathrm{T}})^{\mathrm{T}}$, $H = (H_{11}, \cdots, H_{N1}, \mathbf{H}_{N+1}^{\mathrm{T}}, \cdots, \mathbf{H}_{N+I}^{\mathrm{T}})^{\mathrm{T}}$, and $e = (e_{11}, \cdots, e_{N1}, \mathbf{e}_{N+1}^{\mathrm{T}}, \cdots, \mathbf{e}_{N+I}^{\mathrm{T}})^{\mathrm{T}}$. Here, $\mathbf{H}_i = (H_{i1}, \cdots, H_{it_i})^{\mathrm{T}}$ and $\mathbf{e}_i = (e_{i1}, \cdots, e_{it_i})^{\mathrm{T}}$, $i = N+1, \cdots, N+I$. Then "genotype effect model" (3.1) or "additive effect model" (3.3) can be expressed as $Y = X\phi + H + e$. Let $\hat{\Sigma}_i$ and $\hat{\phi}$ be the maximum likelihood estimations of Σ_i and ϕ. By standard regression theory, the coefficients can be estimated by $\hat{\phi} = [\sum_{i=1}^{N+I} X_i^{\mathrm{T}} \hat{\Sigma}_i^{-1} X_i]^{-1} \sum_{i=1}^{N+I} X_i^{\mathrm{T}} \hat{\Sigma}_i^{-1} \mathbf{y}_i$.

For "genotype effect model" (3.1), denote regression coefficient vector $\eta = (\beta_{11}, \cdots, \beta_{mm}, \beta_{12}, \cdots, \beta_{1m}, \cdots, \beta_{m-1,m})^{\mathrm{T}}$. Let us define a $(J_A - 1) \times J_A$ matrix by

$$T = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 \end{bmatrix}_{(J_A-1)\times J_A}.$$

Then, $(T\eta)^{\mathrm{T}} = (\beta_{11} - \beta_{22}, \cdots, \beta_{11} - \beta_{mm}, \beta_{11} - \beta_{12}, \cdots, \beta_{11} - \beta_{1m}, \cdots, \beta_{11} - \beta_{m-1,m})$. Hence, the hypothesis H_{ad0} is equivalent to $T\eta = (0, \cdots, 0)^{\mathrm{T}}$. By Graybill (1976), Chapter VI, the test statistic of a hypothesis H_{ad0} is noncentral $F(J_A - 1, \sum_{i=1}^{N+I} t_i - J_A)$ defined by

$$F_{m,ad} = \frac{(T\hat{\eta})^{\mathrm{T}}[T(X^{\mathrm{T}}\hat{\Sigma}^{-1}X)^{-1}T^{\mathrm{T}}]^{-1}(T\hat{\eta})}{Y^{\mathrm{T}}[\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X(X^{\mathrm{T}}\hat{\Sigma}^{-1}X)^{-1}X^{\mathrm{T}}\hat{\Sigma}^{-1}]Y} \frac{\sum_{i=1}^{N+I} t_i - J_A}{J_A - 1},$$

where $\Sigma = \mathrm{diag}(\Sigma_1, \cdots, \Sigma_{N+I})$ is the overall variance-covariance matrix with matrices Σ_i on the diagonal, and $\hat{\Sigma}$ is its maximum likelihood estimation. The noncentrality

parameter of above F-statistic is

$$\lambda_{m,ad} = (T\eta)^{\mathrm{T}}[T(X^{\mathrm{T}}\Sigma^{-1}X)^{-1}T^{\mathrm{T}}]^{-1}(T\eta) = (T\eta)^{\mathrm{T}}[T(\sum_{i=1}^{N+I} X_i^{\mathrm{T}}\Sigma_i^{-1}X_i)^{-1}T^{\mathrm{T}}]^{-1}(T\eta).$$

Assume that the dataset is a population sample, i.e., $I = 0$. Under the assumption of large sample size N, we show in Appendix B.3 the following approximation

$$\lambda_{m,ad} \approx \frac{N(1-\varepsilon_A)}{\sigma^2}\Big[\sigma_{ga}^2 R_{AQ}^2 + \sigma_{gd}^2 R_{AQ}^4\Big], \tag{3.7}$$

where R_{AQ}^2 is a general measure of the degree of linkage disequilibrium between marker A and the QTL Q defined by $R_{AQ}^2 = \sum_{g=1}^{m}\sum_{s=1}^{2}[P(Q_sA_g) - P_{A_g}q_s]^2/[P_{A_g}q_s]$ (Hedrick, 1987; Sham et al., 2000). Notice that R_{AQ}^2 is the χ^2 statistic of the $m \times 2$ table of haplotype frequencies of the marker A and trait locus Q. Approximation (3.7) shows that the noncentrality parameter $\lambda_{m,ad}$ is reduced by a factor of $1-\varepsilon_A$. If there is no missing genotype data, i.e., $\varepsilon_A = 0$, approximation (3.7) is exactly the same as that of the "genotype effect model" in Fan et al. (2006). In the presence of missing data, the model developed extends our previous work. In addition, $\lambda_{m,ad}$ is reduced by a factor of R_{AQ}^2 for additive variance σ_{ga}^2, and a factor of R_{AQ}^4 for dominance variance σ_{gd}^2.

For "additive effect model" (3.3), denote $\psi = (\alpha_1, \cdots, \alpha_m)^{\mathrm{T}}$. Let K be a $(m-1)\times m$ matrix defined by

$$K = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 \end{bmatrix}_{(m-1)\times m}.$$

Then, $(K\psi)^{\mathrm{T}} = (\alpha_1 - \alpha_2, \cdots, \alpha_1 - \alpha_m)$. Hence, the hypothesis H_{a0} is equivalent to $K\psi = (0, \cdots, 0)^{\mathrm{T}}$. By Graybill (1976), Chapter VI, the test statistic of the hypothesis H_{a0} is noncentral $F(m-1, \sum_{i=1}^{N+I} t_i - m)$ defined by

$$F_{m,a} = \frac{(K\hat{\psi})^{\mathrm{T}}[K(X^{\mathrm{T}}\hat{\Sigma}^{-1}X)^{-1}K^{\mathrm{T}}]^{-1}(K\hat{\psi})}{Y^{\mathrm{T}}[\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X(X^{\mathrm{T}}\hat{\Sigma}^{-1}X)^{-1}X^{\mathrm{T}}\hat{\Sigma}^{-1}]Y} \frac{\sum_{i=1}^{N+I} t_i - m}{m-1}.$$

Here, the model matrix X is built from the "additive effect model" (3.3). The noncentrality parameter of above F-statistic is $\lambda_{m,a} = (K\psi)^{\mathrm{T}}[K(X^{\mathrm{T}}\Sigma^{-1}X)^{-1}K^{\mathrm{T}}]^{-1}(K\psi)$.

Assume that the dataset is a population sample, i.e., $I = 0$. Under an assumption of large sample size N, we show in Appendix B.4 the following approximation

$$\lambda_{m,a} = \frac{1}{\sigma^2}(K\psi)^{\mathrm{T}}\Big[K(X^{\mathrm{T}}X)^{-1}K^{\mathrm{T}}\Big]^{-1}(K\psi) \approx \frac{N(1-\varepsilon_A)\sigma^2_{ga}}{\sigma^2}R^2_{AQ}. \tag{3.8}$$

Again, the approximation (3.8) shows that the noncentrality parameter $\lambda_{m,a}$ is reduced by a factor of $1 - \varepsilon_A$. If there is no missing genotype data, i.e., $\varepsilon_A = 0$, approximation (3.8) is exactly the same as that of the "additive effect model" in Fan et al. (2006). Besides, $\lambda_{m,a}$ is reduced by a factor of R^2_{AQ} for additive variance. The dominance variance is not present in $\lambda_{m,a}$.

3.2.3 Mixed Effect Models by Two Markers

In addition to marker A, assume that a second marker B is typed, which has n alleles denoted by $B_1, \cdots, B_n$. Suppose that the marker B is also in HWE. Let the frequency of allele B_k be $P_{B_k}, k = 1, 2, \cdots, n$. There are $J_B = n(n+1)/2$ possible genotypes, which can be listed as $B_1B_1, \cdots, B_nB_n, B_1B_2, \cdots, B_1B_n, \cdots, B_{n-1}B_n$. Let y_{ij} be the trait value of the j-th pedigree member of the i-th family with genotype G_{Aij} at marker A and genotype G_{Bij} at marker B. Such as G_{Aij} discussed above, G_{Bij} can be missing. If G_{Bij} is missing, we denote it as $G_{Bij} =?$; and if G_{Bij} is not missing, we denote it by $G_{Bij} \neq ?$. Let us denote the probability that the genotype G_{Bij} is missing by ε_B, i.e., $P(G_{Bij} =?) = \varepsilon_B$. Notice that $P(G_{Bij} \neq ?) = 1 - \varepsilon_B$. In addition, let

$P(G_{Bij} = B_kB_l|G_{Bij} =?)$ or $P(G_{Bij} = B_kB_l|G_{Bij} \neq?)$ be the conditional probability of genotype B_kB_l given $G_{Bij} =?$ or $G_{Bij} \neq?$. Since the missing mechanism is MCAR, the probability

$$\begin{aligned} P(G_{Bij} = B_kB_l|G_{Bij} =?) &= P(G_{Bij} = B_kB_l|G_{Bij} \neq?) \\ &= P(B_kB_l) = \begin{cases} P_{B_k}^2 & \text{if } k = l \\ 2P_{B_k}P_{B_l} & \text{if } k \neq l \end{cases}. \end{aligned}$$

Such as relations (3.4) to define $x_{Aij}^{(g)}$, let us denote the expected number of alleles B_k given the observed genotype G_{Bij} at marker B

$$x_{Bij}^{(k)} = 2 \cdot 1_{(G_{Bij}=B_kB_k)} + \sum_{l\neq k} 1_{(G_{Bij}=B_kB_l)} + 2P_{B_k}1_{(G_{Bij}=?)}. \tag{3.9}$$

The "additive effect model" (13) of Fan et al. (2006) can be extended to

$$y_{ij} = w_{ij}\gamma + \alpha + \sum_{g=1}^{m-1} x_{Aij}^{(g)}\alpha_{Ag} + \sum_{k=1}^{n-1} x_{Bij}^{(k)}\alpha_{Bk} + H_{ij} + e_{ij}, \tag{3.10}$$

where w_{ij} and γ are the same as those in model (3.1), and α, α_{Ag}, and α_{Bk} are regression coefficients. To understand that model (3.10) extends model (13) of Fan et al. (2006), consider the four possible cases as follows.

Case 1: both genotype G_{Aij} and G_{Bij} are not missing, model (3.10) is similar to (13) of Fan et al. (2006). In model (3.10), we model the polygenic effect, which is not modeled in Fan et al. (2006).

Case 2: both genotypes $G_{Aij} =?$ and $G_{Bij} =?$ are missing, model (3.10) becomes

$$y_{ij} = w_{ij}\gamma + \alpha + 2\sum_{g=1}^{m-1} P_{A_g}\alpha_{Ag} + 2\sum_{k=1}^{n-1} P_{B_k}\alpha_{Bk} + H_{ij} + e_{ij}.$$

Case 3: genotype G_{Aij} is not missing and genotype G_{Bij} is missing, (3.10) be-

comes

$$\begin{aligned} y_{ij} &= w_{ij}\gamma + \alpha + \sum_{g=1}^{m-1}\Big[2\cdot 1_{(G_{Aij}=A_gA_g)} + \sum_{h\neq g} 1_{(G_{Aij}=A_gA_h)}\Big]\alpha_{Ag} \\ &\quad +2\sum_{k=1}^{n-1} P_{B_k}\alpha_{Bk} + H_{ij} + e_{ij}. \end{aligned}$$

Case 4: genotype G_{Aij} is missing and genotype G_{Bij} is not missing, (3.10) becomes

$$\begin{aligned} y_{ij} &= w_{ij}\gamma + \alpha + 2\sum_{g=1}^{m-1} P_{A_g}\alpha_{Ag} \\ &\quad +\sum_{k=1}^{n-1}\Big[2\cdot 1_{(G_{Bij}=B_kB_k)} + \sum_{l\neq k} 1_{(G_{Bij}=B_kB_l)}\Big]\alpha_{Bk} + H_{ij} + e_{ij}. \end{aligned}$$

To extend the "genotype effect model" (14) of Fan et al. (2006), let us denote

$$\begin{aligned} z_{Aij}^{(gh)} &= -P_{A_h}^2 1_{(G_{Aij}=A_gA_g)} + P_{A_g}P_{A_h}1_{(G_{Aij}=A_gA_h)} - P_{A_g}^2 1_{(G_{Aij}=A_hA_h)}, \\ z_{Bij}^{(kl)} &= -P_{B_l}^2 1_{(G_{Bij}=B_kB_k)} + P_{B_k}P_{B_l}1_{(G_{Bij}=B_kB_l)} - P_{B_k}^2 1_{(G_{Bij}=B_lB_l)}. \end{aligned} \tag{3.11}$$

If the genotypes G_{Aij} and G_{Bij} are not missing, the variables $x_{Aij}^{(g)}, x_{Bij}^{(k)}, z_{Aij}^{(gh)}$ and $z_{Bij}^{(kl)}$ are the same as those defined in Fan et al. (2006). If the genotype G_{Aij} or G_{Bij} is missing, $x_{Aij}^{(g)}$ or $x_{Bij}^{(k)}$ is simply the expected number $2P_{A_g}$ or $2P_{B_k}$ of alleles A_g or B_k, and $z_{Aij}^{(gh)}$ or $z_{Bij}^{(kl)}$ is 0. The reason that $z_{Aij}^{(gh)}$ is 0 on G_{Aij} =? is as follows: $-P_{A_h}^2 P(G_{Aij} = A_gA_g|G_{Aij} =?) + P_{A_g}P_{A_h}P(G_{Aij} = A_gA_h|G_{Aij} =?) - P_{A_g}^2 P(G_{Aij} = A_hA_h|G_{Aij} =?) = 0$; the same reasoning applies to $z_{Bij}^{(kl)}$. The "genotype effect model" (14) of Fan et al. (2006) can be extended to

$$\begin{aligned} y_{ij} &= w_{ij}\gamma + \alpha + \sum_{g=1}^{m-1} x_{Aij}^{(g)}\alpha_{Ag} + \sum_{k=1}^{n-1} x_{Bij}^{(k)}\alpha_{Bk} \\ &\quad + \sum_{1\le g<h\le m} z_{Aij}^{(gh)}\delta_{Agh} + \sum_{1\le k<l\le n} z_{Bij}^{(kl)}\delta_{Bkl} + H_{ij} + e_{ij}, \end{aligned} \tag{3.12}$$

where δ_{Agh} and δ_{Bkl} are regression coefficients of variables $z_{Aij}^{(gh)}$ and $z_{Bij}^{(kl)}$, respectively; other terms are the same as those of model (3.10).

In the following, we are going to show that model (3.12) extends model (14) of Fan et al. (2006). In total, there are four cases as follows.

Case 1: both genotype G_{Aij} and G_{Bij} are not missing, model (3.12) is similar to (14) of Fan et al. (2006). Here, we add the polygenic additive effect to the model.

Case 2: both genotypes $G_{Aij} =?$ and $G_{Bij} =?$ are missing, model (3.12) becomes

$$y_{ij} = w_{ij}\gamma + \alpha + 2\sum_{g=1}^{m-1} P_{A_g}\alpha_{Ag} + 2\sum_{k=1}^{n-1} P_{B_k}\alpha_{Bk} + H_{ij} + e_{ij}.$$

Case 3: genotype G_{Aij} is not missing and genotype G_{Bij} is missing, (3.12) becomes

$$\begin{aligned} y_{ij} &= w_{ij}\gamma + \alpha + \sum_{g=1}^{m-1}\Big[2\cdot 1_{(G_{Aij}=A_gA_g)} + \sum_{h\neq g} 1_{(G_{Aij}=A_gA_h)}\Big]\alpha_{Ag} \\ &\quad +2\sum_{k=1}^{n-1} P_{B_k}\alpha_{Bk} + \sum_{1\le g<h\le m} z_{Aij}^{(gh)}\delta_{Agh} + H_{ij} + e_{ij}. \end{aligned}$$

Case 4: genotype G_{Aij} is missing and genotype G_{Bij} is not missing, (3.12) becomes

$$\begin{aligned} y_{ij} &= w_{ij}\gamma + \alpha + 2\sum_{g=1}^{m-1} P_{A_g}\alpha_{Ag} \\ &\quad + \sum_{k=1}^{n-1}\Big[2\cdot 1_{(G_{Bij}=B_kB_k)} + \sum_{l\neq k} 1_{(G_{Bij}=B_kB_l)}\Big]\alpha_{Bk} + \sum_{1\le k<l\le n} z_{Bij}^{(kl)}\delta_{Bkl} + H_{ij} + e_{ij}. \end{aligned}$$

Denote $X_{Aij} = (x_{Aij}^{(1)}, \cdots, x_{Aij}^{(m-1)})^{\mathrm{T}}$, $X_{Bij} = (x_{Bij}^{(1)}, \cdots, x_{Bij}^{(n-1)})^{\mathrm{T}}$, and $X_{A\cup B}^{(ij)} = (X_{Aij}^{\mathrm{T}}, X_{Bij}^{\mathrm{T}})^{\tau}$. Let us denote the additive variance-covariance matrix of the indicator variables $x_{Aij}^{(g)}, x_{Bij}^{(k)}$ by

$$V_A = \mathrm{cov}(X_{A\cup B}^{(ij)}, X_{A\cup B}^{(ij)}) = E(X_{A\cup B}^{(ij)}(X_{A\cup B}^{(ij)})^{\tau}) - EX_{A\cup B}^{(ij)}E(X_{A\cup B}^{(ij)})^{\mathrm{T}}.$$

Similarly, let $Z_{Aij} = (z_{Aij}^{(12)}, \cdots, z_{Aij}^{(1m)}, z_{Aij}^{(23)}, \cdots, z_{Aij}^{(2m)}, \cdots, z_{Aij}^{(m-1,m)})^{\mathrm{T}}$, $Z_{Bij} = (z_{Bij}^{(12)}, \cdots, z_{Bij}^{(1n)}, z_{Bij}^{(23)}, \cdots, z_{Bij}^{(2n)}, \cdots, z_{Bij}^{(n-1,n)})^{\mathrm{T}}$, and $Z_{A\cup B}^{(ij)} = (Z_{Aij}^{\mathrm{T}}, Z_{Bij}^{\mathrm{T}})^{\tau}$. Let us denote the

dominance variance-covariance matrix of the indicator variables $z_{Aij}^{(gh)}, z_{Bij}^{(kl)}$ by $V_D = \mathrm{cov}(Z_{A\cup B}^{(ij)}, Z_{A\cup B}^{(ij)})$. The elements of matrices V_A and V_D are provided in Appendix B.5.

