Linkage Analysis of a Quantitative Trait:

Suggested methods for sibling pairs

with at least one member having an extreme trait value

A Dissertation Presented

by

So Youn Shin

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

December 2009

Stony Brook University

The Graduate School

So Youn Shin

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Nancy R. Mendell – Dissertation Advisor
Professor, Applied Mathematics and Statistics

Stephen Finch – Chairperson of Defense
Professor, Applied Mathematics and Statistics

Hongshik Ahn
Professor, Applied Mathematics and Statistics

Deborah L. Levy
Director, Psychology Research Laboratory, McLean Hospital

This Dissertation is accepted by the Graduate School.

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Linkage Analysis of a Quantitative Trait:

Suggested method for sibling pairs

with at least one member having an extreme trait value

by

So Youn Shin

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2009

Most model-free methods for family based genetic linkage analysis require unbiased trait parameter estimates. Using simulation studies, Cuenco et al. (2003) and Bhattacharjee et al. (2008) examined the sensitivity of existing model-free methods to the assumed trait parameter values. They concluded that the misspecification of the trait parameter values could significantly lower the power of linkage analysis methods.

Because trait parameter values of interest are the mean, variance, and correlation, they can be estimated without bias in random samples. However, clinical and epidemiological studies based on selected samples are more powerful, because families are usually selected for having at least one member with an extreme trait value. Parameter estimates based on such truncate samples are biased.

In this paper, we will concentrate on improving the application of existing model-free linkage analysis methods to truncate samples by considering three new approaches for estimating trait parameter values. These three approaches have not previously been applied to linkage analysis, but have been suggested for truncate selection for other purpose (Cohen, 1959; Rao et al., 1968; Mendell and Elston, 1974; Aboueissa and Stoline, 2004). We will evaluate the effect of using five different estimators (the three new approaches based on the truncated likelihood function, and two existing approaches based on the likelihood function and the conditional likelihood function) on the power of five model-free linkage analysis methods.

We note that in our studies, S & P's method seems to be more powerful than Xu and Cuenco's methods, and Xu's method seems more robust to the estimators used (even with sample moments and Shin Nu). In most cases, two of our suggested estimators give greater power (and higher LOD scores) for Xu's method than Peng's CMLEs. However, when S & P's and Cuenco's methods are used, Peng's CMLEs have greater power than our estimators. We recommend using S & P's method with Peng's CMLEs with Shin RM as an initial value for its iteration, because Peng's CMLEs are more robust to the truncation scheme than our estimators.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my dissertation advisor, Dr. Nancy Role Mendell, for her intellectual insight, professional guidance and for lavishing me with an abundance of human kindness. Without her, the journey towards my doctoral degree would have been much more difficult. I would also like to thank my wonderful committee members, Dr. Stephen Finch, Dr. Hongshik Ahn, and Dr. Deborah L. Levy. I have benefitted from their professional advice and support very much.

I would like to express my deep appreciation to my parents and siblings for their continued encouragement, constant support and for their unconditional love. I also want to acknowledge my dearest friend, Brian McGillick, and the people who have been there for me throughout the years, as my sounding board, ever ready with constant understanding, thoughtful advice and the sublime gift of their friendship: Fr. Robert Smith, Julitta Jo, Siwon Song, Heejong Sung, Hyung Yeon Gu, Ji Yeon Hong, Sang Won Kim, and Astrid & Wolfgang Wander.

Lastly, I would like to give my special thanks to Dr. Maripat Quinn and my precious friends who made my Stony Brook experience particularly memorable and enjoyably positive: Tulin Kaman, Eunah Lee, Qilong Yuan, Juhee Hong, Inyoung Kim, Kyewon Lee, Yeongmi Jeon, Hyunsun Lee, Xiaofei Fan, Ming Zhao, Chengrui Huang, Adrienne Chu, Ivan Tovkach, Hye Jin Oh, Choongseok Yoon, Renaud Gauthier, Eric & Anne Waxman, Vincent & Janet Eletto, Alex & Sophie Kaiser, Helen Park, Fr. Thomas

Choi, Fr. Benedict Bae, Sr. Helena, Sr. Mary, St. Therese, St. Vianne, St. Pio, St. Francis, St. Ameliana, Our Lady, and Him.

Chapter 1 Introduction

Genetic linkage analysis is the process of finding the approximate location of genes that determine a trait of interest, by using marker genes whose locations are already known. Linkage analysis is based on the fact that when the trait gene and marker gene are physically close to each other, two related individuals having similar trait values are likely to have the same marker genotypes. Linkage and association studies are two main strategies in gene mapping and play an important role in genetic epidemiology.

Recently, more focus has been on genome-wide association studies (GWAS) which use population-based designs, except in the case of the transmission disequilibrium test (TDT) (Laird and Lange, 2006). However, we should not underestimate the important role of linkage analysis and family-based designs. Identifying a candidate region by linkage analysis can lead to a more cost-effective association study (Clerget-Darpoux and Elston, 2007). Also, since family-based and population-based designs have different strengths and weaknesses, they should complement each other especially now that we have information on hundreds of thousands of sequenced nucleotides (Laird and Lange, 2006).

In this paper, we suggest a linkage analysis method for selected samples. In Chapter 1, we will present background on statistical genetics, model-free linkage analysis methods for quantitative trait locus, and study designs. In Chapter 2, we will show two existing approaches and suggest three new approaches for estimating trait parameter values for selected samples. In Chapter 3, we will explain how to generate simulation data sets. In Chapter 4, we will show the effects on power, using linkage analysis methods, with different trait parameter estimates. In Chapter 5, we will discuss the results of our study and future directions.

1. 1. Quantitative Trait Linkage Analysis in Truncate Samples

There are two types of statistical methods for doing linkage analysis: gene model-based methods and gene model-free methods. Gene model-based methods require knowledge of the inheritance pattern of the trait of interest. For example, they focus on understanding the genetic mechanisms of disorders, usually caused by a single allele, at a single genetic locus. On the other hand, gene model-free methods do not require knowledge of the specific genetic transmission model. For example, most gene model-free methods for a quantitative trait locus (QTL) require only information on trait values in sib pairs and the number of shared alleles Identical by Descent (IBD), at each marker of interest. Thus, gene model-free methods are based on the relation between phenotype differences and genotype differences in relatives.

In this paper, we focus on two of the most commonly used model-free linkage analysis methods for identifying markers linked QTL. The first one is regression based and the second one is likelihood based.

A regression based method was originally suggested by Haseman and Elston (1972) and has been extended by many researchers. The original Haseman and Elston method is intuitively simple and robust with respect to assumptions about the distribution of a quantitative trait. However, it has less power than likelihood based methods, especially in randomly sampled sib pairs (Feingold 2001). Many researchers including Elston et al. (2000), Xu et al. (2000) and Sham and Purcell (2001), have tried to improve the power of regression based methods. However, these latter methods require estimating, or knowing additional parameter values of the trait's distribution, i.e. mean, variance, and correlation. It is not a hard issue to estimate these trait parameter values in random samples. However, estimates of trait parameter values in selected samples are biased, making recent modifications of regression-based methods less robust than the original Haseman and Elston's method.

Likelihood based methods include a variance components method created by Amos (1994) and relatively new, score test based methods, suggested by several researchers (Tang and Siegmund, 2001; Cuenco et al, 2003). Most score test statistics are similar to one another. They have been modifiecd to achieve robustness (Cuenco et al, 2003). The advantage of the score test statistic is that the computation is done only under the null hypothesis of no linkage, not under the alternative. In contrast, the traditional variance components method requires computations under both hypotheses. Since the likelihood based test statistics were developed based on the assumption that the trait value distribution is normal, these methods are sensitive to violations in the assumption of normality and to misspecification of the parameter values of the trait distribution. In other words, if the trait is not normally distributed or the sample is not randomly ascertained, power drops dramatically.

Thus, most model-free methods, including both newly suggested regression based methods and likelihood based methods, require unbiased trait parameter value estimates. Using simulation studies, Cuenco et al. (2003) and Bhattacharjee et al. (2008) examined the sensitivity of existing model-free methods to the assumed trait parameter values. They concluded that misspecification of the trait parameter values could significantly lower the power of linkage analysis methods.