For $k = 1, 2, \cdots, n$, let us denote $D_{B_kQ} = P(Q_1B_k) - q_1P_{B_k}$, which are measures of LD between QTL Q and marker B. Here, $P(Q_1B_k)$ is the frequency of haplotype Q_1B_k. In Appendix B.5, we show that the regression coefficients of models (3.10) and (3.12) are

$$\begin{bmatrix} \alpha_{A1} \\ \vdots \\ \alpha_{A(m-1)} \\ \alpha_{B1} \\ \vdots \\ \alpha_{B(n-1)} \end{bmatrix} = (V_A/2)^{-1} \begin{bmatrix} D_{A_1Q}(1-\varepsilon_A) \\ \vdots \\ D_{A_{m-1}Q}(1-\varepsilon_A) \\ D_{B_1Q}(1-\varepsilon_B) \\ \vdots \\ D_{B_{n-1}Q}(1-\varepsilon_B) \end{bmatrix} \alpha_Q;$$

$$\begin{bmatrix} \delta_{A12} \\ \vdots \\ \delta_{A(m-1)m} \\ \delta_{B12} \\ \vdots \\ \delta_{B(n-1)n} \end{bmatrix} = V_D^{-1} \begin{bmatrix} [P_{A_2}D_{A_1Q} - P_{A_1}D_{A_2Q}]^2(1-\varepsilon_A) \\ \vdots \\ [P_{A_{m-1}}D_{A_mQ} - P_{A_m}D_{A_{m-1}Q}]^2(1-\varepsilon_A) \\ [P_{B_2}D_{B_1Q} - P_{B_1}D_{B_2Q}]^2(1-\varepsilon_B) \\ \vdots \\ [P_{B_{n-1}}D_{B_nQ} - P_{B_n}D_{B_{n-1}Q}]^2(1-\varepsilon_B) \end{bmatrix} \delta_Q. \quad (3.13)$$

Equations (3.13) show that the parameters of LD (i.e., D_{A_gQ} and D_{B_kQ}) and gene effect (i.e., α_Q and δ_Q) are contained in the regression coefficients. Models (3.10) and (3.12) simultaneously take care of the LD and the effects of the putative trait locus Q. The gene substitution effect α_Q is contained only in α_{Ag}, α_{Bk}; and the dominance effect δ_Q is contained only in $\delta_{Agh}, \delta_{Bkl}$. Therefore, V_A is called an additive variance-covariance matrix; and V_D is called a dominance variance-covariance matrix. The model (3.12) orthogonally decomposes genetic effect into summation of additive and

dominance effects.

Based on equations (3.13), we may use models (3.10) and (3.12) to test the association between the trait locus Q and the two markers A and B. Assume that the additive genetic effect is significantly present, but the dominance genetic effect is not significantly present, i.e., $\alpha_Q \neq 0$ but $\delta_Q = 0$. To test association between the markers A & B and the QTL Q, one may test hypotheses $H_{ABa0} : \alpha_{A1} = \cdots = \alpha_{A(m-1)} = \alpha_{B1} = \cdots = \alpha_{B(n-1)} = 0$ vs. H_{ABa1}: at least one α_{Ag}, α_{Bk}'s is not equal to 0. On the other hand, assume that both additive and dominance genetic effects are significantly present at the putative QTL Q, i.e., $\alpha_Q \neq 0$ and $\delta_Q \neq 0$. To test association between the markers A & B and the QTL Q, one may test hypotheses $H_{ABad0} : \alpha_{A1} = \cdots = \alpha_{A(m-1)} = \alpha_{B1} = \cdots = \alpha_{B(n-1)} = \delta_{A12} = \cdots = \delta_{A1m} = \cdots = \delta_{A(m-1)m} = \delta_{B12} = \cdots = \delta_{B1n} = \cdots = \delta_{B(n-1)n} = 0$ vs. H_{ABad1}: at least one $\alpha_{Ag}, \alpha_{Bk}, \delta_{Agh}, \delta_{Bkl}$ is not equal to 0.

Regression Models and F-tests. Based on regression (3.12), one may construct an F-test statistic $F_{AB,ad}$ to test the null hypothesis H_{ABad0} in the same way to construct $F_{m,ad}$ or $F_{m,a}$ (Graybill, 1976, Chapter VI). Under the null hypothesis of H_{ABad0}, $F_{AB,ad}$ is central $F(J_A + J_B - 2, \sum_{i=1}^{N+I} t_i - J_A - J_B + 1)$. Similarly, one may construct an F-test statistic $F_{AB,a}$ to test the null hypothesis H_{ABa0} based on the "additive effect model" (3.10). Under the null hypothesis of H_{ABa0}, $F_{AB,a}$ is central $F(m + n - 2, \sum_{i=1}^{N+I} t_i - m - n + 1)$.

Population Sample and Noncentrality Parameter Approximations. Assume that there are no covariates, and the dataset is a population sample, i.e., $I = 0$. Suppose the sample size N is large enough that the large sample theory applies. Denote $D_{AQ} = (D_{A_1Q}, \cdots, D_{A_{m-1}Q})^{\mathrm{T}}$ and $D_{BQ} = (D_{B_1Q}, \cdots, D_{B_{n-1}Q})^{\mathrm{T}}$; $\Delta_{AQ} = \left([P_{A_2}D_{A_1Q} - P_{A_1}D_{A_2Q}]^2, \cdots, [P_{A_{m-1}}D_{A_mQ} - P_{A_m}D_{A_{m-1}Q}]^2\right)^{\mathrm{T}}$ and $\Delta_{BQ} = \left([P_{B_2}D_{B_1Q} - P_{B_1}D_{B_2Q}]^2, \cdots, [P_{B_{n-1}}D_{B_nQ} - P_{B_n}D_{B_{n-1}Q}]^2\right)^{\mathrm{T}}$. Under the alternative

hypothesis of H_{ABad1}, $F_{AB,ad}$ is noncentral $F(J_A + J_B - 2, N - J_A - J_B + 1)$, and it can be shown that the corresponding noncentrality parameter is approximated by

$$\lambda_{ABad} \approx \frac{N}{\sigma^2}\Bigg[\Big(D_{AQ}^{\mathrm{T}}(1-\varepsilon_A), D_{BQ}^{\mathrm{T}}(1-\varepsilon_B)\Big)(V_A/2)^{-1}\begin{bmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B)\end{bmatrix}\sigma_{ga}^2/(q_1q_2) + \Big(\Delta_{AQ}^{\mathrm{T}}(1-\varepsilon_A), \Delta_{BQ}^{\mathrm{T}}(1-\varepsilon_B)\Big)V_D^{-1}\begin{bmatrix} \Delta_{AQ}(1-\varepsilon_A) \\ \Delta_{BQ}(1-\varepsilon_B)\end{bmatrix}\sigma_{gd}^2/(q_1^2q_2^2)\Bigg].$$

Under the null hypothesis of H_{ABa0}, $F_{AB,a}$ is central $F(m+n-2, N-n-m+1)$. Under the alternative hypothesis of H_{ABa1}, $F_{AB,a}$ is noncentral $F(m+n-2, N-m-n+1)$, and it can be shown that the corresponding noncentrality parameter is approximated by

$$\lambda_{ABa} \approx \frac{N}{\sigma^2}\Big(D_{AQ}^{\mathrm{T}}(1-\varepsilon_A), D_{BQ}^{\mathrm{T}}(1-\varepsilon_B)\Big)(V_A/2)^{-1}\begin{bmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B)\end{bmatrix}\sigma_{ga}^2/(q_1q_2).$$

Pedigree Sample and Noncentrality Parameter Approximations. Consider pedigree data, and assume that there are no covariates. For a relative pair $(1, 2)$ of individuals 1 and 2 who are noninbred relatives, Table 3 gives the conditional probability $P(G_1, G_2|C)$ given their allele IBD sharing status. Here, G_j is genotype of individual j, and C is one event of $(IBD = k), k = 0, 1, 2$. For example, $P(A_gA_g, A_gA_g|IBD = 0) = P_{A_g}^4$, $P(A_gA_g, A_gA_h|IBD = 0) = 2P_{A_g}^3P_{A_h}$ and $P(A_gA_g, A_hA_h|IBD = 0) = P_{A_g}^2P_{A_h}^2$. Utilizing the conditional probabilities of Table 3, the conditional covariances of variables $x_{Aij}^{(g)}, x_{Bij}^{(k)}, z_{Aij}^{(gh)}$ and $z_{Bij}^{(kl)}$ of a relative pair $(1, 2)$ can be calculated and the results are listed in Table 4. Given (IBD=0), the covariances are 0 since the two variables are independent and so unrelated (for instance, $\mathrm{cov}(x_{Ai1}^{(g)}, x_{Ai2}^{(g)}|IBD = 0) = 0$). Other entries of Table 4 can be calculated, accordingly. Based on Table 4, it can be seen that $\mathrm{cov}(X_{A\cup B}^{(i1)}, X_{A\cup B}^{(i2)}|IBD = 0) = \mathrm{cov}(Z_{A\cup B}^{(i1)}, Z_{A\cup B}^{(i2)}|IBD = 0) = 0$ and $\mathrm{cov}(X_{A\cup B}^{(i1)}, Z_{A\cup B}^{(i2)}|IBD = k) = 0$, for $k = 0, 1, 2$.

Table 3: Conditional probability $P(G_1, G_2|C)$ of a relative pair $(1, 2)$ given their allele IBD sharing status. Here, G_j is genotype of individual j, and C is one event of $(IBD = k), k = 0, 1, 2$. In the table, we assume $g \neq h, g \neq g', g \neq h', h \neq g', h \neq h', g' \neq h', k \neq l$.

Conditional Probability	**allele IBD sharing status** C		
	IBD=0	**IBD=1**	**IBD=2**
$P(A_gA_g, A_gA_g\|C)$	$P_{A_g}^4$	$P_{A_g}^3$	$P_{A_g}^2$
$P(A_gA_g, A_gA_h\|C)$	$2P_{A_h}P_{A_g}^3$	$P_{A_h}P_{A_g}^2$	0
$P(A_gA_g, A_hA_h\|C)$	$P_{A_g}^2P_{A_h}^2$	0	0
$P(A_gA_g, A_hA_{h'}\|C)$	$2P_{A_g}^2P_{A_h}P_{A_{h'}}$	0	0
$P(A_gA_h, A_gA_h\|C)$	$4P_{A_g}^2P_{A_h}^2$	$P_{A_g}P_{A_h}^2 + P_{A_g}^2P_{A_h}$	$2P_{A_g}P_{A_h}$
$P(A_gA_h, A_gA_{h'}\|C)$	$4P_{A_g}^2P_{A_h}P_{A_{h'}}$	$P_{A_g}P_{A_h}P_{A_{h'}}$	0
$P(A_gA_h, A_{g'}A_{h'}\|C)$	$4P_{A_g}P_{A_h}P_{A_{g'}}P_{A_{h'}}$	0	0
$P(A_gA_g, B_kB_k\|C)$	$P_{A_g}^2P_{B_k}^2$	$P_{A_g}P_{B_k}P(A_gB_k)$	$P(A_gB_k)^2$
$P(A_gA_g, B_kB_l\|C)$	$2P_{A_g}^2P_{B_k}P_{B_l}$	$P_{A_g}P_{B_l}P(A_gB_k)$ $+P_{A_g}P_{B_k}P(A_gB_l)$	$2P(A_gB_k)P(A_gB_l)$
$P(A_gA_h, B_kB_k\|C)$	$2P_{A_g}P_{A_h}P_{B_k}^2$	$P_{A_g}P_{B_k}P(A_hB_k)$ $+P_{A_h}P_{B_k}P(A_gB_k)$	$2P(A_gB_k)P(A_hB_k)$
$P(A_gA_h, B_kB_l\|C)$	$4P_{A_g}P_{A_h}P_{B_k}P_{B_l}$	$P_{A_g}P_{B_k}P(A_hB_l)$ $+P_{A_g}P_{B_l}P(A_hB_k)$ $+P_{A_h}P_{B_k}P(A_gB_l)$ $+P_{A_h}P_{B_l}P(A_gB_k)$	$2P(A_gB_k)P(A_hB_l)$ $+2P(A_gB_l)P(A_hB_k)$

In addition, we have

$$\begin{aligned}&\mathrm{cov}(X_{A\cup B}^{(i1)}, X_{A\cup B}^{(i2)}|IBD = 1)\\ &= \frac{1}{2}\mathrm{cov}(X_{A\cup B}^{(i1)}, X_{A\cup B}^{(i2)}|IBD = 2), \mathrm{cov}(Z_{A\cup B}^{(i1)}, Z_{A\cup B}^{(i2)}|IBD = 1) = 0.\end{aligned}$$

Let Φ_{12} be their kinship coefficient of individuals 1 and 2, and Δ_{712} be the probability that both alleles shared by the two individuals 1 and 2 are IBD at any locus (Lange, 2002). Then it can be shown that the covariance matrix of variable vectors $X_{A\cup B}^{(i1)}$ and $Z_{A\cup B}^{(i2)}$ is a zero matrix, and

$$\begin{aligned}\mathrm{cov}(X_{A\cup B}^{(i1)}, X_{A\cup B}^{(i2)}) &= 2\Phi_{12}\mathrm{cov}(X_{A\cup B}^{(i1)}, X_{A\cup B}^{(i2)}|IBD = 2) = 2\Phi_{12}V_{A2},\\ \mathrm{cov}(Z_{A\cup B}^{(i1)}, Z_{A\cup B}^{(i2)}) &= \Delta_{712}\mathrm{cov}(Z_{A\cup B}^{(i1)}, Z_{A\cup B}^{(i2)}|IBD = 2) = \Delta_{712}V_{D2}, \qquad (3.14)\end{aligned}$$

Table 4: Conditional expectation of a relative pair $(1,2)$ given their allele IBD sharing status. In the table, we assume $g \neq h, g \neq g', g \neq h', h \neq g', h \neq h', g' \neq h', k \neq l$.

Conditional Expectation	**allele IBD sharing status C (IBD)**		
	0	1	2
$\mathrm{cov}(x_{Ai1}^{(g)}, x_{Ai2}^{(g)}\|C)$	0	$P_{A_g}[1-P_{A_g}](1-\varepsilon_A)^2$	$2P_{A_g}[1-P_{A_g}](1-\varepsilon_A)^2$
$\mathrm{cov}(x_{Ai1}^{(g)}, x_{Ai2}^{(h)}\|C)$	0	$-P_{A_g}P_{A_h}(1-\varepsilon_A)^2$	$-2P_{A_g}P_{A_h}(1-\varepsilon_A)^2$
$\mathrm{cov}(z_{Ai1}^{(gh)}, z_{Ai2}^{(gh)}\|C)$	0	0	$P_{A_g}^2P_{A_h}^2(P_{A_g}+P_{A_h})^2(1-\varepsilon_A)^2$
$\mathrm{cov}(z_{Ai1}^{(gh)}, z_{Ai2}^{(gh')}\|C)$	0	0	$[P_{A_g}P_{A_h}P_{A_{h'}}(1-\varepsilon_A)]^2$
$\mathrm{cov}(z_{Ai1}^{(gh)}, z_{Ai2}^{(g'h')}\|C)$	0	0	0
$\mathrm{cov}(x_{Ai1}^{(g)}, z_{Ai2}^{(gh)}\|C)$	0	0	0
$\mathrm{cov}(x_{Ai1}^{(g)}, z_{Ai2}^{(g'h')}\|C)$	0	0	0
$\mathrm{cov}(x_{Ai1}^{(g)}, x_{Bi2}^{(k)}\|C)$	0	$(1-\varepsilon_A)(1-\varepsilon_B)D_{A_gB_k}$	$2(1-\varepsilon_A)(1-\varepsilon_B)D_{A_gB_k}$
$\mathrm{cov}(x_{Ai1}^{(g)}, z_{Bi2}^{(kl)}\|C)$	0	0	0
$\mathrm{cov}(z_{Ai1}^{(gh)}, z_{Bi2}^{(kl)}\|C)$	0	0	$E[z_{Ai1}^{(gh)}z_{Bi1}^{(kl)}] = E[z_{Ai2}^{(gh)}z_{Bi2}^{(kl)}]$

where the elements of $V_{A2} = \mathrm{cov}(X_{A\cup B}^{(i1)}, X_{A\cup B}^{(i2)}|IBD=2)$ and $V_{D2} = \mathrm{cov}(Z_{A\cup B}^{(i1)}, Z_{A\cup B}^{(i2)}|$ $IBD=2)$ are given by the entries of the last column of Table 4.