Because trait parameter values of interest are the mean, variance, and correlation, they can be estimated without bias in random samples. However, clinical and epidemiological studies based on selected samples are more powerful, because families are usually selected for having at least one member with an extreme trait value. Parameter estimates based on such truncate samples are biased.

One way to obtain trait parameter values for truncate samples is to use information from previous studies of random samples. This approach is fine as long as the populations used in the previous and current studies are the same. Another way to obtain trait parameter values is by using the moments of the truncate sample, although these are biased estimates. Although this approach is not optimal, it is, nevertheless, widely used. When there is no information on the trait distribution or previous studies about the trait of interest, this approach can be used. Another way to obtain trait parameter values, based on the conditional maximum likelihood estimators (CMLEs), was suggested by Peng and Siegmund (2006). As long as the quantitative trait is normally distributed, this approach was shown to be very effective in estimating trait parameter values for truncate samples.

1.2. The Goal of Our Paper

In this paper, we will concentrate on improving the application of existing model-free linkage analysis methods to truncate samples by considering three new approaches for estimating trait parameter values. These three approaches have not previously been applied to linkage analysis, but have been suggested for truncate selection for other purpose (Cohen, 1959; Rao et al., 1968; Mendell and Elston, 1974; Aboueissa and Stoline, 2004). We will evaluate the effect of using five different estimators (the three new approaches based on the truncate likelihood function, and two existing approaches based on the likelihood function and the conditional likelihood function) on the power of the five model-free linkage analysis methods of Haseman and Elston (1972), Xu et al. (2000), Sham and Purcell (2001), Tang and Siegmund (2001) and Cuenco et al (2003).

Although Cuenco et al. (2003) and Bhattacharjee et al. (2008) conducted simulation studies on the power of linkage analysis methods with misspecified trait parameter values, they used randomly picked values, rather than reasonable estimates. Also, although Peng and Siegmund (2006) evaluated the effect of using sample moments and CMLEs, they examined only the power of one linkage analysis method, the Tang and Siegmund method. Our paper moves beyond the existing literature by evaluating five linkage analysis methods with five different trait parameter estimates.

1.3. Background

In this section, we explain some basic concepts in statistical genetics, and then discuss details about existing model-free linkage analysis methods that we will apply to our study. The main sources for this section are Statistics in Human Genetics (Sham, 1997), Quantitative Trait Loci: Methods and Protocols (Camp and Cox, 2002), and Statistical Methods in Genetic Epidemiology (Thomas, 2004).

1.3.1. Terminology in Statistical Genetics

Gregor Mendel, an Augustinian monk, conducted a series of experiments using pea plants. He showed that observable traits (phenotypes) are inherited, by offspring from parents in discrete units that we now call genes. Each gene can have many alternative forms or alleles. About twenty years after Mendel's work, researchers identified the structure and role of chromosomes, which are located inside the nucleus of the cell. The human body's genetic information is contained in two gametes, each of which has 22 pairs of autosomes and one pair of sex chromosomes. The chromosomes contain a long, ladder-like molecule called deoxyribonucleic acid (DNA). The complete genetic sequence for humans is called the human genome. The particular position in the genome is called a locus.

An individual has two alleles at a locus. One allele is transmitted from the two alleles of the father (with equal chance), and the other is transmitted from the two alleles of the mother (with equal chance). This is Mendel's law of segregation. Considering multiple loci, an individual's genotype is formed by two haplotypes (the combination of alleles that are transmitted together), one from the father and the other from the mother. However, when a crossover takes place at an early stage of meiosis, an individual's two haplotypes will be different from those of the parents'. For example, if the paternal and maternal haplotype at loci A and B are $\mathrm{A_f B_f}$ and $\mathrm{A_m B_m}$, offspring may have $\mathrm{A_f B_m}$ and $\mathrm{A_m B_f}$, as well as $\mathrm{A_f B_f}$ and $\mathrm{A_m B_m}$. The recombination fraction is defined as the probability of having (an odd number of) crossovers and is denoted as θ for $0 \le \theta \le 0.5$. When two loci are distant and inherited independently, the probability of crossover is 0.5, the same as the probability of $\mathrm{A_f B_m}$ or $\mathrm{A_m B_f}$ when A and B are on different chromosomes. When two loci are located close enough to each other, it is less likely that the crossover will occur, and the recombination fraction will be close to zero.

In this paper, we will assume Hardy Weinberg Equilibrium (HWE), Linkage Equilibrium (LE) and no epistasis. HWE is the tendency for population genotype frequencies (related to the allele frequencies) to remain unchanged across generations and to be functionally related to the allele frequencies in a specific way. That is, at a bialleleic locus, with alleles $\mathrm{A_1}$ and $\mathrm{A_2}$ having allele frequencies p and $\mathrm{q} = 1 - \mathrm{p}$, the probability of having genotype $\mathrm{A_1A_1}$, $\mathrm{A_1A_2}$, and $\mathrm{A_2A_2}$, are p^2, pq, and q^2, respectively. Assuming there is no selection and that mating is at random with respect to a locus, the three genotypes in the next generation will be $\mathrm{P(A_1A_1) = p^2}$, $\mathrm{P(A_1A_2) = pq}$, and $\mathrm{P(A_2A_2) = q^2}$. Linkage Equilibrium occurs when two alleles at different loci are independent in the population. Thus, the joint occurrence of a gamete with allele $\mathrm{A_i}$, at a locus A with allele frequency $\mathrm{P(A_i)}$, and allele $\mathrm{B_j}$, at a locus B with allele frequency $\mathrm{P(B_j)}$, has frequency of $\mathrm{P(A_i)P(B_j)}$. Epistasis is the interaction between genes. Epistasis occurs when the effect of one gene is modified by another gene. In this paper, we make the assumption that no epistatis is occurring.

1.3.2. Types of Phenotypes

Since the phenotype is the observable trait of a gene, it is easy to understand the concept of qualitative phenotypes as the results of having particular alleles at a trait locus. Traditionally, diseases traits are assumed to be qualitative and dichotomous with one allele being the disease allele and another allele being the healthy allele. Although this is a reasonable formulation for single gene disorders with Mendelian inheritance, more complex models are required to explain the relationship between alleles, disease and disease-related traits.

When phenotypes vary in degree (i.e., are continuous), we call them quantitative phenotypes. Quantitative traits can be more complicated to understand than qualitative traits because they cannot be completely determined by alleles at a single locus. In fact, in the case where there are a large number of genes having small additive effects, the pattern of continuous quantitative traits in a population can be shown to follow a bell curve. This suggests that quantitative traits can be the result of a major locus but also environmental factors and other genes.

1.3.3. The Genetic Model for Quantitative Traits

Our model is based on a general quantitative trait model, which was originally suggested by Fisher in his classic paper (1918), one of the highlights of quantitative genetics. Fisher noted that a quantitative trait value could be determined by a single major genetic effect and environmental effects.

Let us first consider the effect of a single main locus with two alleles. Then we can form a model that includes the environmental effect. Suppose that our trait locus has two alleles A_1 and A_2 with allele frequencies p and $\mathrm{q} = 1 - \mathrm{p}$, and that the mean effect of each genotype has the value a, d and –a, for $\mathrm{A}_1\mathrm{A}_1$, $\mathrm{A}_1\mathrm{A}_2$ and $\mathrm{A}_2\mathrm{A}_2$. These are deviations from the midpoint of two homozygote genotype means. That is, the additive genetic value, a, is half of the difference between two homozygote genotype mean effects. The dominance genetic value, d, is the deviation of the heterozygote mean, from the midpoint of two homozygotes (Camp and Cox, 2002). The dominance genetic value, d, equals a, 0, or –a, depending whether the inheritance of A_1 is dominant, additive, or recessive to A_2, respectively.