Nuclear Family Data. Consider I families each has both parents and s offspring. The total number of individuals is $I(s+2)$. Let us list the $s+1$ individuals of each family as $j = 1, 2, 3, \cdots, s+1$, where individual 1 is the father and individual 2 is the mother, and the offspring are listed as $j = 3, \cdots, s+2$. Suppose that variance-covariance matrices of the I families are the same, i.e., $\Sigma_1 = \cdots = \Sigma_I$. Denote $\Sigma_i^{-1} = \frac{1}{\sigma^2}(\gamma_{hj})_{(s+2)\times(s+2)}$, and let $b = (\gamma_{13} + \cdots + \gamma_{1,s+2}) + (\gamma_{23} + \cdots + \gamma_{2,s+2}) + \sum_{h=3}^{s+2}\sum_{j=h+1}^{s+2}\gamma_{hj}$. If the number of families I is large enough, we show in Appendix

B.6 that the noncentrality parameter of statistic $F_{AB,ad}$ is approximated by

$$\begin{aligned}\lambda_{ABad} \approx\ & \frac{2I\sigma_{ga}^2}{q_1q_2\sigma^2}\Big(D_{AQ}^{\mathrm{T}}(1-\varepsilon_A), D_{BQ}^{\mathrm{T}}(1-\varepsilon_B)\Big)V_A^{-1} \\ & \times\Big(\sum_{k=1}^{s+2}\gamma_{kk}V_A + bV_{A2}\Big)V_A^{-1}\begin{bmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B)\end{bmatrix} \\ & + \frac{I\sigma_{gd}^2}{(q_1q_2)^2\sigma^2}\Big(\Delta_{AQ}^{\mathrm{T}}(1-\varepsilon_A), \Delta_{BQ}^{\mathrm{T}}(1-\varepsilon_B)\Big)V_D^{-1} \\ & \times\Big(\sum_{k=1}^{s+2}\gamma_{kk}V_D + \sum_{k=3}^{s+2}\sum_{l=k+1}^{s+2}\gamma_{kl}V_{D2}/2\Big)V_D^{-1}\begin{bmatrix} \Delta_{AQ}(1-\varepsilon_A) \\ \Delta_{BQ}(1-\varepsilon_B)\end{bmatrix}.\end{aligned} \tag{3.15}$$

Similarly, the noncentrality parameter of statistic $F_{AB,a}$ is approximated by

$$\begin{aligned}\lambda_{ABa} \approx\ & \frac{2I\sigma_{ga}^2}{q_1q_2\sigma^2}\Big[D_{AQ}^{\mathrm{T}}(1-\varepsilon_A), D_{BQ}^{\mathrm{T}}(1-\varepsilon_B)\Big]V_A^{-1} \\ & \times\Big(\sum_{k=1}^{s+2}\gamma_{kk}V_A + bV_{A2}\Big)V_A^{-1}\begin{bmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B)\end{bmatrix}.\end{aligned}$$

Multi-generation Pedigree Data. Consider I families given in graph A or graph B of Figure 3 (Figure 1 in Abecasis et al., 2000b; Fan et al., 2005). For each individual in Figure 3, an ID is assigned. For the grand parents of graph B, both phenotypes and genotypes are unavailable and so no IDs are assigned. The total number of individuals is tI, where $t = 11$ for graph A and $t = 18$ for graph B of Figure 3, respectively. Again, assume that variance-covariance matrices of the I families are the same, i.e., $\Sigma_1 = \cdots = \Sigma_I$. Denote $\Sigma_i^{-1} = \frac{1}{\sigma^2}(\gamma_{hj})_{t\times t}$. If the number of families I is large enough, we can show in the same way as Appendix B.6 that the noncentrality

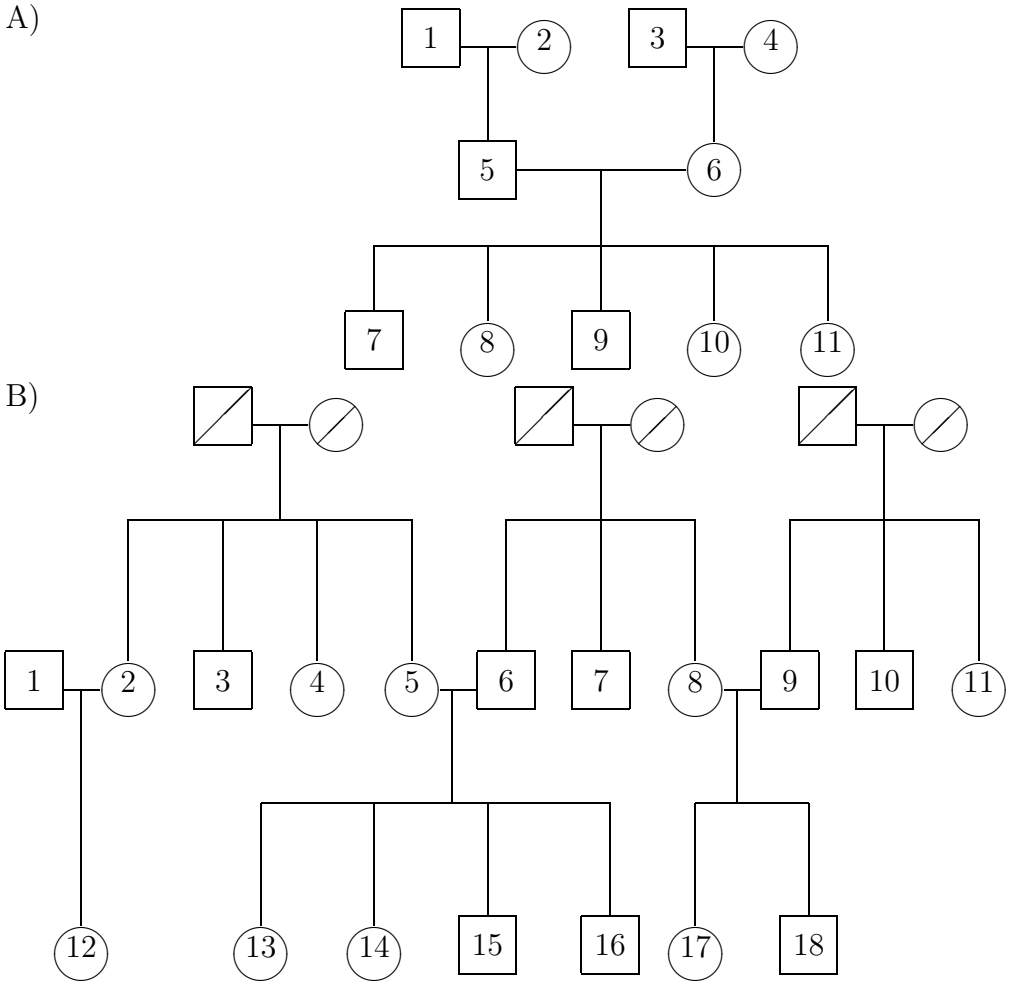

Figure 3: Multi-generation pedigrees used in power calculations and comparison, which are taken from Figure 1 of Abecasis et al. (2000b) or Fan et al. (2005). The number in the box or circle is individual ID.

parameter of statistic $F_{AB,ad}$ is approximated by

$$\begin{aligned}\lambda_{ABad} \approx\ & \frac{2I\sigma^2_{ga}}{q_1q_2\sigma^2}\Big(D^{\mathrm{T}}_{AQ}(1-\varepsilon_A), D^{\mathrm{T}}_{BQ}(1-\varepsilon_B)\Big)V_A^{-1} \\ & \times\Big(\sum_{k=1}^{t}\gamma_{kk}V_A + b_1V_{A2}\Big)V_A^{-1}\begin{bmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B)\end{bmatrix} \\ & + \frac{I\sigma^2_{gd}}{(q_1q_2)^2\sigma^2}\Big(\Delta^{\mathrm{T}}_{AQ}(1-\varepsilon_A), \Delta^{\mathrm{T}}_{BQ}(1-\varepsilon_B)\Big)V_D^{-1} \\ & \times\Big(\sum_{k=1}^{t}\gamma_{kk}V_D + b_2V_{D2}/2\Big)V_D^{-1}\begin{bmatrix} \Delta_{AQ}(1-\varepsilon_A) \\ \Delta_{BQ}(1-\varepsilon_B)\end{bmatrix},\end{aligned} \tag{3.16}$$

where b_1 and b_2 are provided in Appendix B.7. The noncentrality parameter of $F_{AB,a}$ is approximated by

$$\begin{aligned}\lambda_{ABa} &\approx \frac{2I\sigma_{ga}^2}{q_1q_2\sigma^2}\Big(D_{AQ}^{\mathrm{T}}(1-\varepsilon_A), D_{BQ}^{\mathrm{T}}(1-\varepsilon_B)\Big)V_A^{-1} \\ &\quad \times\Big(\sum_{k=1}^{t}\gamma_{kk}V_A + b_1V_{A2}\Big)V_A^{-1}\begin{bmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B)\end{bmatrix}.\end{aligned}$$

3.3 Type I Error Rates and Power Comparison

3.3.1 Type I Error Rates

Simulation studies are performed to evaluate the robustness of the proposed models. We evaluate a marker A which is di-allelic, tri-allelic and quadri-allelic, i.e., $m =$ 2, 3 and 4. For di-allelic marker, equal allele frequencies are assumed, i.e., $P_{A_1} = P_{A_2} = 0.5$; for tri-allelic marker, the allele frequencies are given by $P_{A_1} = P_{A_2} = 0.3$ and $P_{A_3} = 0.4$; and for quadri-allelic marker, equal allele frequencies are assumed, i.e., $P_{A_1} = \cdots = P_{A_4} = 0.25$. Five test cases are considered in type I error rate calculation. Table 5 presents parameters of four test cases. Trait values are constructed by normal distribution with mean 100 and total variance $\sigma^2 = 1$ except for test case of **Admixture**. Here $\sigma^2 = \sigma_{ga}^2 + \sigma_{Ga}^2 + \sigma_e^2$ is the summation of the additive major gene effect σ_{ga}^2 , the variance of polygenic effect σ_{Ga}^2 , and the error variance σ_e^2 . In the test cases of **Null**, **Familiality**, and **Admixture**, no major gene effect is assumed, i.e., $\sigma_{ga}^2 = 0$. In the test cases of **Linkage** and **Composite**, major gene effect is assumed, and recombination fraction $\theta_{AQ} = 0$; in the meantime, linkage equilibrium is assumed between QTL Q and the marker A. In the test case of **Admixture**, population admixture is generated by mixing families equally draw from one of the two sub-populations C and D. In both sub-populations C and D, no major gene effect or familial effect is assumed, i.e., $\sigma_{ga}^2 = \sigma_{Ga}^2 = 0$. However, the

Table 5: The parameters of the simulated genetic cases. The total variance is fixed as $\sigma^2 = \sigma^2_{ga} + \sigma^2_{Ga} + \sigma^2_e = 1$, and $\sigma^2_{gd} = 0$. **Admixture**: no major gene effect or familial effect $\sigma^2_g = \sigma^2_{Ga} = 0$, but with population admixture (see text for explanation).

Test Cases	σ^2_{ga}	σ^2_{Ga}	σ^2_e	σ^2	θ_{AQ}	q_1	$D_{A_gQ}, g = 1, \cdots, m$
Null	0	0	1.0	1.0	0.5	Not Applied	Not Applied
Familiality	0	0.5	0.5	1.0	0.5	Not Applied	Not Applied
Linkage	0.5	0	0.5	1.0	0	0.5	0
Composite	0.2	0.3	0.5	1.0	0	0.5	0

trait mean of sub-population C is fixed as 1 and the variance is fixed as 1. The trait mean of sub-population D is fixed as 0 and the variance is fixed as 1. Therefore, the total variance in the mixing population is $\sigma^2 = 1.25$. The admixture contributed to $(1-0)^2/(4 \times 1.25) = 0.20$ of the total variance.

To calculate the type I error rates of Tables 6 and 7, 1000 datasets are simulated for each test case. Each dataset contains 50 pedigrees of either graph A or graph B of Figure 3, respectively. Using the datasets, we fit the model (3.3) and test the null hypothesis $H_{a0} : \alpha_1 = \cdots = \alpha_m$. Since the QTL Q is in linkage equilibrium with marker A, an empirical test statistic which is larger than the cutting point at a 0.05 significance level is treated as a false positive. Based on likelihood ratio test, type I error rates are calculated as the proportions of the 1000 simulation datasets which give significant result at the 0.05 significant level. The results of type I error rates are presented in Tables 6 and 7. The results show that the type I error rates are around the nominal level 0.05. Hence, the model is reasonably robust. In all the four missing rate cases ($\varepsilon_A = 0.05, 0.10, 0.15, 0.20$), the type I error rates are reasonable. Hence, the missingness does not affect the robustness of the model.

In Table 8, we show the type I error rates using tri-nuclear families. Each tri-nuclear family contains three people, parents and an offspring. Again, 1000 datasets are simulated for each test case. Each dataset contains 50 tri-nuclear pedigrees.

Table 6: Type I error rates (%) at a 0.05 significance level of Small 3-generation pedigree A) based on likelihood ratio tests.

No. of Alleles	**Test Case**	**Error Rates**			
		$\varepsilon_A = 0.05$	$\varepsilon_A = 0.10$	$\varepsilon_A = 0.15$	$\varepsilon_A = 0.20$
Di-allele	Null	4.2	4.9	4.9	4.6
m = 2	Familiality	3.8	4.3	5.2	4.2
	Admixture	4.2	4.5	4.4	4.5
	Linkage	5.7	4.8	5.0	5.3
	Composite	5.2	4.6	4.9	5.0
Tri-allele	Null	4.8	3.9	4.5	3.7
m = 3	Familiality	5.1	5.1	5.4	5.4
	Admixture	4.0	3.7	4.2	4.4
	Linkage	5.6	3.9	5.0	5.8
	Composite	4.5	4.6	5.0	4.9
Quadri	Null	4.9	4.8	5.4	4.6
-allele	Familiality	4.6	5.2	4.6	5.4
m = 4	Admixture	4.4	3.7	4.4	4.4
	Linkage	5.4	5.3	5.7	5.2
	Composite	5.4	5.4	5.4	4.8

Two types of calculation are performed: (1) imputing genotypes which are missing by the proposed method, and keep every individual in the analysis; (2) removing all individuals from analysis whose genotypes are missing. It can be seen that the proposed method can help to get correct type I error rates, by imputing genotypes which are missing, since the type I error rates are around the nominal level 0.05. In the previous approach, an individual is deleted from analysis once his/her genotype is missing; it may inflate type I error rates by the results of Table 8, since the type I error rates are around 0.06.

It worths notice that the calculations are based on tri-nuclear family in Table 8. The sample size of 50 tri-nuclear families is 150, which is moderate. If the individuals with missing genotypes are removed from analysis, the sample sizes will be reduced and this can lead to the inflation of type I error rates. In Table 6, the small three-generation pedigree contains 11 people (and so a sample of 50 families is 550), and in

Table 7: Type I error rates (%) at a 0.05 significance level of Large 3-generation pedigree B) based on likelihood ratio tests.

No. of Alleles	**Test Case**	**Error Rates**			
		$\varepsilon_A = 0.05$	$\varepsilon_A = 0.10$	$\varepsilon_A = 0.15$	$\varepsilon_A = 0.20$
Di-allele	Null	4.1	4.3	4.0	4.0
m = 2	Familiality	5.4	5.6	5.0	4.7
	Admixture	5.4	4.8	4.6	3.5
	Linkage	4.5	4.4	4.7	4.4
	Composite	5.3	5.1	4.8	5.0
Tri-allele	Null	4.8	5.0	5.8	5.0
m = 3	Familiality	5.0	4.8	5.7	5.6
	Admixture	3.8	5.4	5.5	5.8
	Linkage	5.4	4.9	5.2	5.1
	Composite	5.9	5.6	5.6	5.1
Quadri	Null	4.3	4.6	5.6	5.2
-allele	Familiality	4.3	3.9	4.5	4.6
m = 4	Admixture	4.8	5.2	4.3	5.4
	Linkage	5.6	5.8	4.9	4.8
	Composite	5.1	5.2	5.4	4.3

Table 7, the large three-generation pedigree contains 18 individuals (and so a sample of 50 families is 900). Thus, the sample size is already big for the calculations of Table 6 and 7, and the results are reasonable. In one word, the proposed method can be useful for moderate-sized sample with missing genotypes.

3.3.2 Power Comparison

Let us denote the heritability by h^2, which is defined as $h^2 = \sigma^2_{ga}/\sigma^2$ (Falconer and Mackay, 1996). To make power comparison, we plot power curves of two cases using the related noncentrality parameter approximation: a dominant mode of inheritance, $a = 1, d = 1$, and a recessive mode of inheritance, $a = 1, d = -0.5$. Figures 4 and 5 show power curve of population samples at 0.01: Figure 4 are based one models (3.1) and (3.3) using one marker A; Figure 5 are based one models (3.10) and (3.12) using two markers A and B. The power curves of Figure 4 are plotted against the

Table 8: Type I error rates (%) at a 0.05 significance level of 50 tri-nuclear families based on likelihood ratio tests for quadri-allele case, i.e., $m = 4$.

Method	**Test Case**	**Error Rates**			
		$\varepsilon_A = 0.05$	$\varepsilon_A = 0.10$	$\varepsilon_A = 0.15$	$\varepsilon_A = 0.20$
Proposed method: Imputing all genotypes which are missing	Null	5.3	4.1	4.9	4.6
	Familiality	5.6	5.3	5.3	4.4
	Admixture	5.2	5.1	5.1	5.5
	Linkage	5.2	5.1	5.2	5.4
	Composite	5.5	5.5	4.9	4.7
Previous method: Removing all individuals with missing genotypes	Null	6.4	5.6	6.0	5.6
	Familiality	5.7	5.9	5.8	6.2
	Admixture	5.7	5.8	6.0	6.3
	Linkage	6.0	6.0	6.2	6.0
	Composite	5.7	6.0	5.8	5.9

heritability h^2; for graphs I and II, the marker A has three equal frequency alleles; for graphs III and IV, the marker A has four frequency alleles; and the related parameters are given in the legend of the Figure. Two features can be noted from Figure 4: (1) the power based on "genotype effect model" (3.1) is generally lower than that of the "additive effect model" (3.3); (2) the power is reasonably high when the heritability h^2 is larger than 0.15. The power curves of Figure 5 are plotted against the LD measure D_{A_1Q}; for graphs I and II, there are no missing genotypes, i.e., $\varepsilon_A = \varepsilon_B = 0.0$; for graphs III and IV, there are missing genotypes, and $\varepsilon_A = \varepsilon_B = 0.25$. It is obvious that missing genotypes lead to power decreases, since the noncentrality parameter approximations are reduced.