The overall trait mean deviation from the midpoint for this single locus genotype, is computed as

$$p^2a + 2pqd + q^2(-a). \tag{1.3.1}$$

Thus, the variance of trait values, due to this gene, is computed as

$$\sigma_g^2 = 2pq\big(a - d(p - q)\big)^2 + 4p^2q^2d^2. \tag{1.3.2}$$

The first and second terms used in genetic variance are called the *additive* and *dominance* variance components, respectively. Eq. (1.4.2) can be rewritten as

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2, \quad \text{where} \tag{1.3.3}$$

$$\sigma_a^2 = 2pq\big(a - d(p - q)\big)^2 \quad \text{and} \tag{1.3.4}$$

$$\sigma_d^2 = 4p^2q^2d^2. \tag{1.3.5}$$

These two components of the genetic variance are frequently referred to in quantitative genetics. The additive variance component represents the additive effects of the individual alleles at a locus, and the dominance variance component shows the interaction between alleles (Thomas, 2004). When the type of inheritance is additive, i.e. $d = 0$, the additive variance component σ_a^2 is proportional to the additive genetic value, a, and the dominance variance component ${\sigma^2}_d$ becomes 0.

Now let us consider the covariance of sibling trait values, in the case where there is a single main locus. In order to explain the covariance, we use the concept of alleles that are identical by descent (IBD). When two relatives have identical alleles, and these alleles are copies of one allele transmitted from a common ancestor, the relatives are identical by descent. Two relatives can share 0, 1, or 2 alleles IBD at a locus. In sibpairs, the probabilities of sharing 0, 1 and 2 alleles IBD are 0.25, 0.5 and 0.25, respectively. Thus, the expected number of alleles IBD in sibpairs is 1. The expectation of the proportion of alleles IBD, which is half of the expected number of alleles IBD , denoted by π, becomes 0.5. The following table shows the expected proportion of alleles IBD for some types of relative pairs:

Table 1.1. Probability of sharing alleles IBD and the expected of the proportion of alleles IBD for different type of relatives (Thomas, 2004)

Type of relative	$E(\pi)$	f_0	f_1	f_2
Monozygotic twins	1	0	0	1
Dizygotic twins	0.5	0.25	0.5	0.25
Full sibs	0.5	0.25	0.5	0.25
Half sibs	0.25	0.5	0.5	0
Grandparents and grand child	0.25	0.5	0.5	0
Random unrelated individuals	0	1	0	0

π: Proportion of alleles IBD, f_0, f_1, f_2: Probability of sharing 0,1, or 2 alleles IBD

Now, let us go back to the covariance of trait values in pairs of relatives resulting from a single main locus. The covariance is dependent on the number of alleles IBD. If the number of alleles IBD, at this locus, equals zero, there would be no contribution to the covariance of the trait values by this gene. If the number of alleles IBD is two, the covariance due to the genotype effect will be exactly the same as the variance $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$. Lastly, if the number of alleles IBD is 1, the covariance will be half of the additive variance component $0.5\sigma_a^2$, since only one allele contributes to the covariance and thus, the effect of interactions between alleles does not have to be considered. The covariance of trait values between two relatives, due to the effect of a single gene, has the form of

$$\mathrm{Cov(two\ relatives)} = f_1(0.5\sigma_a^2) + f_2\left(\sigma_a^2 + \sigma_d^2\right) = \pi\sigma_a^2 + f_2\sigma_{d,}^2 \qquad (1.3.6)$$

where π is the proportion of alleles IBD, and f_1 and f_2 are the probability of sharing 1 and 2 alleles IBD at a locus, respectively. For example, in sibling pairs, the expectation of covariance between two siblings will be $0.5\sigma_a^2 + 0.25\sigma_d^2$ according to Table 1.4.1.

As we mentioned at the beginning of this section, our quantitative trait model assumes the effects of both genotypes and environments. Assuming the HWE, LE and no epistasis, the trait value of an i-th relative has the form of

$$x_i = \mu + g_i + e_i \qquad (1.3.7)$$

where μ is a constant overall trait mean, g_i is the effect of a single genotype, and e_i is the environmental effect, uncorrelated to the genetic effect. Without loss of generality, $E(g_i) = E(e_i) = 0$ can be assumed. Letting X be the random variable of the trait value, we have

$$E(X) = \mu \quad \text{and} \qquad (1.3.8)$$

$$\mathrm{Var}(X) = \sigma_g^2 + \sigma_e^2 = \sigma_a^2 + \sigma_d^2 + \sigma_e^2 \equiv \sigma_x^2. \qquad (1.3.9)$$

Letting X_1, X_2 be the random variable of trait values of two relatives, we have

$$\mathrm{Cov}(X_1, X_2|\pi) = \pi\sigma_a^2 + f_2\sigma_d^2 + \rho_e\sigma_e^2 \ \equiv \rho_\pi\sigma_x^2 \qquad (1.3.10)$$

where $\rho_e = \mathrm{corr}(e_1, e_2)$ is the correlation due to environmental effects, and ρ_π is the conditional correlation between two relatives, X_1 and X_2, given genotype information.

In sibpairs, the marginal covariance can be written as

$$\mathrm{Cov}(X_1, X_2) = 0.5\sigma_a^2 + 0.25\sigma_d^2 + \rho_e\sigma_e^2 \equiv \rho\sigma_x^2 \qquad (1.3.11)$$

where ρ is the marginal correlation between two siblings X_1 and X_2. By substituting $\rho_e\sigma_e^2 = \rho\sigma_x^2 - 0.5\sigma_a^2 - 0.25\sigma_d^2$ from equation (1.3.11) to equation (1.3.10), the conditional covariance between sibpairs is now

$$\mathrm{Cov}(X_1, X_2|\pi) = \rho_\pi\sigma_x^2 = \rho\sigma_x^2 + (\pi - 0.5)\sigma_a^2 + (f_2 - 0.25)\sigma_d^2. \qquad (1.3.12)$$

Tang and Siegmund (2001) suggested rewriting (1.3.12) as

$$\mathrm{Cov}(X_1, X_2|\pi) = \rho_\pi\sigma_x^2 = \rho\sigma_x^2 + (\pi - 0.5)\sigma_g^2 - (f_1 - 0.5)\sigma_d^2/2 \qquad (1.3.13)$$

so the terms involving π have a mean of zero and are uncorrelated.

Before moving to linkage analysis methods for quantitative traits, we introduce one more term, heritability. Heritability is used to measure the contribution of genetic factors in the variability of a trait. It is defined as the ratio of variance due to genetic effects to the overall trait variance. As a narrow definition, it is the ratio of the additive variance component to the overall trait variance, and can be expressed as

$$h^2 = \frac{\sigma_a^2}{\sigma_x^2}. \qquad (1.3.14)$$

Heritability of a trait can be different in different populations.

1.4. Gene Model-Free Linkage Analysis Methods

1.4.1. The Regression Based Method for Linkage

This method is based on the regression of quantitative trait values in siblings, X_1 and X_2, on the proportion of alleles IBD, π, at a trait locus. Let

$$\mathbf{x_i} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_i = \begin{pmatrix} x_{1_i} \\ x_{2_i} \end{pmatrix} = \begin{pmatrix} \mu + g_{1_i} + e_{1_i} \\ \mu + g_{2_i} + e_{2_i} \end{pmatrix} \qquad (1.4.1)$$

be the observed trait value of the i-th sibpair described in the previous section. Also, we denote by

$$x_{D_i} = \left(x_{1_i} - x_{2_i}\right)^2, \qquad (1.4.2)$$

the squared trait difference, and by

$$x_{S_i} = \left(\left(x_{1_i} - \mu\right) + \left(x_{2_i} - \mu\right)\right)^2, \qquad (1.4.3)$$

the mean corrected squared trait sum of the i-th sibpair.

The basic idea of the regression based methods is the following: the expectation of $x_{D_i} = \left(x_{1_i} - x_{2_i}\right)^2$ can be written in the form of a simple regression model (assuming no dominance variance) as

$$E(X_D) = \alpha_D + \beta_D \pi \qquad (1.4.4)$$

where π is the proportion of alleles IBD at a trait locus. It can be shown that the sample slope is an unbiased estimate of genetic variance. That is

$$\beta_D = -2\sigma_g^2. \tag{1.4.5}$$

Haseman and Elston (1972) originally proposed the regression model, given the marker genotype in the form of

$$E(X_D) = \alpha_D + \beta_D \pi + \gamma_D f_1 \tag{1.4.6}$$

where π is the proportion of alleles IBD at a marker locus, and f_1 is the probability of sharing 2 alleles IBD at a marker locus. They also considered dominance variance and showed that

$$\alpha_D = 2(1-\rho_e)\sigma_e^2 + 2\Psi\sigma_g^2 + 2\Psi(1-2\Psi)\sigma_d^2\,, \tag{1.4.7}$$

$$\beta_D = 2(1-2\Psi){\sigma^2}_g\,, \qquad \text{and} \tag{1.4.8}$$

$$\gamma_D = (1-2\Psi)^2\sigma_d^2. \tag{1.4.9}$$

Here, $\Psi = \theta^2 + (1-\theta)^2$ and θ denotes the recombination fraction between the trait and the marker loci. Since θ is between 0 and 0.5, Ψ also has values between 0 and 0.5. Thus, β_D can equal 0 or have a negative value. It will be 0 when there is no linkage between the trait and the marker locus when the recombination fraction equals 0.5. Haseman and Elston tested the null hypothesis of no linkage, $H_0\colon \beta_D = 0$ against the alternative hypothesis, $H_A\colon \beta_D < 0$. Here the test statistic is a one sided t-statistic for the slope estimate $\hat{\beta}_D$.

Many regression-based methods have been proposed with the objective of increasing power to detect linkage. Two variations by Xu et al. (2000) and Sham and Purcell (2001) have shown to have greater power than others methods, according to Cuenco et al (2003).

Xu et al. (2000) performed two regressions of $x_{D_i} = \left(x_{1_i} - x_{2_i}\right)^2$ and $x_{S_i} = \left(\left(x_{1_i} - \mu\right) + \left(x_{2_i} - \mu\right)\right)^2$ on π_i separately, and proposed a new test statistic based on the weighted average of the slope estimates $\hat{\beta}_D$ and $\hat{\beta}_S$. (The regression coefficient β_S is $-2(1-2\Psi){\sigma^2}_g$ at a marker locus (Drigalenko, 1998) and the slope estimate $\hat{\beta}_S$ is a one sided t-statistic for the hypothesis $H_0\colon \beta_S = 0$ vs. $H_A\colon \beta_S > 0$.) The new test statistic has the form of

$$\begin{aligned}\hat{\beta} &= w\left(-\hat{\beta}_D\right) + (1-w)\hat{\beta}_S \\ &= \frac{\hat{\sigma}_S^2 - \hat{\sigma}_{DS}^2}{\hat{\sigma}_D^2 + \hat{\sigma}_S^2 - 2\hat{\sigma}_{DS}^2}\left(-\hat{\beta}_D\right) + \frac{\hat{\sigma}_D^2 - \hat{\sigma}_{DS}^2}{\hat{\sigma}_D^2 + \hat{\sigma}_S^2 - 2\hat{\sigma}_{DS}^2}\hat{\beta}_S\end{aligned} \qquad (1.4.10)$$

where $w = \frac{\hat{\sigma}_S^2 - \hat{\sigma}_{DS}^2}{\hat{\sigma}_D^2 + \hat{\sigma}_S^2 - 2\hat{\sigma}_{DS}^2}$, $\hat{\sigma}_D^2 = \widehat{Var}(\hat{\beta}_D)$, $\hat{\sigma}_S^2 = \widehat{Var}(\hat{\beta}_S)$ and $\hat{\sigma}_{DS}^2 = \widehat{Cov}(\hat{\beta}_D, \hat{\beta}_S)$. The null hypothesis is $H_0\colon \beta = 0$ against the alternative hypothesis of $H_A\colon \beta > 0$.

Sham and Purcell (2001) regressed the weighted linear combination of x_{D_i} and x_{S_i} on the mean corrected proportion of alleles IBD, $\pi_i - 0.5$, shown as

$$E\left(\frac{X_S}{(1+\rho)^2} - \frac{X_D}{(1-\rho)^2} + \frac{4\rho}{1-\rho^2}\right) = \alpha + \beta(\pi - 0.5) \qquad (1.4.11)$$

where ρ is the marginal correlation. The test statistic is a one sided t-statistic for the slope estimate $\hat{\beta}$, with null hypothesis $H_0\colon \beta = 0$, against the alternative hypothesis $H_A\colon \beta > 0$. Sham and Purcell also showed that this regression-based method has power equivalent to likelihood-based methods. The Sham and Purcell model assumes that the trait variance equals one. We note that the linear combination of x_{D_i} and $x_{S_i}$of the regression model is from the bivariate normal distribution. In order to evaluate power, with misspecified trait parameter values, including the variance, in our simulation, we will use the modified Sham and Purcell method as follows

$$\mathrm{E}\left(\frac{\mathrm{X_S}}{\sigma_x^2(1+\rho)^2}-\frac{\mathrm{X_D}}{\sigma_x^2(1-\rho)^2}+\frac{4\rho}{1-\rho^2}\right)=\alpha+\beta(\pi-0.5) \qquad (1.4.12)$$

with the slope estimate $\hat{\beta}$ with null hypothesis $\mathrm{H_0}: \beta = 0$ against the alternative hypothesis $\mathrm{H_A}: \beta > 0$.

1.4.2. Likelihood Based Method

This method is based on the likelihood function of trait values under the assumption of multivariate normality. Specifically, the joint distribution is bivariate normal in sib pairs, and multivariate normal in a pedigree of k relatives.

$$\ln \mathrm{L} = c - \frac{1}{2}\sum_{i=1}^{N} \ln \det(\boldsymbol{\Sigma}_{\mathbf{i}}\,) - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x_i}-\boldsymbol{\mu}_{\mathbf{i}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{i}}{}^{-1}(\mathbf{x_i}-\boldsymbol{\mu}_{\mathbf{i}}) \qquad (1.4.13)$$

where $\mathbf{x_i}$ is the trait value vector of the i-th pedigree, $\boldsymbol{\mu}_{\mathbf{i}} = \mathrm{E}(\mathbf{x_i})$ is the mean vector and $\boldsymbol{\Sigma}_{\mathbf{i}}$ is the covariance matrix of the i-th pedigree. Note that this likelihood can be applied to any size pedigree, as long as the assumption of multivariate normality holds. The variance and the covariance between two pedigree members are given in equations (1.3.9) and (1.3.10). Amos (1994) showed the covariance at a marker locus, as well as at a trait locus, by assigning coefficients of σ_a^2 and σ_d^2 as the function of proportion of alleles IBD and the recombination fraction θ. However, he suggested assuming $\theta = 0$ in hypothesis tests for linkage when the data are from only one type of pedigree, where the unique estimates of θ and σ_a^2 are not guaranteed. In our simulation, we consider only sibpairs, so we will assume that $\theta = 0$ and use the covariance of equation (1.3.10) for the following likelihood-based methods.

Using this log likelihood function, Amos (1994) tested the null hypothesis of no linkage $H_0: \sigma_a^2 = 0$ against the alternative hypothesis $H_A: \sigma_a^2 > 0$, using the χ^2 log likelihood ratio test statistic, $-2(\ln L - \ln L_0)$. He suggested estimating the linkage parameters σ_a^2 (as well as other parameters) by using generalized estimating equation (GEE) approaches.

Recently, many researchers prefer score statistics to log likelihood ratio statistics because of the simplicity of computation. The score test is asymptotically equivalent to the likelihood ratio test, but the computation of the maximum likelihood estimates should be done only under the null hypothesis (Carroll et al., 2006). In general, if we are testing $H_0: \eta = \eta_0$, the score test statistic is defined as

$$Z_S = S(\eta_0)/\sqrt{I_n(\eta_0)} \tag{1.4.14}$$

where

$$S(\eta) = \frac{\partial}{\partial \eta} \ln L(\eta|X) \quad \text{and} \tag{1.4.15}$$

$$I_n(\eta) = \mathrm{Var}_\theta\big(S(\eta)\big) = -E_\eta\left(\frac{\partial^2}{\partial \eta^2} \ln L(\eta|X)\right) \tag{1.4.16}$$

(Casella and Berger, 2002).