The power curves of Figures 6, 7, and 8 are based on pedigree data: Figure 6 is based on nuclear families in which each family consists of two parents and two children; Figure 7 is based 30 small 3-generation pedigrees (Graph I, Figure 3); and Figure 8 is based 30 large 3-generation pedigrees (Graph II, Figure 3). Such as the population data, the three Figures show that missing genotypes lead to power decreases. In addition, the power based on "genotype effect model" (3.12) is generally

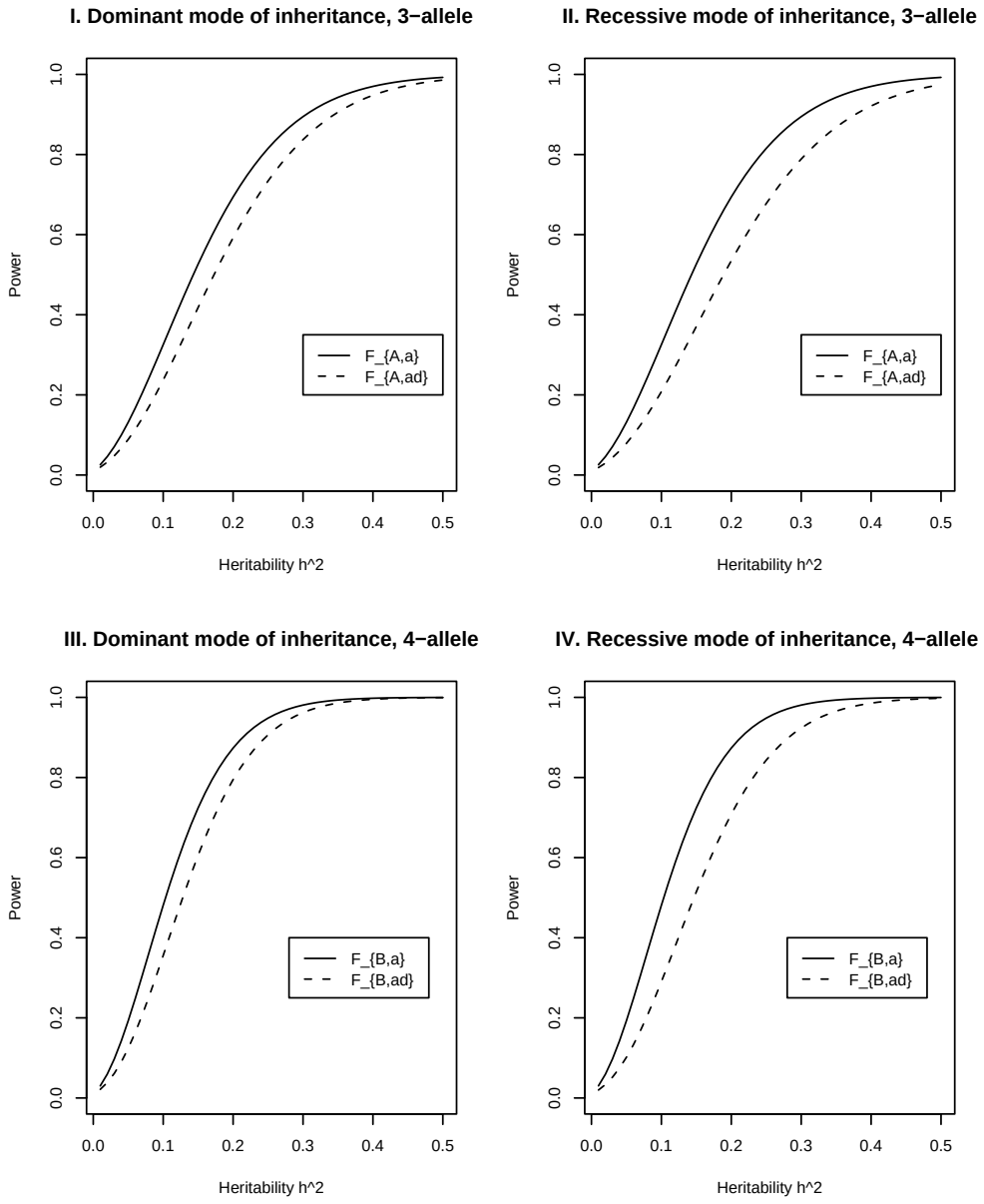

Figure 4: Power curve of population sample at 0.01 level based on models (3.1) and (3.3), where $N = 250, \varepsilon_A = 0.1, q_1 = 0.5, \sigma^2_{Ga} = 0.10$. For graphs I and II, the marker A has three alleles and $P_{A_i} = 1/3, i = 1, 2, 3, D_{A_1Q} = 0.12, D_{A_2Q} = D_{A_3Q} = -0.06$; for graphs III and IV, the marker A has four alleles, $P_{A_i} = 0.25, i = 1, \cdots, 4, D_{A_1Q} = -D_{A_2Q} = D_{A_3Q} = -D_{A_4Q} = 0.08$; for dominant mode of inheritance of graphs I and III, $a = 1, d = 1$; for recessive mode of inheritance of graphs II and IV, $a = 1, d = -0.5$.

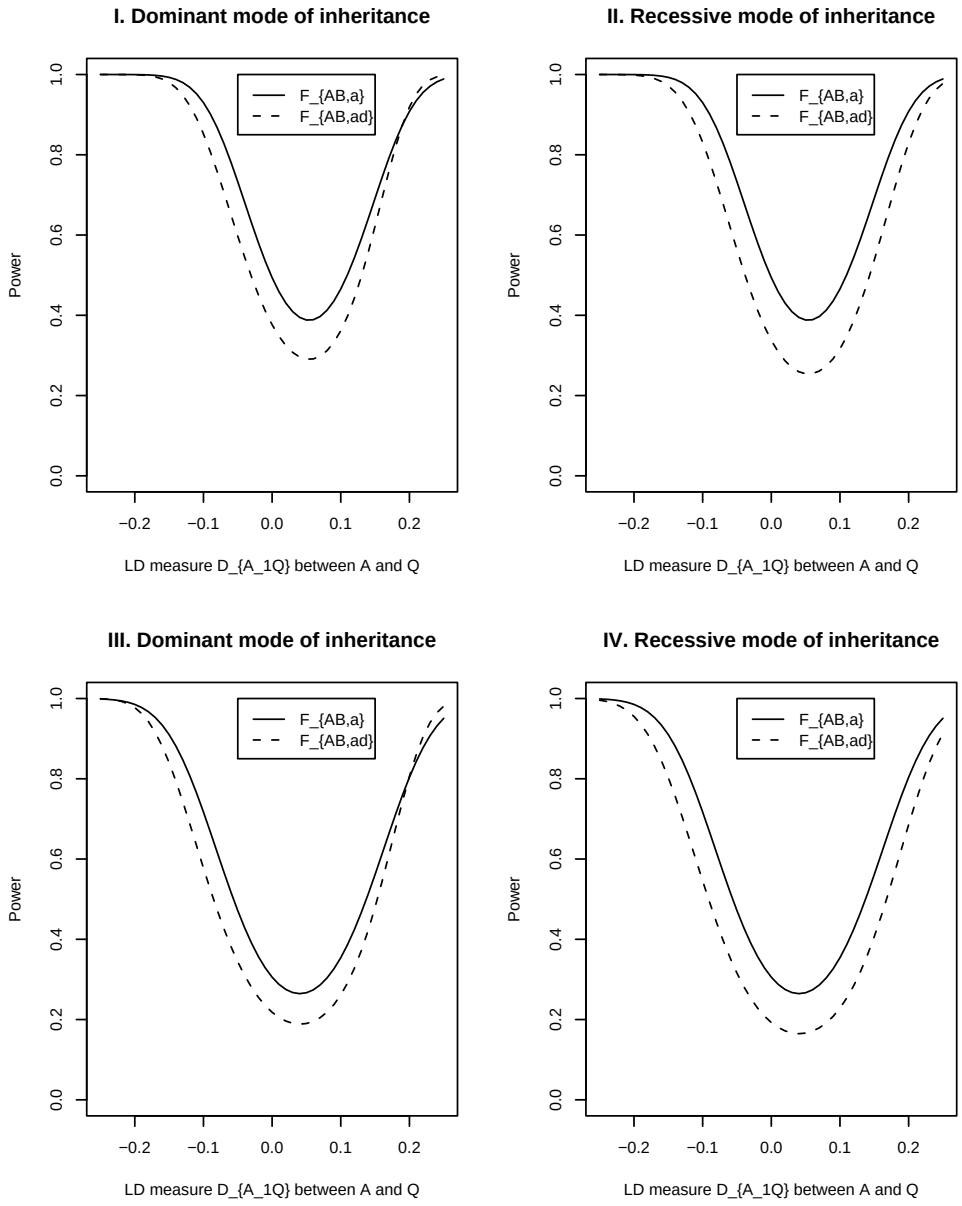

Figure 5: Power curve of population sample at 0.01 level based on models (3.10) and (3.12), where $N = 200, m = 2, n = 3, q_1 = q_2 = P_{A_1} = P_{A_2} = 0.5, P_{B_i} = 1/3, i = 1, 2, 3, D_{B_1Q} = D_{B_2Q} = 0.06, D_{A_1B_1} = D_{A_1B_2} = 0.05, \sigma^2_{Ga} = 0.10, h^2 = 0.15$. For graphs I and II, $\varepsilon_A = \varepsilon_B = 0.0$; for graphs III and IV, $\varepsilon_A = \varepsilon_B = 0.25$; for dominant mode of inheritance of graphs I and III, $a = 1, d = 1$; for recessive mode of inheritance of graphs II and IV, $a = 1, d = -0.5$.

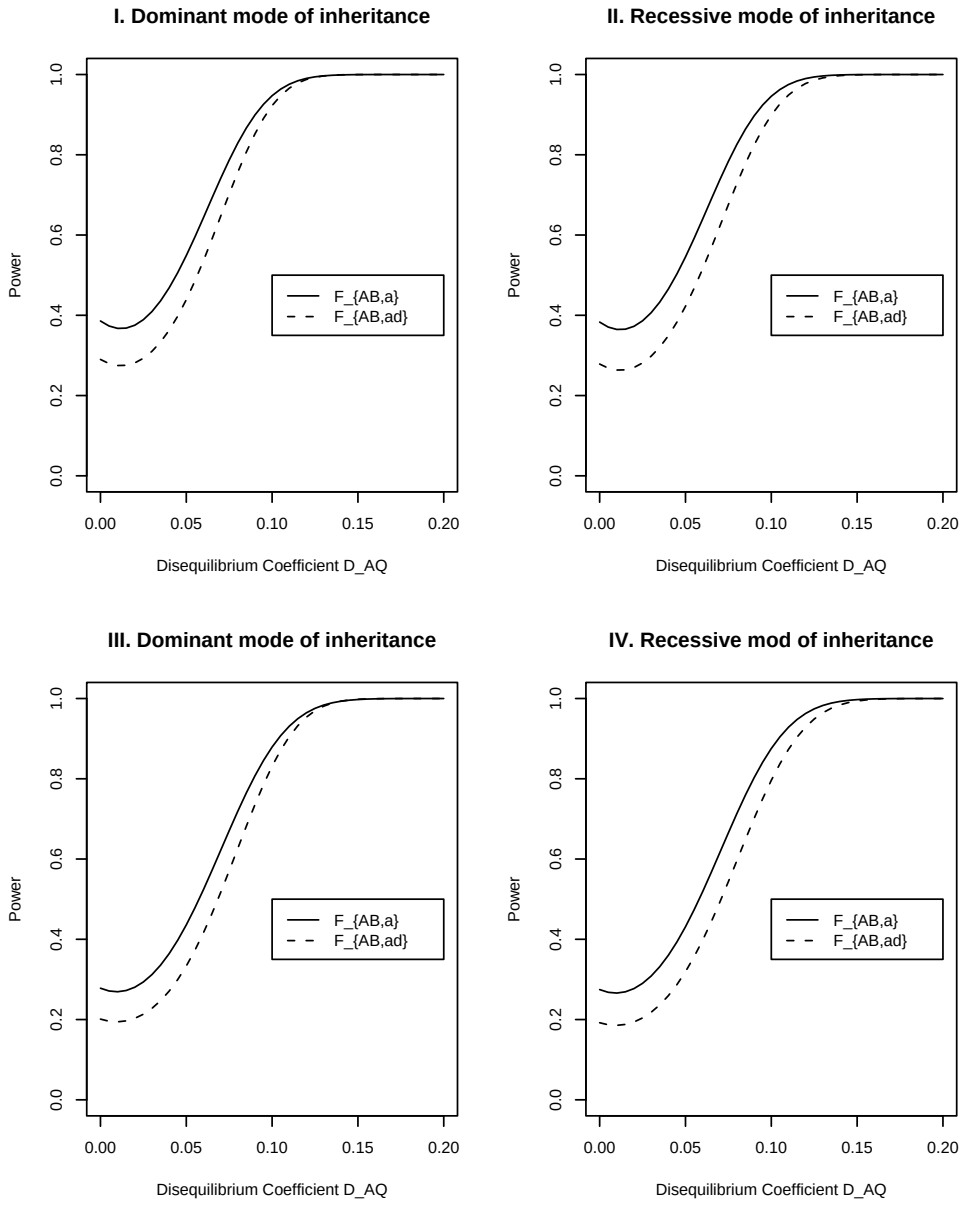

Figure 6: Power curve of 200 nuclear families at 0.01 level based on models (3.10) and (3.12), where $m = n = 2, q_1 = P_{A_1} = P_{B_1} = 0.5, D_{A_1B_1} = 0.05, D_{B_1Q} = 0.06, \sigma^2_{Ga} = 0.10, h^2 = 0.15$. Here, each nuclear family has two children. For graphs I and II, $\varepsilon_A = \varepsilon_B = 0$; for graphs III and IV, $\varepsilon_A = \varepsilon_B = 0.25$; for dominant mode of inheritance of graphs I and III, $a = 1, d = 1$; for recessive mode of inheritance of graphs II and IV, $a = 1, d = -0.5$.

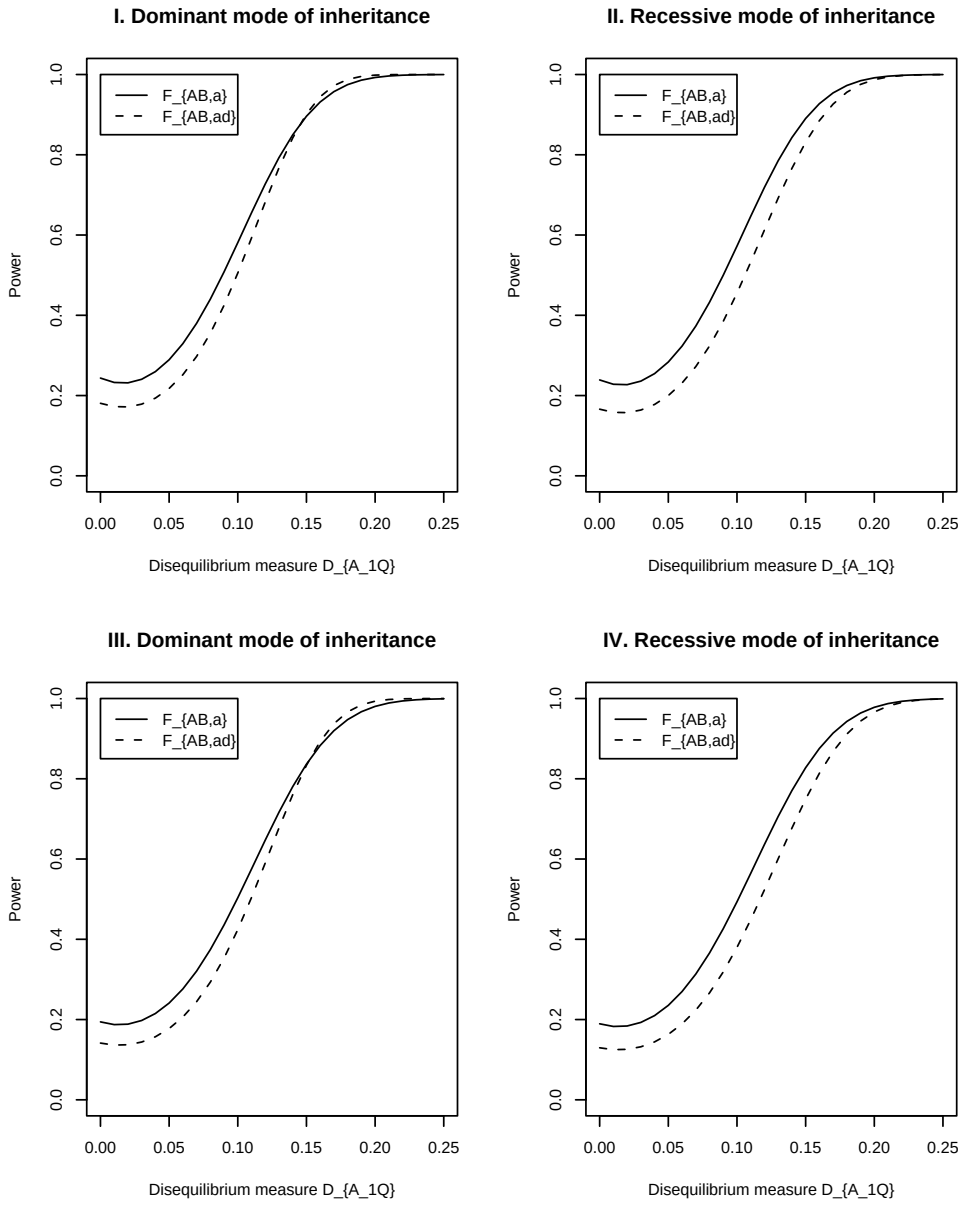

Figure 7: Power curve of 45 small 3-generation pedigrees (Graph A, Figure 3) at 0.01 level based on models (3.10) and (3.12), where $m = n = 2, q_1 = P_{A_1} = P_{B_1} = 0.5, D_{B_1Q} = 0.08, D_{A_1B_1} = 0.05, \sigma^2_{Ga} = 0.10, h^2 = 0.15$. For graphs I and II, $\varepsilon_A = \varepsilon_B = 0.0$; for graphs III and IV, $\varepsilon_A = \varepsilon_B = 0.25$; for dominant mode of inheritance of graphs I and III, $a = 1, d = 1$; for recessive mode of inheritance of graphs II and IV, $a = 1, d = -0.5$.