As we mentioned earlier, score statistics in linkage analysis are similar to one another and have been modified to achieve robustness (Cuenco et al, 2003). Two of them, by Tang and Siegmund (2001), and Cuenco et al. (2003), are shown below.

From sibling pairs of $\mathbf{x_i} = \begin{pmatrix} x_{1_i} \\ x_{2_i} \end{pmatrix}$ for $i = 1, \cdots, n$, with $\boldsymbol{\mu} = \begin{pmatrix} \mu \\ \mu \end{pmatrix}$ and $\boldsymbol{\Sigma_i} = \begin{pmatrix} \sigma_x^2 & \rho_{\pi_i}\sigma_x^2 \\ \rho_{\pi_i}\sigma_x^2 & \sigma_x^2 \end{pmatrix}$, where ρ_{π_i} is the conditional correlation between two siblings for a given genotype, the likelihood function in equation (1.4.13) becomes

$$\begin{aligned} \ln L = c + \sum_{i=1}^{N} \Bigg(&- \ln \sigma^2 - \frac{1}{2}\ln\left(1 - \rho_{\pi_i}^2\right) - \frac{\left(x_{1_i} - x_{2_i}\right)^2}{4\sigma_x^2\left(1 - \rho_{\pi_i}\right)} \\ &- \frac{\left(x_{1_i} + x_{2_i} - 2\mu\right)^2}{4\sigma_x^2\left(1 - \rho_{\pi_i}\right)} \Bigg) \quad \text{and thus,} \end{aligned} \tag{1.4.17}$$

$$\begin{aligned} S(\sigma_a^2) &= \frac{\partial}{\partial \sigma_a^2} \ln L \\ &= \sum_{i=1}^{N} \left(\frac{\rho_{\pi_i}}{1 - \rho_{\pi_i}^2} - \frac{\left(x_{1_i} - x_{2_i}\right)^2}{4\sigma_x^2\left(1 - \rho_{\pi_i}\right)^2} + \frac{\left(x_{1_i} + x_{2_i} - 2\mu\right)^2}{4\sigma_x^2\left(1 - \rho_{\pi_i}\right)^2} \right) \left(\frac{\pi_i - 0.5}{\sigma_x^2} \right) \end{aligned} \tag{1.4.18}$$

since

$$\frac{\partial \rho_{\pi_i}}{\partial \sigma_a^2} = \frac{\pi_i - 0.5}{\sigma_x^2} \tag{1.4.19}$$

from equation (1.4.13) assuming no dominance, as Tang and Siegmund (2001) suggested.

Setting the trait variance equal to 1.0, Tang and Siegmund (2001)'s suggested a robust score statistic in the form of

$$\frac{\Sigma \left(\frac{\rho}{1 - \rho^2} - \frac{\left(x_{1_i} - x_{2_i}\right)^2}{4(1 - \rho)^2} + \frac{\left(x_{1_i} + x_{2_i} - 2\mu\right)^2}{4(1 + \rho)^2} \right) (\pi_i - 0.5)}{\frac{1}{2\sqrt{2}} \sqrt{\Sigma \left(\frac{\rho}{1 - \rho^2} - \frac{\left(x_{1_i} - x_{2_i}\right)^2}{4(1 - \rho)^2} + \frac{\left(x_{1_i} + x_{2_i} - 2\mu\right)^2}{4(1 + \rho)^2} \right)^2}} \tag{1.4.20}$$

and Cuenco et al. (2003) suggested another score statistic by using the empirical standard deviation of π instead of $\frac{1}{2\sqrt{2}}$, in the form of

$$\frac{\sum\left(\frac{\rho}{1-\rho^2}-\frac{\left(x_{1_i}-x_{2_i}\right)^2}{4(1-\rho)^2}+\frac{\left(x_{1_i}+x_{2_i}-2\mu\right)^2}{4(1+\rho)^2}\right)(\pi_i-0.5)}{\sqrt{\frac{\sum(\pi_i-0.5)^2}{n}}\sqrt{\sum\left(\frac{\rho}{1-\rho^2}-\frac{\left(x_{1_i}-x_{2_i}\right)^2}{4(1-\rho)^2}+\frac{\left(x_{1_i}+x_{2_i}-2\mu\right)^2}{4(1+\rho)^2}\right)^2}}. \qquad (1.4.21)$$

Note that the score statistics include only a marginal correlation, because the computation is done under the null hypothesis of no linkage.

In our simulation, we would like to evaluate different estimates for trait parameter values, including the variance, as well as the mean and the correlation. Thus, we will use a modified version of Tang and Siegmund's method and Cuenco et al's method of

$$\frac{\sum\left(\frac{\rho}{1-\rho^2}-\frac{\left(x_{1_i}-x_{2_i}\right)^2}{4\sigma_x^2(1-\rho)^2}+\frac{\left(x_{1_i}+x_{2_i}-2\mu\right)^2}{4\sigma_x^2(1+\rho)^2}\right)\left(\frac{\pi_i-0.5}{\sigma_x^2}\right)}{\frac{1}{2\sqrt{2}}\sqrt{\sum\left(\frac{\rho}{1-\rho^2}-\frac{\left(x_{1_i}-x_{2_i}\right)^2}{4\sigma_x^2(1-\rho)^2}+\frac{\left(x_{1_i}+x_{2_i}-2\mu\right)^2}{4\sigma_x^2(1+\rho)^2}\right)^2\left(\frac{1}{\sigma_x^2}\right)^2}} \text{ and} \qquad (1.4.22)$$

$$\frac{\sum\left(\frac{\rho}{1-\rho^2}-\frac{\left(x_{1_i}-x_{2_i}\right)^2}{4\sigma_x^2(1-\rho)^2}+\frac{\left(x_{1_i}+x_{2_i}-2\mu\right)^2}{4\sigma_x^2 4(1+\rho)^2}\right)\left(\frac{\pi_i-0.5}{\sigma_x^2}\right)}{\sqrt{\frac{\sum(\pi_i-0.5)^2}{n}}\sqrt{\sum\left(\frac{\rho}{1-\rho^2}-\frac{\left(x_{1_i}-x_{2_i}\right)^2}{4\sigma_x^2(1-\rho)^2}+\frac{\left(x_{1_i}+x_{2_i}-2\mu\right)^2}{4\sigma_x^2(1+\rho)^2}\right)^2\left(\frac{1}{\sigma_x^2}\right)^2}}. \qquad (1.4.23)$$

1.5. Sampling Designs

There are several sampling approaches in linkage analysis. In many cases, the sampling is based on trait values. For example, in clinical studies, families are often selected when at least one member has extreme trait values. This type of ascertainment based on a single-proband is called truncate sampling in our paper. It has been known that single-proband sampling has the advantage of greater power than random population sampling (Feingold, 2001). When a pedigree has a constant sibship size, the power gains are even greater (Feingold, 2001). However, the trait distribution of this selected sample is different from that of the population.

Chapter 2 Methods

In this chapter, we will show two existing approaches and propose three new approaches for estimating trait parameter values in truncate samples, assuming that the trait distribution in sibpairs is bivariate normal. The two existing estimates are based on the likelihood function for random samples and the conditional likelihood function. On the other hand, the three new estimating methods are based on the likelihood function for truncate samples, where at least one sibling has an extreme trait value. In these three new methods, the trait mean and the trait variance will be estimated together, but the correlation will be estimated separately. In this chapter, we will focus on estimating only the marginal correlation ρ, not the conditional correlation, because the model-free linkage analysis methods we use in our study, require only a marginal correlation in their test statistics.

2.1. Existing Approaches

2.1.1. Sample Moments

Working with sample moments is the most commonly used way to estimate parameter values, especially for random samples. However, this may not be a good approach for truncate samples because it does not allow for the selection. Assuming a single multivariate normal distribution, we can easily get the maximum likelihood estimates for trait parameter values. We will first review the MLEs for general pedigrees with sibship of size s.

Suppose that the trait value $\mathbf{X} \in \mathbb{R}^{\mathbf{s}}$ (column vector), of s siblings, comes from a distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^{\mathbf{s}}$ and $\mathrm{s} \times \mathrm{s}$ dimensional covariance matrix $\boldsymbol{\Sigma}$. If we assume a multivariate normal distribution for $\mathbf{X}$, the probability density function (pdf) of $\mathbf{X}$ is written as

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{s/2}} \cdot \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{2.1}$$

From the independent and identically distributed n observations of $\mathbf{x_i}$ for $\mathrm{i} = 1, \cdots, \mathrm{n}$, we get the log likelihood function of

$$\begin{aligned} \ln \mathrm{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{\mathrm{sn}}{2}\ln 2\pi - \frac{\mathrm{n}}{2}\ln\det(\boldsymbol{\Sigma}) \\ -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathbf{x_i}-\boldsymbol{\mu})(\mathbf{x_i}-\boldsymbol{\mu})^{\mathrm{T}}\right). \end{aligned} \tag{2.2}$$

We can derive the maximum likelihood estimates (MLEs) of the population mean vector and the covariance matrix as

$$\hat{\boldsymbol{\mu}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbf{x_i} \quad \text{and} \tag{2.3}$$

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - \widehat{\boldsymbol{\mu}})(\mathbf{x_i} - \widehat{\boldsymbol{\mu}})^{T}. \qquad (2.4)$$

If we consider sibpairs (i.e. $s = 2$), the MLEs of three population parameter values μ, σ and ρ from n observations of random sample $\mathbf{x_i} = \binom{x_{1_i}}{x_{2_i}}$ for $i = 1, \cdots, n$, along with $\boldsymbol{\mu} = \binom{\mu}{\mu}$ and $\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$ will be

$$\hat{\mu} = \frac{1}{2n}\sum_{i=1}^{n}(x_{1_i} + x_{2_i}), \qquad (2.5)$$

$$\hat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n}\left\{\left(x_{1_i} - \hat{\mu}\right)^2 + (x_{2_i} - \hat{\mu})^2\right\} \quad \text{and} \qquad (2.6)$$

$$\hat{\rho} = \frac{1}{\hat{\sigma}^2} \cdot \frac{1}{n}\sum_{i=1}^{n}\left(x_{1_i} - \hat{\mu}\right)(x_{2_i} - \hat{\mu}). \qquad (2.7)$$

However, in truncate samples, these estimates are no longer MLEs, even though Peng and Siegmund (2006) and Bhattacharjee et al. (2008) retained the term “MLEs”. In order to avoid confusion, we will call the estimates shown in equations (2.5), (2.6) and (2.7) “sample moments”.

2.1.2. Conditional Maximum Likelihood Estimates

Peng's conditional MLEs (Peng and Siegmund, 2006) can be used when the proband is known (Bhattacharjee et al., 2008). As in the previous section, suppose that trait value $\mathbf{X} \in \mathbb{R}^{\mathbf{s}}$ (column vector) obtained from a sibship of size s, is from the distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^{\mathbf{s}}$ and $\mathrm{s} \times \mathrm{s}$ dimensional covariance matrix $\mathbf{\Sigma}$, assuming a multivariate normal distribution. $\mathbf{X}$ can be rewritten as $\begin{pmatrix} \mathrm{X}_1 \\ \mathbf{X_2} \end{pmatrix}$, where $\mathrm{X}_1 \in \mathbb{R}$ is the proband's trait value, and $\mathbf{X_2} \in \mathbb{R}^{\mathrm{s}-1}$ is his/her sibling's trait value. Peng's conditional log likelihood function of $\mathbf{X_2}$ given $\mathrm{X}_1 = \mathrm{x}_1$ is

$$\begin{aligned} \ln \mathrm{L}(\boldsymbol{\mu}_\mathbf{c}, \mathbf{\Sigma}_\mathbf{c}) = -\frac{\mathrm{sn}}{2}\ln 2\pi - \frac{\mathrm{n}}{2}\ln\det(\mathbf{\Sigma}_\mathbf{c}) \\ -\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}_\mathbf{c}{}^{-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathbf{x}_{\mathbf{2}_\mathbf{i}} - \boldsymbol{\mu}_\mathbf{c}\right)\left(\mathbf{x}_{\mathbf{2}_\mathbf{i}} - \boldsymbol{\mu}_\mathbf{c}\right)^{\mathrm{T}}\right). \end{aligned} \tag{2.8}$$

The equation above is similar to equation (2.2) but with the conditional mean $\boldsymbol{\mu}_\mathbf{c} = \mathrm{E}(\mathbf{X_2}|\mathrm{X}_1 = \mathrm{x}_1)$ and the conditional covariance matrix $\mathbf{\Sigma}_\mathbf{c} = \mathrm{Cov}(\mathbf{X_2}|\mathrm{X}_1 = \mathrm{x}_1)$.

Let us now consider n observations of sibpairs $\mathbf{x_i} = \begin{pmatrix} \mathrm{x}_{1_\mathrm{i}} \\ \mathrm{x}_{2_\mathrm{i}} \end{pmatrix}$ for $\mathrm{i} = 1, \cdots, \mathrm{n}$, where $\boldsymbol{\mu} = \begin{pmatrix} \mu \\ \mu \end{pmatrix}$ and $\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$. Then the equation (2.8) reduces to

$$\begin{aligned} \ln \mathrm{L}(\mu, \sigma^2, \rho) = -\mathrm{n}\ln 2\pi - \frac{\mathrm{n}}{2}\ln\left(\sigma^2(1-\rho^2)\right) \\ -\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{\left(\mathrm{x}_{2_\mathrm{i}} - \left(\mu + \rho\left(\mathrm{x}_{1_\mathrm{i}} - \mu\right)\right)\right)^2}{2\sigma^2(1-\rho^2)}. \end{aligned} \tag{2.9}$$

where

$$\mu_c = E(X_2|X_1 = x_1) = \mu + \rho(x_1 - \mu) \quad \text{and} \tag{2.10}$$

$$\Sigma_c = Var(X_2|X_1 = x_1) = \sigma^2(1 - \rho^2). \tag{2.11}$$

Peng and Siegmund (2006) then suggests the use of a numerical iterative method to obtain the estimates of μ, σ and ρ, maximizing the conditional log likelihood function. These estimates are called conditional MLEs.

2.2. Suggested Approaches

The following three new methods use the likelihood function for truncate samples, where at least one sibling has an extreme trait value, i.e. a trait value greater than a threshold value T. Like existing methods, we assume that the trait value $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ of sibpairs has a bivariate normal distribution with mean $\boldsymbol{\mu} = \begin{pmatrix} \mu \\ \mu \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$.

If we define a random variable $X^* = \begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix}$ of a truncate sibpair where X_1^* and X_2^* represent a proband's trait and his/her sibling's trait, the pdf of X_1^* is

$$f(\mathrm{x}_1^*) = \begin{cases} \dfrac{f(\mathrm{x}_1^*;\mu,\sigma^2)}{\int_{\mathrm{T}}^{\infty} f(\mathrm{x}_1^*;\mu,\sigma^2)} = \dfrac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\mathrm{x}_1^*-\mu)^2}{2\sigma^2}\right)}{1-\Phi\left(\frac{\mathrm{T}-\mu}{\sigma}\right)} & \text{if } \mathrm{x}_1^* > T \\ 0 & \text{if } \mathrm{x}_1^* \le \mathrm{T} \end{cases} \quad \text{and} \tag{2.12}$$

and its log likelihood function from n observations of $\mathrm{x}_\mathrm{i}^* = \begin{pmatrix} \mathrm{x}_{1_\mathrm{i}}^* \\ \mathrm{x}_{2_\mathrm{i}}^* \end{pmatrix}$ for $\mathrm{i} = 1, \cdots, \mathrm{n}$, is

$$\begin{aligned} \ln \mathrm{L}(\mu,\sigma^2) = -\mathrm{n}\ln\left(1-\Phi\left(\frac{\mathrm{T}-\mu}{\sigma}\right)\right) - \frac{\mathrm{n}}{2}\ln 2\pi - \frac{\mathrm{n}}{2}\ln\sigma^2 \\ -\frac{1}{2\sigma^2}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{x}_{1_\mathrm{i}}^* - \mu\right)^2. \end{aligned} \tag{2.13}$$

This likelihood function will be used to get “truncated maximum likelihood estimates” for μ and σ^2 as shown in subsections 2.2.1, 2.2.2, and 2.2.3. At the end of this section we explain how to derive the correlation estimate for truncate samples.