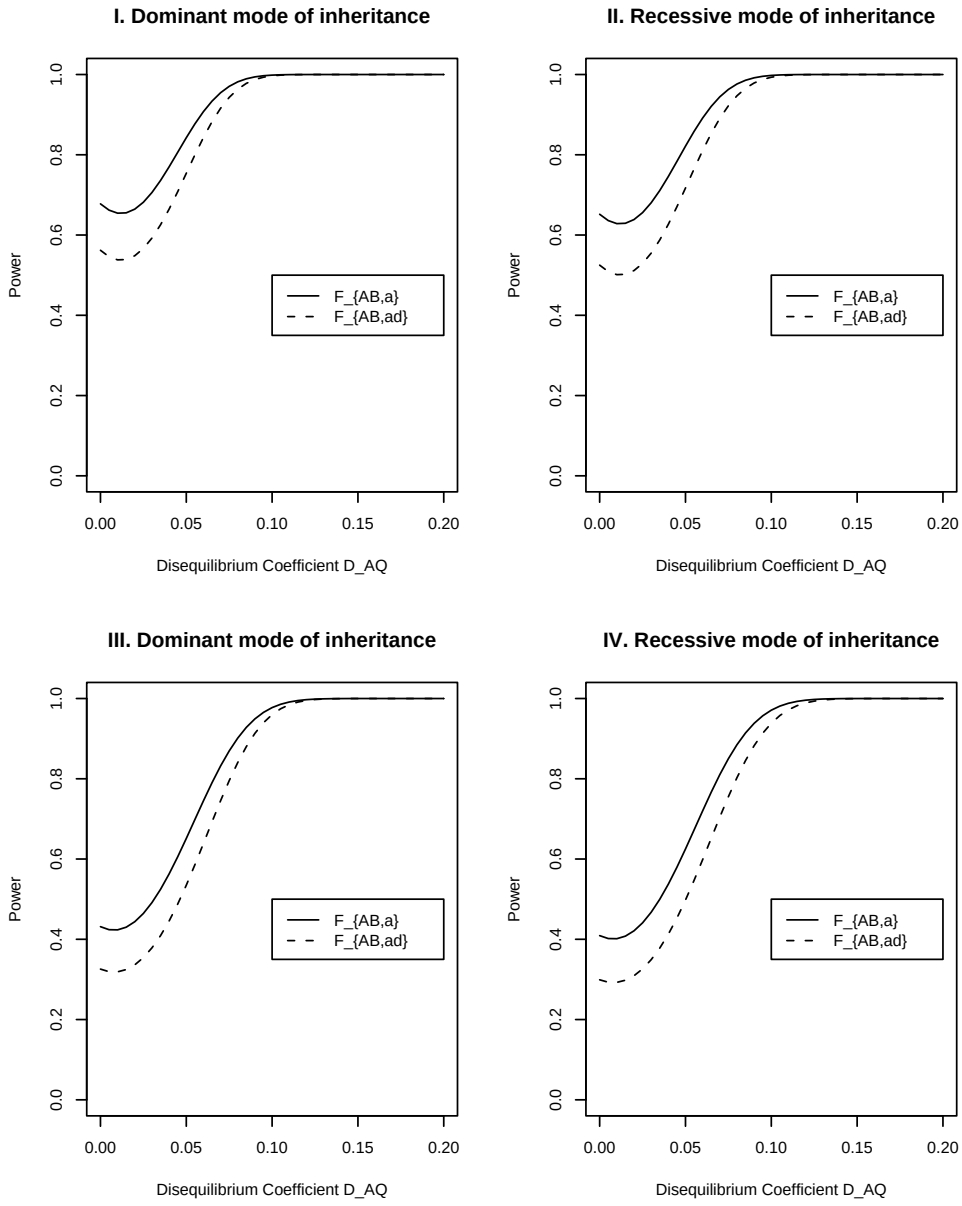

Figure 8: Power curve of 30 large 3-generation pedigrees (Graph B, Figure 3) at 0.01 level based on models (3.10) and (3.12), where $m = n = 2, q_1 = P_{A_1} = P_{B_1} = 0.5, D_{B_1Q} = 0.06, D_{A_1B_1} = 0.05, \sigma^2_{Ga} = 0.10, h^2 = 0.15$. For graphs I and II, $\varepsilon_A = \varepsilon_B = 0.0$; for graphs III and IV, $\varepsilon_A = \varepsilon_B = 0.25$; for dominant mode of inheritance of graphs I and III, $a = 1, d = 1$; for recessive mode of inheritance of graphs II and IV, $a = 1, d = -0.5$.

lower than that of the "additive effect model" (3.10).

3.4 Examples

3.4.1 The European ACE Data

The proposed method is applied to analyze angiotensin-1 converting enzyme data (Farrall et al., 1999; Keavney et al., 1998). The data consist of 83 extended families with between 4 and 18 members. Circulating ACE levels were measured for 405 individuals. Ten bi-allelic polymorphisms in the ACE gene were genotyped. There is missing genotype information at markers. Although we can not rigorously show that the missingness is MCAR, it is roughly correct since there is no systematic pattern in the missingness; actually, either founder's or non-founder's genotypes or both can be missing in a pedigree. In addition, the missingness is different from marker to marker, i.e., genotypes of an individual at some markers are missing, and are not missing at other markers. In our previous study, all individuals with missing genotypes are deleted, and so the total number of individuals is different from marker to marker (Fan et al., 2005; refer to column 2 of Table 9). For instance, there are 4 individuals whose genotype information is missing at marker I/D and the total number $N = 401$ in the previous study; at marker G2350A, on the other hand, there are more missing genotype data, and $N = 365$. In this dissertation research, all 405 individuals are used in the analysis using the developed variance component models for each marker.

Before fitting the models, multi-point IBD at each marker are calculated by Merlin (Abecasis et al., 2002). Variance component linkage analysis shows that additive variances are significantly larger than 0, but dominant variances are not significantly larger than 0 (Abecasis et al., 2000b). Hence, dominant effects can be excluded from regression equation, and the total variance is modeled as $\sigma^2 = \sigma^2_{ga} + \sigma^2_{Ga} + \sigma^2_e$. Table 9 shows LD analysis of the ACE gene by individual marker. To make com-

Table 9: Linkage disequilibrium analysis of the European ACE data by individual marker. The **AbAw's lod** is taken from Table 4 of Abecasis et al. (2000b). The **lod without missing** is taken from Table 5, column 4, Fan et al. (2005), which is calculated by deleting all individuals when their genotypes are missing. The **lod with missing** is calculated based on the model developed in this chapter. Abbreviation: **ind.** is individuals.

	Previous Results			**lod w/ missing**
Marker A	**# of ind.**	**AbAw's lod**	**lod w/o missing**	**# of ind.=405**
T-5491C	391	9.86	13.91	13.34
A-5466C	392	9.04	14.06	14.25
T-3892C	400	12.49	18.27	17.58
A-240T	401	10.81	13.35	13.66
T-93C	392	10.93	13.00	13.13
T-1237C	377	11.52	20.59	17.94
G2215A	372	14.91	27.01	24.78
I/D	401	15.76	27.37	27.59
G2350A	365	14.40	28.01	26.13
4656(CT)3/2	390	14.22	27.93	27.16

parison with the "AbAw" approach, results of AbAw's lod are taken from Table 4 of Abecasis et al. (2000b). After fitting the proposed models in this article, **lod** is calculated by $LRT/(2\log 10)$, where $LRT = 2(L_1 - L_0)$, L_1 is the log-likelihood under $y_{ij} = \beta + x_{Aij}\alpha_A + G_{ij} + e_{ij}$, and L_0 is the log-likelihood under $y_{ij} = \beta + G_{ij} + e_{ij}$. The **lod without missing** is taken from Table 5, column 4, Fan et al. (2005), which is calculated by deleting all individuals when their genotypes are missing. The **lod with missing** is calculated based on the model developed in this chapter.

The lod scores calculated by the proposed method in this article are similar to those in our previous study for most markers (column 4 and column 5, Table 9). Hence, whether the individuals with missing genotypes are removed from the analysis or not does not influence the conclusion we reached. The results of Table 9 confirm the finding that the association is strongest around the G2215A, I/D, G2350A and 4656(CT)3/2 polymorphisms (Abecasis et al., 2000b). Therefore, these markers are likely in complete LD with the trait alleles. In addition, the lod scores based on

our approach are generally higher than those of the "AbAw" approach. The Lod scores calculated by the proposed method at three markers, T-1237C, G2215A and G2350A, show big decreases compared with those of our previous study. This is most likely due to the fluctuations from the missingness. In our previous study, we found that allele I at marker I/D is almost always present with allele A at marker G2350A, and allele D at marker I/D is almost always present with allele G at marker G2350A (Fan et al., 2005). The frequency of the four haplotypes of these two markers are as follows: $P(IA) = 0.478875, P(IG) = 0.002817, P(DG) = 0.515494, P(DA) = 0.002817$. Besides, the measure of LD is 0.2468478 between the markers I/D and G2350A. The two markers are almost in complete LD with each other. However, the lod scores of these two markers are different from each other, which is most likely due to that there are more missing genotypes at marker G2350A.

3.4.2 The Nigeria ACE Data

This dataset consists of 233 Nigerian families with 1694 individuals, and 786 individuals are phenotyped with plasma ACE concentrations. This dataset were collected through the multi-center International Collaborative Study on Hypertension in Blacks sampling frame (Cox et al., 2002). The samples were genotyped at 35 markers in families of African ancestry, resulting in significant genetic diversity and a strong ACE association signal. Since 7 markers have no location information, we only use the remaining 28 markers in analysis (Table 1, Cox et al., 2002). As with the Oxford European ACE data, there are missing genotypes for the data. Hence, we analyze the data using the proposed methods.

Variance component linkage analysis shows that additive variance is significantly larger than 0, but dominant variance and polygenic variance are not significantly larger than 0. Hence, the total variance is modeled as $\sigma^2 = \sigma^2_{ga} + \sigma^2_e$. After fit-

Table 10: Linkage disequilibrium analysis of the Nigerian ACE data by individual marker. The **Previous** method is to calculate by deleting all individuals when their genotypes are missing. The **Proposed** method is to calculate based on the model developed in Chapter IV. For **Proposed** method, # of individuals = 786.

Method	**Previous**				**Proposed**	
Marker A	**# of ind.**	**AbAw Lod**	**Lod**	**F-stat**	**Lod**	**F-stat**
A11377T	696	4.79	10.41	48.62	10.52	49.74
T11434C	666	0.93	0.51	1.54	0.28	1.13
T16778C	671	0.01	0.42	1.15	0.23	0.94
G16804A	678	1.16	1.19	4.83	0.73	3.24
C17105T	686	0.03	0.25	0.16	0.03	0.00
A17554G	693	*	0.28	0.24	0.06	0.11
A21235G	694	3.81	8.60	40.10	8.40	39.48
D/I-23069	695	3.95	9.33	43.69	9.06	42.68
G23454C	660	3.49	8.66	40.99	7.80	37.25
A23462C	657	4.33	10.18	48.59	9.04	43.39
A23495G	692	13.62	31.73	162.45	31.12	157.16
G24188A	673	2.18	8.55	40.62	8.03	38.01
A24409G	679	0.39	0.95	4.24	0.91	4.07
A24418G	677	0.00	0.18	0.55	0.28	1.12
C28919T	675	0.04	0.06	0.13	0.05	0.06
G28952C	679	*	0.80	3.40	0.47	2.03
T29035C	681	0.55	2.92	13.27	2.87	13.13
C29097T	678	3.87	11.51	55.23	11.46	54.72
C29302T	672	0.01	2.03	9.23	1.92	8.77
29349delT	678	*	1.80	8.10	1.67	7.58
G29373A	673	0.03	0.06	0.03	0.04	0.03
C29809T	675	9.77	24.80	124.65	24.60	122.45
31839insC	652	0.40	0.57	2.24	0.35	1.44
$tgccc_{2/1}$-31933	654	0.03	0.66	2.63	0.70	3.08
A31958G	654	8.58	31.27	162.64	30.21	154.04
C32128A	651	2.55	6.93	32.39	6.42	30.16
G32178A	667	*	0.18	0.79	0.20	0.76
$(CT)_{3/2}$-32915	674	1.14	2.05	7.64	1.90	8.63

ting the proposed models in this chapter, lod is calculated by $LRT/(2\log 10)$, where $LRT = 2(L_1 - L_0)$, L_1 is the log-likelihood under $y_{ij} = \alpha + x^{(1)}_{Aij}\alpha_A + e_{ij}$, and L_0 is the log-likelihood under $y_{ij} = \alpha + e_{ij}$. The **Previous** methods are calculated by deleting all individuals when their genotypes are missing. The **Proposed** methods

are calculated based on the model developed in this chapter. Table 10 shows LD analysis of the ACE gene by individual marker. The most strongly associated polymorphisms include A23495G, A31958G, C29809T, C29097T, A11377T, etc, in the order of lod scores and F-statistics. After identifying the most strongly associated polymorphism A23495G, we treat it as marker A. The other markers are treated as marker B, and we fit model $y_{ij} = \alpha + x^{(1)}_{Aij}\alpha_A + x^{(1)}_{Bij}\alpha_B + e_{ij}$. Table 11 shows the results by testing $H_0 : \alpha_B = 0$. In addition to marker A = A23495G, Table 11 shows the most strongly polymorphisms include A31958G, C29809T, A11377T, C29097T, C32128A, at a significance level 0.001. Based on the results of Tables 10 and 11, we treat polymorphism A23495G as marker A and polymorphism A31958G as B, and continue to add more markers in the analysis. Table 12 shows that the polymorphism A11377T is the most strongly associated one at a significance level 0.001, in addition to polymorphisms A = A23495G and B = A31958G. Table 13 show that no more polymorphism is associated with the the plasma ACE concentrations at a significance level 0.001, in addition to polymorphisms A = A23495G, B = A31958G and C = A11377T.

3.5 Discussion

In searching for common genes of complex traits, large samples are needed that are likely to come only from combining family and population based data. In addition, sophisticated methods are needed to analyze these combinations. The statistical and mathematical methods and models must account for missing data, and must account for environmental covariates which are certain to play a role in complex diseases. In this research, we have extended previous variance component models for combined linkage and association mapping of QTL in the presence of missing genotypes. Under an assumption of MCAR, two regression models, “genotype effect

Table 11: Linkage disequilibrium analysis of the Nigerian ACE data by two markers. Regressions are given by: (1) $y_{ij} = \alpha + e_{ij}$; (2) $y_{ij} = \alpha + x^{(1)}_{Aij}\alpha_A + e_{ij}$; (3) $y_{ij} = \alpha + x^{(1)}_{Aij}\alpha_A + x^{(1)}_{Bij}\alpha_B + e_{ij}$. **log-L** is log-likelihood. **lod**=$LRT/(2\log 10)$ and **LRT**=$2(L_1 - L_0)$, where L_0 is the log-likelihood under the null hypothesis H_0, and L_1 is that under the alternative. For example, $143.320 = 2(-362.844 + 434.504)$ and $44.618 = 2(-340.535 + 362.844)$, in the case of $A = A23495G, B = A31958G$. The results are calculated based on the proposed models. The number of individual is 786.

Markers	**Reg.**	**log-L**	H_0	**LRT**	**lod**	**p-value**
	(1)	-434.504				
A: A23495G	(2)	-362.844	$\alpha_A = 0$	143.32	31.12	< 0.001
B: A31958G	(3)	-340.535	$\alpha_B = 0$	44.618	9.689	< 0.001
C29809T	(3)	-349.093	$\alpha_B = 0$	27.502	5.972	< 0.001
A11377T	(3)	-350.772	$\alpha_B = 0$	24.144	5.243	< 0.001
C29097T	(3)	-352.789	$\alpha_B = 0$	20.110	4.367	< 0.001
C32128A	(3)	-354.131	$\alpha_B = 0$	17.426	3.784	< 0.001
$(CT)_{3/2}$-32915	(3)	-357.719	$\alpha_B = 0$	10.250	2.226	0.001
T29035C	(3)	-358.159	$\alpha_B = 0$	9.370	2.035	0.002
T16778C	(3)	-358.557	$\alpha_B = 0$	8.574	1.862	0.003
C28919T	(3)	-359.106	$\alpha_B = 0$	7.476	1.623	0.006
A24418G	(3)	-359.604	$\alpha_B = 0$	6.480	1.407	0.011
G24188A	(3)	-359.793	$\alpha_B = 0$	6.102	1.325	0.014
29349delT	(3)	-360.239	$\alpha_B = 0$	5.210	1.131	0.022
A21235G	(3)	-360.511	$\alpha_B = 0$	4.666	1.013	0.031
D/I-23069	(3)	-360.515	$\alpha_B = 0$	4.658	1.011	0.031
A23462C	(3)	-360.607	$\alpha_B = 0$	4.474	0.972	0.034
G28952C	(3)	-361.29	$\alpha_B = 0$	3.108	0.675	0.078
G23454C	(3)	-361.39	$\alpha_B = 0$	2.908	0.631	0.088
G29373A	(3)	-361.703	$\alpha_B = 0$	2.282	0.496	0.131
C17105T	(3)	-361.802	$\alpha_B = 0$	2.084	0.453	0.149
C29302T	(3)	-362.167	$\alpha_B = 0$	1.354	0.294	0.245
A24409G	(3)	-362.449	$\alpha_B = 0$	0.790	0.172	0.374
G32178A	(3)	-362.559	$\alpha_B = 0$	0.570	0.124	0.450
$\text{tgccc}_{2/1}$-31933	(3)	-362.615	$\alpha_B = 0$	0.458	0.099	0.499
G16804A	(3)	-362.679	$\alpha_B = 0$	0.330	0.072	0.566
A17554G	(3)	-362.719	$\alpha_B = 0$	0.250	0.054	0.617
31839insC	(3)	-362.800	$\alpha_B = 0$	0.088	0.019	0.767
T11434C	(3)	-362.839	$\alpha_B = 0$	0.010	0.002	0.920

model" and "additive effect model", are proposed to model the association between the markers and the trait locus. If the marker genotypes are not missing, the model

Table 12: Linkage disequilibrium analysis of the Nigerian ACE data by three markers. Regressions are given by: (1) $y_{ij} = \alpha + e_{ij}$; (2) $y_{ij} = \alpha + x^{(1)}_{Aij}\alpha_A + e_{ij}$; (3) $y_{ij} = \alpha + x^{(1)}_{Aij}\alpha_A + x^{(1)}_{Bij}\alpha_B + e_{ij}$; (4) $y_{ij} = \alpha + x^{(1)}_{Aij}\alpha_A + x^{(1)}_{Bij}\alpha_B + x^{(1)}_{Cij}\alpha_C + e_{ij}$. The results are calculated based on the proposed models. The notations are the same as Table 11.

Markers		**Reg.**	**log-L**	H_0	**LRT**	**lod**	**p-value**
		(1)	-434.504				
A:	A23495G	(2)	-362.844	$\alpha_A = 0$	143.32	31.12	< 0.001
B:	A31958G	(3)	-340.535	$\alpha_B = 0$	44.618	9.689	< 0.001
C:	A11377T	(4)	-334.407	$\alpha_C = 0$	12.256	2.661	< 0.001
	C32128A	(4)	-336.764	$\alpha_C = 0$	7.542	1.638	0.006
	T16778C	(4)	-337.827	$\alpha_C = 0$	5.416	1.176	0.020
	delT29349	(4)	-337.900	$\alpha_C = 0$	5.270	1.144	0.022
	$(CT)_{3/2}$-32915	(4)	-337.946	$\alpha_C = 0$	5.178	1.124	0.023
	T29035C	(4)	-338.158	$\alpha_C = 0$	4.754	1.032	0.029
	C29097T	(4)	-338.475	$\alpha_C = 0$	4.120	0.895	0.042

is exactly the same as those of previous study, i.e., the number of genotypes or alleles is used as weight to model the effect of the genotypes or alleles in single marker case. If the marker genotypes are missing, the expected number of genotypes or alleles is used as weight to model the effect of the genotypes or alleles. The "genotype effect model" can be used to model the additive and dominance effects simultaneously; the "additive effect model" only takes care of additive effect. Based on the two models, F-test statistics are proposed to test association between the QTL and markers. The noncentrality parameter approximations of F-test statistics are derived to make power calculation and comparison, which show that the power of the F-tests is reduced due to the missingness. In addition, likelihood ratio test statistics can be used to test the association. Under the assumption that the genotype data are MCAR, simulation studies are performed to calculate the type I error rates to evaluate the robustness of the proposed models. It is found the type I error rates are reasonable. The method is applied to analyze the angiotensin-1 converting enzyme data.

In this chapter, the genotypes are assumed to be missing completely at random.

Table 13: Linkage disequilibrium analysis of the Nigerian ACE data by four markers. Regressions are given by: (1) $y_{ij} = \alpha + e_{ij}$; (2) $y_{ij} = \alpha + x^{(1)}_{Aij}\alpha_A + e_{ij}$; (3) $y_{ij} = \alpha + x^{(1)}_{Aij}\alpha_A + x^{(1)}_{Bij}\alpha_B + e_{ij}$; (4) $y_{ij} = \alpha + x^{(1)}_{Aij}\alpha_A + x^{(1)}_{Bij}\alpha_B + x^{(1)}_{Cij}\alpha_C + e_{ij}$. (5) $y_{ij} = \alpha + x^{(1)}_{Aij}\alpha_A + x^{(1)}_{Bij}\alpha_B + x^{(1)}_{Cij}\alpha_C + x^{(1)}_{Dij}\alpha_D + e_{ij}$. The results are calculated based on the proposed models. The notations are the same as Table 11.

Markers		**Reg.**	**log-L**	H_0	**LRT**	**lod**	**p-value**
		(1)	-434.504				
A:	A23495G	(2)	-362.844	$\alpha_A = 0$	143.32	31.12	< 0.001
B:	A31958G	(3)	-340.535	$\alpha_B = 0$	44.618	9.689	< 0.001
C:	A11377T	(4)	-334.407	$\alpha_C = 0$	12.256	2.661	< 0.001
D:	delT29349	(5)	-331.276	$\alpha_D = 0$	6.262	1.360	0.012
D:	T16778C	(5)	-332.148	$\alpha_D = 0$	4.518	0.981	0.034
D:	$(CT)_{3/2}$-32915	(5)	-332.309	$\alpha_D = 0$	4.196	0.911	0.041
D:	C32128A	(5)	-333.060	$\alpha_D = 0$	2.694	0.585	0.101
D:	T29035C	(5)	-333.242	$\alpha_D = 0$	2.330	0.506	0.127
D:	C29097T	(5)	-334.342	$\alpha_D = 0$	0.130	0.028	0.718

This assumption is roughly true for some study, such as the angiotensin-1 converting enzyme data. For other studies, the missingness can be systematic, e.g., some or all of the founder genotypes can be missing (Wang and Elston, 2005). It is unclear how this will affect the proposed models. In practice, the assumption of MCAR is unlikely to be true. For instance, consider a tri-nuclear family. Assume that the two parents' genotypes are 1/2 and 1/1, and the genotype of the offspring is missing. It is easy to see that the genotype of the offspring can be 1/1 or 2/1 with a probability 0.5 for each genotype. Hence, the right imputing method is to assign a weight of 0.5 for each of these two genotypes. Besides, other genotypes such as 2/2 should be assigned a weight of 0. In short, information of genotypes of family members can be used to infer the missing genotype of another family member. In this way, it is very likely that the power can be improved. However, it won't be easy to get neat noncentrality parameter approximations as the ones we have in the research under an assumption of MCAR. Instead, simulation study is a possible method for the investigation. We

leave this issue to be investigated in the future study.

The proposed models can be used to analyzed either single marker or multiple genotype data. However, the models can only be used to analyze one phenotype. As the ability to generate more genetics data, both for phenotypes and for genotypes, increases, additional statistical methods are required to evaluate multiple phenotypes, multiple genotypes. More research is necessary to extend existing theoretical methods to new data analytic situations of interest, including analyzing multivariate phenotypes in a complex disease setting.

CHAPTER IV

SUMMARY AND FUTURE RESEARCH

4.1 Summary

In this dissertation, we first have studied a semiparametric measurement error problem in a canonical exponential family framework. A functional method has been developed for generalized partially linear models with independent normal measurement error on the latent variable. Based upon the parametric idea of sufficiency scores, we have constructed unbiased score functions by conditioning on parametric-dependent sufficient statistics. Therefore, no distribution assumptions are made on the latent variable. Simulation studies and real data analyses showed that the proposed method performs better than the naive method. We have also proved asymptotic properties of the estimators.

Next, in a statistical genetics study, two regression models have been developed to investigate the impact of missing genotype for a high resolution combined linkage and association mapping of QTL. Based on the two models, F-test statistics or likelihood ratio test statistics have been used to test the association between the QTL and markers. Simulation studies have shown that the proposed method can help to get correct type I error rate for a moderate size dataset, although it cannot improve power. The method has been applied to analyze the ACE data.

4.2 Future Research

In Chapter II, we have made an assumption that the variance of measurement error Σ_{uu} is known. The reason for this assumption is for simplicity. The current research is a starting point for future research when we treat Σ_{uu} as another parameter. Given

replicated measurements, an unbiased estimating equation for Σ_{uu} is constructed in (2.12). This is a natural extension of current research.

In Chapter III, the missing mechanism is assumed as missing completely at random (MCAR). In practice, the assumption of MCAR is unlikely to be true. As we discussed in Section 3.5, information of genotypes of family members can be used to infer the missing genotype of another family member. It is very likely that the power can be improved by utilizing the pedigree structure. This is potential area for future research.

REFERENCES

Abecasis, G. R., Cardon, L. R., and Cookson, W. O. C. (2000a). A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics* **66**, 279–292.

Abecasis, G. R., Cherny, S. S., Cookson, W. O. C., and Cardon, L. R. (2002). Merlin — rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101.

Abecasis, G. R., Cookson, W. O. C., and Cardon, L. R. (2000b). Pedigree tests of linkage disequilibrium. *European Journal of Human Genetics* **8**, 545–551.

Allison, D. B. (2001). Joint tests of linkage and association for quantitative traits. *Theoretical Population Biology* **60**, 239–251.

Almasy, L. and Blangero, J. (1998). Multipoint quantitative trait linkage analysis in general pedigrees. *American Journal of Human Genetics* **62**, 1198–1211.

Almasy, L., Williams, J. T., Dyer, T. D., and Blangero, J. (1999). Quantitative trait locus detection using combined linkage/disequilibrium analysis. *Genetic Epidemiology* **17 (Suppl 1)**, S31–S36.

Amos, C. I. (1994). Robust variance-components approach for assessing linkage in pedigrees. *American Journal of Human Genetics* **54**, 534–543.

Amos, C. I. and Elston, R. C. (1989). Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genetic Epidemiology* **6**, 349–360.

Boerwinkle, E., Chakraborty, E., and Sing, C. F. (1986). The use of measured genotype information in the analysis of quantitative phenotype in man. I. models and analytical methods. *Annals of Human Genetics* **50**, 181–194.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Second Edition. New York: Chapman and Hall CRC Press.

Carroll, R. J., Ruppert, D., and Welsh, A. H. (1998). Local estimating equations. *Journal of the American Statistical Association* **93**, 214–227.

Carroll, R. J., Spiegelman, C., Lan, K. K., Bailey, K. T., and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika* **71**, 19–26.

Claeskens, G. and Carroll, R. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* **in press**.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics* **31**, 1852–1884.

Cox, R., Bouzekri, N., Martin, S., Southam, L., Hugill, A., Golamaully, M., Richard Cooper, R., Adeyemo, A., Soubrier, F., Ward, R., Lathrop, G. M., Matsuda, F., and Farrall, M. (2002). Angiotensin-1-converting enzyme (ACE) plasma concentration is influenced by multiple ACE-linked quantitative trait nucleotides. *Human Molecular Genetics* **11**, 2969–2977.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. 4th edition, London: Longman.

Fan, R. and Jung, J. (2003). High resolution joint linkage disequilibrium and linkage

mapping of quantitative trait loci based on sibship data. *Human Heredity* **56**, 166–187.

Fan, R., Jung, J., and Jin, L. (2006). High resolution association mapping of quantitative trait loci, a population based approach. *Genetics* **172**, 663–686.

Fan, R., Spinka, C., Jin, L., and Jung, J. (2005). Pedigree linkage disequilibrium mapping of quantitative trait loci. *European Journal of Human Genetics* **13**, 216–231.

Fan, R. and Xiong, M. (2002). High resolution mapping of quantitative trait loci by linkage disequilibrium analysis. *European Journal of Human Genetics* **10**, 607–615.

Fan, R. and Xiong, M. (2003). Combined high resolution linkage and association mapping of quantitative trait loci. *European Journal of Human Genetics* **11**, 125–137.

Farrall, M., Keavney, B., McKenzie, C. A., Delèpine, M., Matsuda, F., and Lathrop, G. M. (1999). Fine mapping of an ancestral recombination break-point in DCP1. *Nature Genetics* **23**, 270–271.

Feingold, E. (2002). Invited editorial: Regression-based quantitative-trait-locus mapping in the 21st century. *American Journal of Human Genetics* **71**, 217–222.

Fulker, D. W., Cherny, S. S., and Cardon, L. R. (1995). Multiple interval mapping of quantitative trait loci, using sib-pairs. *American Journal of Human Genetics* **56**, 1224–1233.

Fulker, D. W., Cherny, S. S., Sham, P. C., and Hewitt, J. K. (1999). Combined

linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics* **64**, 259–267.

Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

George, V., Tiwari, H. K., Zhu, X., and Elston, R. C. (1999). A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *American Journal of Human Genetics* **65**, 236–245.

Goldgar, D. E. (1990). Multipoint analysis of human quantitative genetic variation. *American Journal of Human Genetics* **47**, 957–967.

Graybill, F. A. (1976). *Theory and Application of the Linear Model*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* **2**, 3–19.

Hedrick, P. W. (1987). Gametic disequilibrium measures: Proceed with caution. *Genetics* **117**, 331–341.

Jung, J., Fan, R., and Jin, L. (2005). Combined linkage and association mapping of quantitative trait loci by multiple markers. *Genetics* **170**, 881–898.

Keavney, B., McKenzie, C. A., Connell, J. M., Julier, C., Peter, J., Ratcliffe, P. J., Sobel, E., Lathrop, M., and Farrall, M. (1998). Measured haplotype analysis of the angiotension-1 converting enzyme gene. *Human Molecular Genetics* **7**, 1745–1751.

Lange, K. (2002). *Mathematical and Statistical Methods for Genetic Analysis*. Second edition. New York: Springer.

Li, M., Boehnke, M., and Abecasis, G. R. (2005). Joint modeling of linkage and association: Identifying SNPs responsible for a linkage signal. *American Journal of Human Genetics* **76**, 934–949.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Second edition. Hoboken, NJ: John Wiley & Sons.

Liu, L. (2007). Estimation of generalized partially linear models with measurement error using sufficiency scores. *Statistics and Probability Letters* **in press**.

Maity, A., Ma, Y., and Carroll, R. J. (2007). Efficient estimation of population-level summaries in general semiparametric regression models. *Journal of the American Statistical Association* **102**, 123–139.

Pinherio, J. C. and Bates, D. M. (2000). *Mixed-effects Models in S and S-plus*. New York: Springer.

Pratt, S. C., Daly, M. J., and Kruglyak, L. (2000). Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. *American Journal of Human Genetics* **66**, 1153–1157.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semi-parametric models. *Journal of the American Statistical Association* **89**, 501–511.

Sham, P. C., Cherny, S. S., Purcell, S., and Hewitt, J. K. (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *American Journal of Human Genetics* **66**, 1616–1630.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74**, 703–714.

Wang, T. and Elston, R. C. (2005). The bias introduced by population stratification in IBD based linkage analysis. *Human Heredity* **60**, 134–142.

Xiong, M. and Jin, L. (2000). Combined linkage and linkage disequilibrium mapping for genome screens. *Genetic Epidemiology* **19**, 211–234.

APPENDIX A

CONDITIONS AND PROOFS IN CHAPTER II

A.1 Regularity Conditions

The regularity conditions given below are based on Claeskens and Carroll (2007) and Maity, Ma, and Carroll (2007).

(C1) The bandwidth sequence $h_n \to 0$ as $n \to \infty$, in such a way that $nh_n/\log^2(n) \to \infty$ and $h_n \geq \{\log(n)/n\}^{1-2/\lambda}$ for λ as in condition (C4) such that $nh_n^4 \to 0$.

(C2) The kernel function K is a symmetric, continuously differentiable pdf on $[-1, 1]$ taking on the value zero at the boundaries. The design density f_z is differentiable on $B = [b_1, b_2]$, the derivative is continuous, and $\inf_{z\in B} f_z(z) > 0$. The function $\theta(\cdot, \kappa)$ has second continuous derivatives on B and is also twice differentiable with respect to κ.

(C3) For $\kappa \neq \kappa_1$, the Kullback-Leibler distance between $\Psi\{\cdot, \cdot, \theta(\cdot, \kappa), \kappa\}$ and $\Psi\{\cdot, \cdot, \theta(\cdot, \kappa_1), \kappa_1\}$ is strictly positive. For every (y, δ), third partial derivatives of $\Psi\{y, \delta, \theta(z), \kappa\}$ with respect to κ exist and are continuous in κ. The 4th partial derivative exists for almost all (y, δ). Further, mixed partial derivatives $\frac{\partial^{r+s}}{\partial\kappa^r\partial v^s}\Psi\{y, \delta, v, \kappa)\}|_{v=\theta(z)}$, with $0 \leq r, s \leq 4, r + s \leq 4$ exist for almost all (y, δ) and $E\{\sup_\kappa \sup_v \left|\frac{\partial^{r+s}}{\partial\kappa^r\partial v^s}\Psi\{y, \delta, v, \kappa)\}\right|^2\} < \infty$.

(C4) Denote $\{\theta_0(z), \kappa_0\}$ as the true function and parameter. There exists a neighborhood $\mathcal{N}\{\theta_0(z), \kappa_0\}$ such that

$$\max_{k=1,2} \sup_{z\in B} \left\| \sup_{(\theta,\kappa)\in\mathcal{N}\{\theta_0(z),\kappa_0\}} \left| \frac{\partial^k}{\partial\theta^k} \log\{\Psi(Y, \Delta, \theta, \kappa)\} \right| \right\|_{\lambda,z} < \infty$$

for some $\lambda \in (2, \infty]$, where $\|\cdot\|_{\lambda,z}$ is the L^λ-norm, conditional on $Z = z$. Further,

$$\sup_{z\in B} E_z \left[\sup_{(\theta,\kappa)\in\mathcal{N}\{\theta_0(z),\kappa_0\}} |\frac{\partial^3}{\partial\theta^3}\log\{\Psi(Y,\Delta,\theta,\kappa)\}| \right] < \infty.$$

A.2 Sketch Proof of Theorem 2.4.3

The calculations given below basically follow those of Carroll et al. (1998), with the uniformity of the expansion following as in Claeskens and Van Keilegom (2003), see also Maity et al. (2007).

For given $\kappa = (\beta^{\mathrm{T}}, \phi)^{\mathrm{T}}$ and any z_0, we let $\alpha_0 = \theta_0(z_0)$, $\alpha_1 = h\theta_0'(z_0)$. Denote argument $(\cdot, x) = \{Y_i, \Delta_i(\beta), x, \kappa\}$. Taking a simple Taylor expansion to equation (2.5) with respect to (α_0, α_1) gives

$$\begin{aligned}
0 &= n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)(1, \frac{Z_i - z_0}{h})^{\mathrm{T}}\Psi_\theta(\cdot, \widehat{\alpha}_0 + \widehat{\alpha}_1\frac{Z_i - z_0}{h}) \\
&= n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)(1, \frac{Z_i - z_0}{h})^{\mathrm{T}}\Psi_\theta(\cdot, \alpha_0 + \alpha_1\frac{Z_i - z_0}{h}) + n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0) \\
&\quad \times(1, \frac{Z_i - z_0}{h})^{\mathrm{T}}(1, \frac{Z_i - z_0}{h})\Psi_{\theta\theta}(\cdot, \alpha_0 + \alpha_1\frac{Z_i - z_0}{h}) \begin{bmatrix} \widehat{\alpha}_0 - \alpha_0 \\ \widehat{\alpha}_1 - \alpha_1 \end{bmatrix} + o_p(n^{-1/2}). \quad \text{(A.1)}
\end{aligned}$$

Note that

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)(1, \frac{Z_i - z_0}{h})^{\mathrm{T}}(1, \frac{Z_i - z_0}{h})\Psi_{\theta\theta}(\cdot, \alpha_0 + \alpha_1\frac{Z_i - z_0}{h}) \\
\xrightarrow{p} \quad & f_z(z_0)E[\Psi_{\theta\theta}\{Y, \Delta(\beta), \theta(Z), \kappa\}|Z = z_0] \cdot \boldsymbol{I}_2,
\end{aligned}$$

where $\boldsymbol{I}_2$ is 2×2 identity matrix. Further by Taylor series expansion,

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)(1, \frac{Z_i - z_0}{h})^{\mathrm{T}} \Psi_\theta(\cdot, \alpha_0 + \alpha_1 \frac{Z_i - z_0}{h}) \\
= \; & n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)(1, \frac{Z_i - z_0}{h})^{\mathrm{T}} \Psi_\theta\{\cdot, \theta_0(z_0) + \theta_0'(z_0)(Z_i - z_0)\} \\
= \; & n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)(1, \frac{Z_i - z_0}{h})^{\mathrm{T}} \Psi_\theta\{\cdot, \theta_0(Z_i) - \frac{1}{2}\theta_0''(z_0)(Z_i - z_0)^2\} + o_p(n^{-1/2}) \\
= \; & n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)(1, \frac{Z_i - z_0}{h})^{\mathrm{T}} \\
& \times [\Psi_\theta\{\cdot, \theta_0(Z_i)\} - \frac{1}{2}\theta_0''(z_0)(Z_i - z_0)^2 \Psi_{\theta\theta}\{\cdot, \theta_0(Z_i)\}] + o_p(n^{-1/2}) \\
= \; & n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)(1, \frac{Z_i - z_0}{h})^{\mathrm{T}} \Psi_\theta\{\cdot, \theta_0(Z_i)\} \\
& -\frac{h^2}{2}\theta_0''(z_0) f_z(z_0) E\{\Psi_{\theta\theta}(\bullet)|Z = z_0\}(1, 0)^{\mathrm{T}} + o_p(n^{-1/2}).
\end{aligned}
$$

Plug in (A.1) and solve only for the element $(\widehat{\alpha}_0 - \alpha_0)$, we have

$$
\begin{aligned}
\widehat{\theta}(z_0, \kappa_0) - \theta_0(z_0) &= \widehat{\alpha}_0 - \alpha_0 \\
&= \frac{h^2}{2}\theta_0''(z_0) - \frac{n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)\Psi_\theta(\bullet_i)}{f_z(z_0) E\{\Psi_{\theta\theta}(\bullet)|Z = z_0\}} + o_p(n^{-1/2}).
\end{aligned}
$$

Hence Theorem 2.4.3 follows.