2.2.1. Shin’s Application to Rao and Mendell (Shin RM)

It has been shown by many investigators, including Rao et al. (1968) and Mendell and Elston (1974) that the estimates of μ and σ^2 by the method of moments, from n observations of $\mathrm{x}_{1_\mathrm{i}}^*$ for $\mathrm{i} = 1, \cdots, \mathrm{n}$, are

$$\hat{\mu} = \overline{\mathrm{x}_1^*} - \lambda\hat{\sigma} \quad \text{and} \tag{2.14}$$

$$\hat{\sigma}^2 = \frac{1}{1-\lambda(\lambda - \mathrm{z}_\mathrm{T})}\mathrm{s}_1^{*2} \tag{2.15}$$

where

$$\overline{x_1^*} = \frac{1}{n}\sum_{i=1}^{n} x_{1_i}^*, \tag{2.16}$$

$$s_1^{*\,2} = \frac{1}{n}\sum_{i=1}^{n}(x_{1_i}^* - \overline{x_1^*})^2, \tag{2.17}$$

$$z_T \equiv \frac{T-\mu}{\sigma} \quad \text{and} \tag{2.18}$$

$$\lambda \equiv \frac{\varphi(z_T)}{1-\Phi(z_T)}. \tag{2.19}$$

Here, z_T is the standardized threshold of T and λ is the ratio of the density of probability to the tail area, where $\varphi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ and $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{u^2}{2}\right)du$ are the pdf and cdf of a standard normal random variable.

It can be shown that from the log likelihood function for truncate samples in equation (2.13) that these sample moments are MLEs of μ and σ^2. We set the first derivatives of the likelihood function with respect to μ and σ^2 to be zero as follows:

$$\frac{\partial \ln L}{\partial\, \mu} = -\frac{n\lambda}{\sigma} + \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_{1_i}^* - \mu\right) \equiv 0 \quad \text{and} \tag{2.20}$$

$$\frac{\partial \ln L}{\partial\, \sigma^2} = -\frac{n\lambda z_T}{2\sigma^2} - \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(x_{1_i}^* - \mu\right)^2 \equiv 0. \tag{2.21}$$

Here z_T and λ are defined as in equations (2.18) and (2.19), respectively. Given the truncate sample with a threshold value T, we get MLEs of the original population parameters

$$\hat{\mu} = \overline{x_1^*} - \lambda\hat{\sigma} \quad \text{and} \tag{2.22}$$

$$\widehat{\sigma}^2 = \frac{1}{1-\lambda(\lambda - z_T)} s_1^{*2} \tag{2.23}$$

which are the same as the sample moments of equations (2.14) and (2.15).

These estimates can be derived simply and analytically as shown above. However, it is natural to wonder how one could obtain the standardized threshold value z_T $(\text{and } \lambda = \lambda(z_T).)$ In practice, we usually do not know μ and σ^2 (which is the reason these values are being estimated), so z_T is unlikely to be obtained since it corresponds to the threshold value of T.

However, we often have a set value of $p \equiv \Pr(X_1 > T)$ or a rough idea of p. For example, you may have chosen T because it is known that roughly 10% of the population is on or above T. In this case, $z_T = -\Phi^{-1}(p)$. Or if we estimate p by $\hat{p}$, then we get the estimates $\widehat{z_T} = -\Phi^{-1}(\hat{p})$.

So we have reason to suppose that we know p as well as T, when we apply our estimates for the mean and variance in equations (2.22) and (2.23).

2.2.2. Shin’s Application to Cohen and Aboueissa (Shin CA)

In this section, we will show the computer algorithm for obtaining Cohen’s MLEs (Cohen, 1959) for the mean and variance from normally distributed singly censored or singly truncated samples as suggested by Aboueissa and Stoline (2004). This algorithm was originally written for censored samples, but we can easily modify the method for truncate samples by substituting the censoring level with $p = \Pr(X_1 > T)$. Cohen (1959) used the same likelihood function (2.13) from n observations of $x^*_{1_i}$ for $i = 1, \cdots, n$, and developed the system of equations

$$\hat{\mu} = \overline{x_1^*} - \xi(\overline{x_1^*} - T) \tag{2.24}$$

$$\hat{\sigma}^2 = {s_1^*}^2 + \xi(\overline{x_1^*} - T)^2 \tag{2.25}$$

providing the tabled values for

$$\xi = \frac{\frac{1-p}{p}\lambda}{\frac{1-p}{p}\lambda - z_T} \tag{2.26}$$

where $\overline{x_1^*} = \frac{1}{n}\sum_{i=1}^{n} x^*_{1_i}$, ${s_1^*}^2 = \frac{1}{n}\sum_{i=1}^{n}(x^*_{1_i} - \overline{x_1^*})^2$, $z_T = \frac{T-\mu}{\sigma}$, $\lambda = \frac{\varphi(z_T)}{1-\Phi(z_T)}$ are defined the same as in the previous section in equations (2.16), (2.17), (2.18) and (2.19).

Aboueissa and Stoline (2004) suggested a new algorithm that does not require the auxiliary table for $\xi = \frac{\frac{1-p}{p}\lambda}{\frac{1-p}{p}\lambda - z_T}$ and claimed that his replacement method was superior to existing replacement methods. His algorithm is based on solving the equation

$$\frac{1 - \frac{1-p}{p}\lambda\left(\frac{1-p}{p}\lambda - z_T\right)}{\left(\frac{1-p}{p}\lambda - z_T\right)^2} = \frac{{s_1^*}^2}{(\overline{x_1^*} - T)^2} \equiv \gamma \tag{2.27}$$

which comes from Cohen's derivation for z_T, which can be written as the second degree polynomials of z_T,

$$\gamma z_T^2 + (\gamma + 1)\left(\frac{1-p}{p}\lambda\right)^2 - (2\gamma + 1)\left(\frac{1-p}{p}\lambda\right)z_T - 1 = 0. \tag{2.28}$$

Once z_T is estimated numerically using the equation above, ξ, $\hat{\mu}$, and $\hat{\sigma}^2$ can be obtained from equations (2.26), (2.24) and (2.25), respectively.

Note that again we assume that p as well as T are known. However, z_T can be estimated by using equation (2.28) differently from its application in the previous section, where $\widehat{z_T} = -\Phi^{-1}(\hat{p})$.

2.2.3. Shin's application to the Newton-Raphson algorithm (Shin Nu)

The second method is the same as the first one except for the estimating algorithm for z_T given p. If we only know the threshold value of T on and above which we selected our truncate sample, the numerical iterative method would give us the trait values of μ and σ^2, maximizing the log likelihood function of equation (2.13).

There are many numerical algorithms to get MLEs, but since Peng and Siegmund (2006) recommended using the Newton-Raphson algorithm to get his conditional MLEs, we would like to apply the same algorithm to get MLEs of log likelihood functions for truncate samples.

The basic structure of an iteration of the Newton-Raphson algorithm is included in the R software package, as a nonlinear minimization function (Schnabel, 1985). Since we would like to maximize the log likelihood function of equation (2.13), we denote this function by $q(\omega) = -\ln L(\omega)$, where $\omega = (\ \mu, \sigma^2)$. Using the initially estimated parameter values of ω, for example, we can simply use the sample moments as $\omega_0 = \left(\overline{x_1^*}, {s_1^*}^2\right)$ and compute a gradient and a hessian of $q(\omega_0)$. We then compute the next estimate of ω by using the multiplication of the inverse of a hessian and a gradient.

Again, by evaluating the gradient of a new estimate, we decide to stop or return to the first step with the new estimate as an initial ω_0. When we cannot compute the exact gradient or hessian, we use an approximation of it.