A.3 Sketch Proof of Theorem 2.4.2

To prove Theorem 2.4.2 we apply the result in Appendix A.2 and Lemma 2.4.1 to calculate an expansion of $\widehat{\kappa}$. Define

$$
\begin{aligned}
\Psi^*_{\kappa\kappa}\{Y, \Delta(\beta), \widehat{\theta}(Z, \kappa), \kappa\} &= \frac{d}{d\kappa}\Psi_\kappa\{Y, \Delta(\beta), \widehat{\theta}(Z, \kappa), \kappa\} \\
&= \Psi_{\kappa\Delta}(\cdot)\frac{\partial \Delta}{\partial \kappa} + \Psi_{\kappa\kappa}(\cdot) + \Psi_{\kappa\theta}(\cdot)\frac{\partial \widehat{\theta}(\cdot)}{\partial \kappa}.
\end{aligned}
$$

Apply Taylor expansion to (2.7) to get

$$\begin{aligned} 0 &= n^{-1/2}\sum_{i=1}^{n}\Psi_{\kappa}\{Y_i,\Delta_i(\widehat{\beta}),\widehat{\theta}(Z_i,\widehat{\kappa}),\widehat{\kappa}\} \\ &= n^{-1/2}\sum_{i=1}^{n}\Psi_{\kappa}\{Y_i,\Delta_i(\beta_0),\widehat{\theta}(Z_i,\kappa_0),\kappa_0\} \\ &\quad +n^{-1}\sum_{i=1}^{n}\Psi^{*}_{\kappa\kappa}\{Y_i,\Delta_i(\beta_0),\widehat{\theta}(Z_i,\kappa_0),\kappa_0\}n^{1/2}(\widehat{\kappa}-\kappa_0)+o_p(1). \qquad \text{(A.2)} \end{aligned}$$

Hence, by Lemma 2.4.1 we have

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\Psi^{*}_{\kappa\kappa}\{Y_i,\Delta_i(\beta_0),\widehat{\theta}(Z_i,\kappa_0),\kappa_0\} = E[\Psi^{*}_{\kappa\kappa}(\bullet)] = \mathcal{F}$$

So we plug $\mathcal{F}$ in (A.2) and apply Taylor expansion again

$$\begin{aligned} &-\mathcal{F}n^{1/2}(\widehat{\kappa}-\kappa_0) \\ &\quad = n^{-1/2}\sum_{i=1}^{n}\Psi_{\kappa}(\bullet_i)+n^{-1/2}\sum_{i=1}^{n}\Psi_{\kappa\theta}(\bullet_i)\{\widehat{\theta}(Z_i,\kappa_0)-\theta(Z_i)\}+o_p(1). \end{aligned}$$

Therefore, by Theorem 2.4.3 we have that

$$\begin{aligned} &-\mathcal{F}n^{1/2}(\widehat{\kappa}-\kappa_0) \\ &\quad = n^{-1/2}\sum_{i=1}^{n}\Psi_{\kappa}(\bullet_i)+(n^{1/2}h^2/2)n^{-1}\sum_{i=1}^{n}\Psi_{\kappa\theta}(\bullet_i)\theta_0''(Z_i)+B_n+o_p(1). \end{aligned}$$

The second term in the right hand side goes to zero if we assume $nh^4 \to 0$. By the symmetry of kernel $K_h(\cdot)$,

$$\begin{aligned} B_n &= -n^{-1/2}\sum_{i=1}^{n}\Psi_{\kappa\theta}(\bullet_i)\frac{n^{-1}\sum_{j=1}^{n}K_h(Z_j-Z_i)\Psi_{\theta}(\bullet_j)}{f_z(Z_i)E\{\Psi_{\theta\theta}(\bullet)|Z=Z_i\}} \\ &= -n^{-1/2}\sum_{j=1}^{n}\Psi_{\theta}(\bullet_j)n^{-1}\sum_{i=1}^{n}\frac{K_h(Z_i-Z_j)\Psi_{\kappa\theta}(\bullet_i)}{f_z(Z_i)E\{\Psi_{\theta\theta}(\bullet)|Z=Z_i\}} \\ &= -n^{-1/2}\sum_{j=1}^{n}\Psi_{\theta}(\bullet_j)\frac{E\{\Psi_{\kappa\theta}(\bullet)|Z=Z_j\}}{E\{\Psi_{\theta\theta}(\bullet)|Z=Z_j\}}+o_p(1) \\ &= -n^{-1/2}\sum_{j=1}^{n}\Psi_{\theta}(\bullet_j)\mathcal{U}(Z_j)+o_p(1). \end{aligned}$$

Hence Theorem 2.4.2 follows.

APPENDIX B

PROOFS IN CHAPTER III

B.1 Regression Coefficients for Genotype Effect Model

Multiplying both sides of the "genotype effect model" (3.1) by $1_{(G_{Aij}=A_gA_h)}$ and taking expectation lead to

$$\begin{aligned} E(y_{ij}1_{(G_{Aij}=A_gA_h)}) &= w_{ij}\gamma E[1_{(G_{Aij}=A_gA_h)}] + E[1_{(G_{Aij}=A_gA_h)}]\beta_{gh} \\ &= \begin{cases} (1-\varepsilon_A)[w_{ij}\gamma+\beta_{gg}]P_{A_g}^2 & \text{if } g=h \\ (1-\varepsilon_A)[w_{ij}\gamma+\beta_{gh}]\cdot 2P_{A_g}P_{A_h} & \text{if } g\neq h \end{cases}. \end{aligned} \quad \text{(C.1)}$$

Let G_{Qij} be genotype of the j-th individual of the i-th family at the trait locus Q. A true random effect model describing the trait value is $y_{ij} = w_{ij}\gamma + g_{ij} + H_{ij} + e_{ij}$, where

$$g_{ij} = \begin{cases} a & G_{Qij} = Q_1Q_1 \\ d & G_{Qij} = Q_1Q_2 \\ -a & G_{Qij} = Q_2Q_2 \end{cases}.$$

Since the missing mechanism is MCAR, we have

$$\begin{aligned} P(G_{Qij}=Q_1Q_1, G_{Aij}=A_gA_g|G_{Aij}\neq ?) &= [P(Q_1A_g)]^2, \\ P(G_{Qij}=Q_1Q_2, G_{Aij}=A_gA_g|G_{Aij}\neq ?) &= 2P(Q_1A_g)P(Q_2A_g), \\ P(G_{Qij}=Q_2Q_2, G_{Aij}=A_gA_g|G_{Aij}\neq ?) &= [P(Q_2A_g)]^2. \end{aligned}$$

Utilizing relations $P(Q_1A_g) = D_{A_gQ} + P_{A_g}q_1$ and $P(Q_2A_g) = -D_{A_gQ} + P_{A_g}q_2$ gives

$$\begin{aligned}
E(y_{ij}1_{(G_{Aij}=A_gA_g)}) &= w_{ij}\gamma E[1_{(G_{Aij}=A_gA_g)}] + E[g_{ij}1_{(G_{Aij}=A_gA_g)}] \\
&= w_{ij}\gamma P(A_gA_g|G_{Aij} \neq ?)P(G_{Aij} \neq ?) \\
&\quad + E[g_{ij}1_{(G_{Aij}=A_gA_g)}|G_{Aij} \neq ?]P(G_{Aij} \neq ?) \\
&= (1-\varepsilon_A)\Big[w_{ij}\gamma P_{A_g}^2 + a[P(Q_1A_g)]^2 \\
&\quad + d \cdot 2P(Q_1A_g)P(Q_2A_g) - a[P(Q_2A_g]^2\Big] \\
&= (1-\varepsilon_A)\Big[w_{ij}\gamma P_{A_g}^2 + \mu P_{A_g}^2 \\
&\quad + 2D_{A_gQ}\alpha_Q P_{A_g} - \delta_Q D_{A_gQ}^2\Big]. \qquad \text{(C.2)}
\end{aligned}$$

Equating equations (C.1) and (C.2), we show the equation (3.5) when $g = h$. Now assume that $g \neq h$. Since the missing mechanism is MCAR, we have

$$\begin{aligned}
P(G_{Qij} = Q_1Q_1, G_{Aij} = A_gA_h|G_{Aij} \neq ?) &= 2P(Q_1A_g)P(Q_1A_h); \\
P(G_{Qij} = Q_1Q_2, G_{Aij} = A_gA_h|G_{Aij} \neq ?) &= 2P(Q_1A_g)P(Q_2A_h) \\
&\quad + 2P(Q_1A_h)P(Q_2A_g); \\
P(G_{Qij} = Q_2Q_2, G_{Aij} = A_gA_h|G_{Aij} \neq ?) &= 2P(Q_2A_g)P(Q_2A_h).
\end{aligned}$$

Utilizing relations $P(Q_1A_g) = D_{A_gQ} + P_{A_g}q_1$, $P(Q_2A_g) = -D_{A_gQ} + P_{A_g}q_2$, $P(Q_1A_h) = D_{A_hQ} + P_{A_h}q_1$, $P(Q_2A_h) = -D_{A_hQ} + P_{A_h}q_2$ gives

$$\begin{aligned}
E(y_{ij}1_{(G_{Aij}=A_gA_h)}) &= w_{ij}\gamma E[1_{(G_{Aij}=A_gA_h)}] + E[g1_{(G_{Aij}=A_gA_h)}] \\
&= w_{ij}\gamma P(A_gA_h|G_{Aij} \neq ?)P(G_{Aij} \neq ?) \\
&\quad + E[g1_{(G_{Aij}=A_gA_h)}|G_{Aij} \neq ?]P(G_{Aij} \neq ?) \\
&= (1-\varepsilon_A)\Big[w_{ij}\gamma \cdot 2P_{A_g}P_{A_h} \\
&\quad + 2a\Big\{P(Q_1A_g)P(Q_1A_h) - P(Q_2A_g)P(Q_2A_h)\Big\} \\
&\quad + d\Big\{2P(Q_1A_g)P(Q_2A_h) + 2P(Q_2A_g)P(Q_1A_h)\Big\}\Big] \\
&= (1-\varepsilon_A)\Big[2P_{A_g}P_{A_h}w_{ij}\gamma + 2P_{A_g}P_{A_h}\mu \\
&\quad + 2\alpha_Q\Big(D_{A_gQ}P_{A_h} + D_{A_hQ}P_{A_g}\Big) - 2\delta_Q D_{A_gQ}D_{A_hQ}\Big]. \quad \text{(C.3)}
\end{aligned}$$

Equating equations (C.1) and (C.3), we show the equation (3.5) when $g \neq h$.

B.2 Regression Coefficients for Additive Effect Model

In relations (C.1), replacing β_{gh} by $\alpha_g + \alpha_h$ and taking summation lead to

$$\begin{aligned}
E(y_{ij}1_{(G_{Aij}\neq ?)}) &= \sum_{1\leq g\leq h\leq m} E(y_{ij}1_{(G_{Aij}=A_gA_h)}) \\
&= (1-\varepsilon_A)\sum_{g=1}^{m}\sum_{h=1}^{m}\Big(w_{ij}\gamma + \alpha_g + \alpha_h\Big)P_{A_g}P_{A_h} \\
&= (1-\varepsilon_A)\Big(w_{ij}\gamma + 2\sum_{g=1}^{m}\alpha_g P_{A_g}\Big).
\end{aligned}$$

Since the missing mechanism is MCAR, one has $E(y_{ij}1_{(G_{Aij}\neq ?)}) = E(y_{ij}|G_{Aij} \neq ?)(1-\varepsilon_A) = (1-\varepsilon_A)Ey_{ij} = (1-\varepsilon_A)\Big(w_{ij}\gamma + \mu\Big)$. Thus, $\sum_{g=1}^{m}\alpha_g P_{A_g} = \mu/2$.

Again, replacing β_{gh} by $\alpha_g + \alpha_h$ in relations (C.1) and taking summation with

respect to h lead to

$$\begin{aligned} E\Big[y_{ij}1_{(G_{Aij}=A_gA_g)} + \frac{1}{2}\sum_{h\neq g} y_{ij}1_{(G_{Aij}=A_gA_h)}\Big] &= (1-\varepsilon_A)\sum_{h=1}^{m}\Big(w_{ij}\gamma + \alpha_g + \alpha_h\Big)P_{A_g}P_{A_h} \\ &= (1-\varepsilon_A)P_{A_g}\Big(w_{ij}\gamma + \alpha_g + \sum_{h=1}^{m}\alpha_h P_{A_h}\Big) \\ &= (1-\varepsilon_A)P_{A_g}\Big(w_{ij}\gamma + \alpha_g + \mu/2\Big). \quad \text{(C.4)} \end{aligned}$$

Notice $\sum_{g=1}^{m} D_{A_gQ} = 0$. Taking summation of relations (C.2) and (C.3) leads to

$$\begin{aligned} & E\Big[y_{ij}1_{(G_{Aij}=A_gA_g)} + \frac{1}{2}\sum_{h\neq g} y_{ij}1_{(G_{Aij}=A_gA_h)}\Big] \\ = \;& (1-\varepsilon_A)P_{A_g}\Big[w_{ij}\gamma + \mu + D_{A_gQ}\alpha_Q/P_{A_g}\Big]. \quad \text{(C.5)} \end{aligned}$$

Equating the right-hand terms of relations (C.4) and (C.5) leads to (3.6).

B.3 Noncentrality Parameter for Genotype Effect Model

Assume that there are no covariates, and the dataset is a population sample. Then the model matrix of "genotype effect model" (3.1) is

$$X_i = X_{Ai1}^{\mathrm{T}} = (x_{Ai1}^{(11)}, \cdots, x_{Ai1}^{(mm)}, x_{Ai1}^{(12)}, \cdots, x_{Ai1}^{(1m)}, \cdots, x_{Ai1}^{(m-1,m)}),$$

for $i = 1, \cdots, N$. To show noncentrality parameter approximation (3.7), we first notice the following relation

$$E[X_1X_1^{\mathrm{T}}] = (1-\varepsilon_A)\mathrm{diag}(P_{A_1}^2, v^{\mathrm{T}}) + \varepsilon_A \begin{pmatrix} P_{A_1}^2 \\ v \end{pmatrix}(P_{A_1}^2, v^{\mathrm{T}}), \quad \text{(C.6)}$$

where v is a column vector given by

$$v^{\mathrm{T}} = \Big(P_{A_2}^2, \cdots, P_{A_m}^2, 2P_{A_1}P_{A_2}, \cdots, 2P_{A_1}P_{A_m}, \cdots, 2P_{A_{m-1}}P_{A_m}\Big).$$

In addition, $\mathrm{diag}(P_{A_1}^2, v^{\mathrm{T}})$ is a diagonal matrix, whose elements on the diagonal are given by the elements of $(P_{A_1}^2, v^{\mathrm{T}})$. We may verify (C.6) by

$$\begin{aligned} E[(x_{A11}^{(gh)})^2] &= E\{1_{(G_{A11}=A_gA_h)}\} + P(A_gA_h)^2 E\{1_{(G_{A11}=?)}\} \\ &= P(A_gA_h)(1-\varepsilon_A) + P(A_gA_h)^2\varepsilon_A, \end{aligned}$$

and for $(g,h) \neq (k,l)$,

$$E[x_{A11}^{(gh)} x_{A11}^{(kl)}] = P(A_gA_h)P(A_kA_l)E\{1_{(G_{A11}=?)}\} = P(A_gA_h)P(A_kA_l)\varepsilon_A.$$

Let us denote

$$u = \Big(P_{A_2}^{-2}, \cdots, P_{A_m}^{-2}, [2P_{A_1}P_{A_2}]^{-2}, \cdots, [2P_{A_1}P_{A_m}]^{-2}, \cdots, [2P_{A_{m-1}}P_{A_m}]^{-2}\Big).$$

Applying the law of large number and a fact of inverse matrix that $(M+ab^{\mathrm{T}})^{-1} = M^{-1} - (M^{-1}a)(b^{\mathrm{T}}M^{-1})/(1+b^{\mathrm{T}}M^{-1}a)$, we can calculate the following approximation

$$\begin{aligned} T(X^{\mathrm{T}}X)^{-1}T^{\mathrm{T}} &\approx T\Big[NE\Big(X_1X_1^{\mathrm{T}}\Big)\Big]^{-1}T^{\mathrm{T}} \\ &= N^{-1}\cdot T\Big[(1-\varepsilon_A)\mathrm{diag}(P_{A_1}^2, v^{\mathrm{T}}) + \varepsilon_A \begin{pmatrix} P_{A_1}^2 \\ v \end{pmatrix} (P_{A_1}^2, v^{\mathrm{T}})\Big]^{-1} T \\ &= [(1-\varepsilon_A)N]^{-1}\cdot T\Big[\mathrm{diag}(P_{A_1}^{-2}, u^{\mathrm{T}}) - \varepsilon_A \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (1,\cdots,1)\Big]T \\ &= [(1-\varepsilon_A)N]^{-1}\cdot T\mathrm{diag}(P_{A_1}^{-2}, u^{\mathrm{T}})T. \end{aligned}$$

Utilizing above relation, we may show noncentrality parameter approximation (3.7) in the same way as Appendix III, Fan et al. (2006).

B.4 Noncentrality Parameter for Additive Effect Model

Assume that there are no covariates, and the dataset is a population sample. Then the model matrix of "additive effect model" (3.3) is $X_i = Z_{Ai1}^{\mathrm{T}} = (x_{Ai1}^{(1)}, \cdots, x_{Ai1}^{(m)})$,

$i = 1, \cdots, N$. To show noncentrality parameter approximation (3.8), we first notice the following relation

$$E[Z_{A11}Z_{A11}^{\mathrm{T}}] = 2(1-\varepsilon_A)\Big[\mathrm{diag}(P_{A_1},\cdots,P_{A_m}) + \begin{pmatrix} P_{A_1} \\ \vdots \\ P_{A_m} \end{pmatrix}(P_{A_1},\cdots,P_{A_m})\Big]$$
$$+4\varepsilon_A \begin{pmatrix} P_{A_1} \\ \vdots \\ P_{A_m} \end{pmatrix}(P_{A_1},\cdots,P_{A_m}),$$

which can be verified by

$$\begin{aligned} E[(x_{A11}^{(g)})^2] &= 4E\{1_{(G_{A11}=A_gA_g)}\} + \sum_{h\neq g} E\{1_{(G_{A11}=A_gA_h)}\} + 4P_{A_g}^2 E1_{(G_{A11}=?)} \\ &= 2(1-\varepsilon_A)P_{A_g}[1+P_{A_g}] + 4P_{A_g}^2\varepsilon_A, \end{aligned}$$

and for $h \neq g$,

$$E[x_{A11}^{(g)}x_{A11}^{(h)}] = (1-\varepsilon_A)\cdot 2P_{A_g}P_{A_h} + 4P_{A_g}P_{A_h}\varepsilon_A.$$

Let $X = (Z_{A11},\cdots,Z_{AN1})^{\mathrm{T}}$. Applying the law of large number and a fact of inverse matrix $(M+ab^{\mathrm{T}})^{-1} = M^{-1} - (M^{-1}a)(b^{\mathrm{T}}M^{-1})/(1+b^{\mathrm{T}}M^{-1}a)$, we can calculate the

following approximation

$$\begin{aligned}
K(X^{\mathrm{T}}X)^{-1}K^{\mathrm{T}} &\approx K\Big[NE\Big(Z_{A11}Z_{A11}^{\mathrm{T}}\Big)\Big]^{-1}K^{\mathrm{T}} \\
&= N^{-1}\cdot K\Big[2(1-\varepsilon_A)\mathrm{diag}(P_{A_1},\cdots,P_{A_m}) \\
&\quad +2(1+\varepsilon_A)\begin{pmatrix} P_{A_1} \\ \vdots \\ P_{A_M} \end{pmatrix}(P_{A_1},\cdots,P_{A_m})\Big]^{-1}K \\
&= [2(1-\varepsilon_A)N]^{-1}\cdot K\Big[\mathrm{diag}(P_{A_1}^{-1},\cdots,P_{A_m}^{-1}) \\
&\quad -(1+\varepsilon_A)\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}(1,\cdots,1)/2\Big]K \\
&= [2(1-\varepsilon_A)N]^{-1}\cdot K\mathrm{diag}(P_{A_1}^{-1},\cdots,P_{A_m}^{-1})K.
\end{aligned}$$

Utilizing above relation, we may show noncentrality parameter approximation (3.8) in the same way as Appendix IV, Fan et al. (2006).

B.5 Regression Coefficients for Two-Marker Models

For $g = 1, 2, \cdots, m, k = 1, \cdots, n$, let us denote $D_{A_gB_k} = P(A_gB_k) - P_{A_g}P_{B_k}$, which are measures of LD between markers A and B. Here, $P(A_gB_k)$ is frequency of haplotype A_gB_k. It can be shown that for $g \neq h, k \neq l, h \neq h', l \neq l', (g,h) \neq$

$(g', h'), (k, l) \neq (k', l')$,

$$
\begin{aligned}
&E\{x_{Aij}^{(g)}\} = 2P_{A_g}; \quad E\{x_{Bij}^{(k)}\} = 2P_{B_k}; \quad E\{z_{Aij}^{(gh)}\} = 0; \quad E\{z_{Bij}^{(kl)}\} = 0;\\
&E(x_{Aij}^{(g)})^2 = (1-\varepsilon_A)(2P_{A_g}^2 + 2P_{A_g}) + 4P_{A_g}^2\varepsilon_A;\\
&E[x_{Aij}^{(g)}x_{Aij}^{(h)}] = 2P_{A_g}P_{A_h}(1-\varepsilon_A) + 4P_{A_g}P_{A_h}\varepsilon_A;\\
&E(x_{Bij}^{(k)})^2 = (1-\varepsilon_B)(2P_{B_k}^2 + 2P_{B_k}) + 4P_{B_k}^2\varepsilon_B;\\
&E[x_{Bij}^{(k)}x_{Bij}^{(l)}] = 2P_{B_k}P_{B_l}(1-\varepsilon_B) + 4P_{B_k}P_{B_l}\varepsilon_B;\\
&E(z_{Aij}^{(gh)})^2 = (1-\varepsilon_A)P_{A_g}^2P_{A_h}^2[P_{A_g} + P_{A_h}]^2;\\
&E(z_{Bij}^{(kl)})^2 = (1-\varepsilon_B)P_{B_k}^2P_{B_l}^2[P_{B_k} + P_{B_l}]^2;\\
&E[x_{Aij}^{(g)}z_{Aij}^{(gh)}] = E[x_{Aij}^{(g)}z_{Aij}^{(hh')}] = E[x_{Bij}^{(k)}z_{Bij}^{(kl)}] = E[x_{Bij}^{(k)}z_{Bij}^{(ll')}] = 0\\
&E[x_{Aij}^{(g)}z_{Bij}^{(kl)}] = E[x_{Bij}^{(k)}z_{Aij}^{(gh)}] = E[z_{Aij}^{(gh)}z_{Aij}^{(g'h')}] = E[z_{Bij}^{(kl)}z_{Bij}^{(k'l')}] = 0;\\
&E[x_{Aij}^{(g)}x_{Bij}^{(k)}] = 2D_{A_gB_k}(1-\varepsilon_A)(1-\varepsilon_B) + 4P_{A_g}P_{B_k}, E[z_{Aij}^{(gh)}z_{Aij}^{(gh')}]\\
&\qquad\qquad = (P_{A_g}P_{A_h}P_{A_h'})^2(1-\varepsilon_A);\\
&E[z_{Bij}^{(kl)}z_{Bij}^{(kl')}] = (P_{B_k}P_{B_l}P_{B_l'})^2(1-\varepsilon_B);\\
&E[z_{Aij}^{(gh)}z_{Bij}^{(kl)}] = \Big[P_{A_h}\Big(P_{B_l}D_{A_gB_k} - P_{B_k}D_{A_gB_l}\Big) - P_{A_g}\Big(P_{B_l}D_{A_hB_k} - P_{B_k}D_{A_hB_l}\Big)\Big]^2\\
&\qquad\qquad \times(1-\varepsilon_A)(1-\varepsilon_B);\\
&E[y_{ij}x_{Aij}^{(g)}] = 2P_{A_g}(w_{ij}\gamma + \mu) + 2\alpha_QD_{A_gQ}(1-\varepsilon_A);\\
&E[y_{ij}x_{Bij}^{(k)}] = 2P_{B_k}(w_{ij}\gamma + \mu) + 2\alpha_QD_{B_kQ}(1-\varepsilon_B);\\
&E[y_{ij}z_{Aij}^{(gh)}] = \delta_Q\Big[P_{A_g}D_{A_hQ} - P_{A_h}D_{A_gQ}\Big]^2(1-\varepsilon_A);\\
&E[y_{ij}z_{Bij}^{(kl)}] = \delta_Q\Big[P_{B_k}D_{B_lQ} - P_{B_l}D_{B_kQ}\Big]^2(1-\varepsilon_B). \qquad \text{(C.7)}
\end{aligned}
$$

The quantities in (C.7) imply that the elements of V_A are given by

$$
\begin{aligned}
\mathrm{cov}(x_{Aij}^{(g)}, x_{Aij}^{(h)}) &= -2P_{A_g}P_{A_h}(1-\varepsilon_A), \mathrm{var}(x_{Aij}^{(g)}) = 2P_{A_g}(1-P_{A_g})(1-\varepsilon_A),\\
\mathrm{cov}(x_{Aij}^{(g)}, x_{Bij}^{(k)}) &= 2D_{A_gB_k}(1-\varepsilon_A)(1-\varepsilon_B),\\
\mathrm{cov}(x_{Bij}^{(k)}, x_{Bij}^{(l)}) &= -2P_{B_k}P_{B_l}(1-\varepsilon_B), \mathrm{var}(x_{Bij}^{(k)}) = 2P_{B_k}(1-P_{B_k})(1-\varepsilon_B).
\end{aligned}
$$

Since $EZ_{A\cup B}^{(ij)}$ is a vector of zero's by the quantities in (C.7), it can be shown that $V_D = \text{cov}(Z_{A\cup B}^{(ij)}, Z_{A\cup B}^{(ij)}) = E(Z_{A\cup B}^{(ij)}(Z_{A\cup B}^{(ij)})^\tau)$. Moreover, the quantities in (C.7) imply that the covariance matrix $\text{cov}(X_{A\cup B}^{(ij)}, Z_{A\cup B}^{(ij)})$ is a zero matrix. In addition, the covariances between the trait value y_{ij} and variables $x_{Aij}^{(g)}, x_{Bij}^{(k)}, z_{Aij}^{(gh)}$ and $z_{Bij}^{(kl)}$ are

$$\text{cov}(y_{ij}, x_{Aij}^{(g)}) = 2\alpha_Q(1-\varepsilon_A)D_{A_gQ}, \qquad \text{cov}(y_{ij}, x_{Bij}^{(k)}) = 2\alpha_Q(1-\varepsilon_B)D_{B_kQ},$$
$$\text{cov}(y_{ij}, z_{Aij}^{(gh)}) = E[y_{ij}z_{Aij}^{(gh)}], \qquad \text{cov}(y_{ij}, z_{Bij}^{(kl)}) = E[y_{ij}z_{Bij}^{(kl)}].$$

Taking variance-covariance between y_{ij} and $x_{Aij}^{(g)}, x_{Bij}^{(k)}, z_{Aij}^{(gh)}, z_{Bij}^{(kl)}$ based on relation (3.12), we may get the regression coefficients (3.13) of models (3.10) and (3.12).

B.6 Noncentrality Parameter for Nuclear Family

Notice $\Sigma_i^{-1} = \frac{1}{\sigma^2}(\gamma_{hj})_{(s+2)\times(s+2)}$. Let X_i be the model matrix of family $i = 1, 2, \cdots, I$. Then $X_i = (\mathbf{1_{s+2}}, X_{Ai}, X_{Bi}, Z_{Ai}, Z_{Bi})$, where $\mathbf{1_{s+2}}$ is a $(s+2)$-dimension column vector of 1's and

$$X_{Ai} = \begin{bmatrix} x_{Ai1}^{(1)} & \cdots & x_{Ai1}^{(m-1)} \\ \vdots & \cdots & \vdots \\ x_{Ai,s+2}^{(1)} & \cdots & x_{Ai,s+2}^{(m-1)} \end{bmatrix}; \quad X_{Bi} = \begin{bmatrix} x_{Bi1}^{(1)} & \cdots & x_{Bi1}^{(n-1)} \\ \vdots & \cdots & \vdots \\ x_{Bi,s+2}^{(1)} & \cdots & x_{Bi,s+2}^{(n-1)} \end{bmatrix};$$
$$Z_{Ai} = \begin{bmatrix} z_{Ai1}^{(12)} & \cdots & z_{Ai1}^{(m-1,m)} \\ \vdots & \cdots & \vdots \\ z_{Ai,s+2}^{(12)} & \cdots & z_{Ai,s+2}^{(m-1,m)} \end{bmatrix}; \quad Z_{Bi} = \begin{bmatrix} z_{Bi1}^{(12)} & \cdots & z_{Bi1}^{(n-1,n)} \\ \vdots & \cdots & \vdots \\ z_{Bi,s+2}^{(12)} & \cdots & z_{Bi,s+2}^{(n-1,n)} \end{bmatrix}.$$

Denote $\gamma = \sum_{k=1}^{s+2}\sum_{l=1}^{s+2}\gamma_{kl}$. Applying law of large number leads to an approxi-

mation as

$$\sum_{i=1}^{I} X_i^{\mathrm{T}}\Sigma_i^{-1}X_i/I \approx \tag{C.8}$$
$$\frac{1}{\sigma^2}\begin{bmatrix} \gamma & \gamma[E(X_{A\cup B}^{(11)})]^{\mathrm{T}} & O_1 \\ \gamma E(X_{A\cup B}^{(11)}) & \sum_{k=1}^{s+2}\gamma_{kk}V_A + bV_{A2} + \gamma E(X_{A\cup B}^{(11)})[E(X_{A\cup B}^{(11)})]^{\mathrm{T}} & O_2 \\ O_3 & O_4 & \sum_{k=1}^{s+2}\gamma_{kk}V_D + \sum_{k=3}^{s+2}\sum_{l=k+1}^{s+2}\gamma_{kl}V_{D2}/2 \end{bmatrix},$$

where O_i, $i = 1, 2, 3, 4$ are zero vectors or matrices, and $E(X_{A\cup B}^{(11)}) = (2P_{A_1}, \cdots, 2P_{A_{m-1}}, 2P_{B_1}, \cdots, 2P_{B_{n-1}})^{\mathrm{T}}$.

Let

$$S = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

be the test matrix corresponding to hypothesis H_{ABad0}, and

$$\phi = (\alpha, \alpha_{A1}, \cdots, \alpha_{A(m-1)}, \alpha_{B1}, \cdots, \alpha_{B(m-1)}, \delta_{A12}, \cdots, \delta_{A(m-1)m}, \delta_{B12}, \cdots, \delta_{B(n-1)n})^{\mathrm{T}}$$

be the column vector of regression coefficient of "genotype effect model" (3.12). Utilizing regression coefficients (3.13), we may show (3.15) by plugging approximation (C.8) into $\lambda_{ABad} = (S\phi)^{\mathrm{T}}[S(\sum_{i=1}^{I} X_i^{\mathrm{T}}\Sigma_i^{-1}X_i)^{-1}S^{\mathrm{T}}]^{-1}(S\phi)$. One may want to notice that we may use Theorem 8.5.11, Harville (1997), to calculate the inverse of the right-hand matrix of (C.8).

B.7 Constants in Noncentrality Parameter Approximation for Pedigree Data

For pedigrees in graph A of Figure 3, the constants b_1 and b_2 of $\lambda_{AB,ad}$ in (3.16) are given by

$$\begin{aligned} b_1 &= [\gamma_{15} + (\gamma_{17} + \cdots + \gamma_{1,11})/2] + [\gamma_{25} + (\gamma_{27} + \cdots + \gamma_{2,11})/2] \\ &\quad + [\gamma_{36} + (\gamma_{37} + \cdots + \gamma_{3,11})/2] + [\gamma_{46} + (\gamma_{47} + \cdots + \gamma_{4,11})/2] \\ &\quad + (\gamma_{57} + \cdots + \gamma_{5,11}) + (\gamma_{67} + \cdots + \gamma_{6,11}) + \sum_{k=7}^{11} \sum_{l=k+1}^{11} \gamma_{kl}, \\ b_2 &= \sum_{k=7}^{11} \sum_{l=k+1}^{11} \gamma_{kl}/2. \end{aligned}$$

For pedigrees in graph B of Figure 3, constants b_1 and b_2 are given by

$$\begin{aligned} b_1 &= \gamma_{1,12} + [\gamma_{2,12} + (\gamma_{2,13} + \cdots + \gamma_{2,16})/2] + [\gamma_{3,12} + \cdots + \gamma_{3,16}]/2 \\ &\quad + [\gamma_{4,12} + \cdots + \gamma_{4,16}]/2 + [\gamma_{5,12}/2 + (\gamma_{5,13} + \cdots + \gamma_{5,16})] \\ &\quad + [(\gamma_{6,13} + \cdots + \gamma_{6,16}) + (\gamma_{6,17} + \gamma_{6,18})/2] + [\gamma_{7,13} + \cdots + \gamma_{7,18}]/2 \\ &\quad + [(\gamma_{8,13} + \cdots + \gamma_{8,16})/2 + (\gamma_{8,17} + \gamma_{8,18})] + (\gamma_{9,17} + \gamma_{9,18}) + (\gamma_{10,17} + \gamma_{10,18})/2 \\ &\quad + (\gamma_{11,17} + \gamma_{11,18})/2 + (\gamma_{12,13} + \cdots + \gamma_{12,16})/4 + (\gamma_{13,14} + \gamma_{13,15} + \gamma_{13,16}) \\ &\quad + (\gamma_{14,15} + \gamma_{14,16}) + \gamma_{15,16} + [\gamma_{13,17} + \cdots + \gamma_{16,17}]/4 \\ &\quad + [\gamma_{13,18} + \cdots + \gamma_{16,18}]/4 + \gamma_{17,18}, \\ b_2 &= [(\gamma_{13,14} + \gamma_{13,15} + \gamma_{13,16}) + (\gamma_{14,15} + \gamma_{14,16}) + \gamma_{15,16}]/2 + \gamma_{17,18}/2. \end{aligned}$$

VITA

Lian Liu was born in Tianjin, China. He graduated from Yaohua High School in Tianjin, China in 1998. He received a Bachelor of Science degree in probability and statistics from Peking University in Beijing, China in July 2002, and a Master of Science degree in statistics from Texas A&M University in College Station, Texas, under the direction of Dr. Daren B. H. Cline in May 2004. He continued his studies under the direction of Dr. Raymond J. Carroll, and received a Doctor of Philosophy degree in statistics from Texas A&M University in August 2007. His permanent address is:

3-3-401 Runtai Garden, Tiantai Road, Hebei District

Tianjin, China, 300230.