2.2.4. Estimating Correlation for Truncate Samples

Now let us estimate the correlation, ρ, using n pairs of observations $x_i^* = \begin{pmatrix} x_{1_i}^* \\ x_{2_i}^* \end{pmatrix}$ for $i = 1, \cdots, n$. We assumed in the beginning of section 2.2. that X_1 and X_2 follows a bivariate normal distribution, with $\mu = \mu_1 = \mu_2$, $\sigma = \sigma_1 = \sigma_2$ and ρ. In that case the conditional distribution of X_2 given $X_1 = x_1$ is normal and the conditional expectation has the form of

$$E(X_2|X_1 = x_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}\left(x_1 - \mu_1\right) = \mu(1-\rho) + \rho x_1. \qquad (2.29)$$

Since the assumption of normality implies using the simple linear regression model

$$E(X_2|X_1 = x_1) = \alpha + \beta x_1, \qquad (2.30)$$

we calculate $\alpha = \mu(1-\rho)$ and $\beta = \rho$.

Equations (2.29) and (2.30) still hold for $X^* = \begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix}$ where $X_1^*|X_1 > T$ and $X_2^*|X_1^*$ as

$$E(X_2^*|X_1^* = x_1^*) = \mu(1-\rho) + \rho x_1^*, \qquad (2.31)$$

$$E(X_2^*|X_1^* = x_1^*) = \alpha + \beta x_1^*. \qquad (2.32)$$

Thus, we can get the MLE of β in the regression model in the form of

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_{2_i}^* - \overline{x_2^*})(x_{1_i}^* - \overline{x_1^*})}{\sum_{i=1}^{n}(x_{1_i}^* - \overline{x_1^*})^2} \tag{2.33}$$

using the log likelihood function of

$$\ln L\left(\alpha, \beta, \sigma_{x_2^*}^2\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{\sum_{i=1}^{n}\left(x_{2_i}^* - \alpha - \beta x_{1_i}^*\right)^2}{2\sigma_{x_2^*}^2} \tag{2.34}$$

where $\sigma_{x_2^*}^2$ is the variance of X_2^*. Thus, the correlation estimate in our suggested methods for our truncate samples is

$$\hat{\rho} = \hat{\beta} = \frac{\sum_{i=1}^{n}(x_{2_i}^* - \overline{x_2^*})(x_{1_i}^* - \overline{x_1^*})}{\sum_{i=1}^{n}(x_{1_i}^* - \overline{x_1^*})^2}. \tag{2.35}$$

Chapter 3 Simulations

We conducted a simulation study in order to investigate the power of linkage analysis methods with five different nuisance parameter estimates. We generated artificial data for simple nuclear families with two siblings and picked families in which at least one sibling had an extreme trait value.

The steps in the simulation of our data set are as follows. Based on the genetic model described in Chapter 1, we first generated the parents' genotypes at a trait locus. Then the offspring's genotype and quantitative phenotype were generated. Quantitative phenotype values were standardized using the overall mean and variance. This standardization step was done in order to simplify the comparisons of various estimators. We also incorporated additional environmental correlation between siblings. Once the simulation at the trait locus was completed, we simulated a marker locus taking into account the recombination fraction between two loci. Finally, we stored truncate samples where at least one sibling had a quantitative trait greater than a given threshold value. The probability of alleles shared IBD was computed by Merlin software using the parents' genotypes. Details of the algorithm are described below.

Assuming HWE and random mating, we considered two alleles A_1 and A_2 with allele frequencies p and $q = 1 - p$, respectively, at a trait locus. If we denote G_i (i=1,2,3) as the genotype at a trait locus, the probability of having genotype G_i is the following:

$$\mathrm{P}(\mathrm{G_i}) = \begin{cases} \mathrm{P}(\mathrm{G_1}) \equiv \mathrm{P}(\mathrm{A_1A_1}) = \mathrm{p}^2 \\ \mathrm{P}(\mathrm{G_2}) \equiv \mathrm{P}(\mathrm{A_1A_2}) = \mathrm{pq} \\ \mathrm{P}(\mathrm{G_3}) \equiv \mathrm{P}(\mathrm{A_2A_2}) = \mathrm{q}^2. \end{cases} \tag{3.1}$$

First, the genotype of parents at a trait locus was simulated using a uniform random number generator. If a uniform random number fell between 0 and p, we assigned $\mathrm{A_1}$ to the first allele $\mathrm{A_1}$. Otherwise, we assigned $\mathrm{A_2}$ as the genotype. Similarly, we generated another uniform random number to assign the second allele. Once both parents' genotypes were simulated, we simulated the offsprings' genotypes. There were four possible genotypes with equal chances of being transmitted. If a random number fell between 0 and 0.25, the offspring received the first paternal and the first maternal alleles. If a random number fell between 0.25 and 0.5, the offspring received the first paternal and the second maternal alleles. If a random number fell between 0.5 and 0.75, the offspring received the second paternal and the first maternal alleles. Finally, if a random number fell between 0.75 and 1, the offspring received the second paternal and the second maternal alleles.

Once the offspring's genotypes were simulated, we simulated quantitative trait values. As mentioned in Section 1.2.4, the actual distribution of a quantitative trait is a *mixture of Gaussians*. Given the genotype, the trait value is equal to the genotype mean effect plus a normally distributed environmental effect. Let the mean and the variance of environmental effect be 0 and σ_e^2, respectively. Then, the distribution of trait value can be written as

$$\begin{aligned} &\mathrm{X}|\mathrm{G_1} \sim \mathrm{N}(\ \ \mathrm{a}, \sigma_e^2) \\ &\mathrm{X}|\mathrm{G_2} \sim \mathrm{N}(\ \ \mathrm{d}, \sigma_e^2) \\ &\mathrm{X}|\mathrm{G_3} \sim \mathrm{N}(-\mathrm{a}, \sigma_e^2). \end{aligned} \tag{3.2}$$

The weight in the mixture is the probability of having each genotype. Thus, we get the probability density function of a trait value, x, as

$$
\begin{aligned}
p(\mathrm{x}) &= \sum_{\mathrm{i}=1}^{3} \mathrm{P}(\mathrm{G_i}) \cdot p(\mathrm{x}|\mathrm{G_i}) \\
&= \mathrm{p}^2 \cdot f(\mathrm{x}|\mathrm{a}, \sigma_{\mathrm{e}}^2) + 2\mathrm{pq} \cdot f(\mathrm{x}|\mathrm{d}, \sigma_{\mathrm{e}}^2) + \mathrm{q}^2 \cdot f(\mathrm{x}|-\mathrm{a}, \sigma_{\mathrm{e}}^2)
\end{aligned}
\tag{3.3}
$$

where $f(\mathrm{x}|\mu, \sigma^2)$ is the pdf of normal distribution with mean μ and variance σ^2.

The genotype mean effect for each genotype can be computed analytically, given the allele frequency, heritability and inheritance type. In our simulation, we considered allele frequencies of 0.1 and 0.01, heritability values of 0.2 and 0.4, and three inheritance types where A_1 is dominant, additive, and recessive to A_2. The allele frequencies were chosen in order to simulate a trait with a rare allele frequency. The heritability value of 0.2 were chosen because they have been used by Cuenco et al. (2003). We considered heritability value of 0.4 as well. Heritability can be written in terms of allele frequency when we know the overall trait variance, as seen in equation (1.3.14)

$$
\mathrm{h}^2 = \frac{\sigma_{\mathrm{a}}^2}{\sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2} = \frac{2\mathrm{pq}\big(\mathrm{a} - \mathrm{d}(\mathrm{p} - \mathrm{q})\big)^2}{2\mathrm{pq}\big(\mathrm{a} - \mathrm{d}(\mathrm{p} - \mathrm{q})\big)^2 + 4\mathrm{p}^2\mathrm{q}^2\mathrm{d}^2 + \sigma_{\mathrm{e}}^2}. \tag{3.4}
$$

This equation can be rewritten in terms of genotype mean effect, a, as follows

$$
\mathrm{a}^2 = \begin{cases}
\dfrac{\mathrm{h}^2\sigma_{\mathrm{e}}^2}{(1-\mathrm{h}^2)2\mathrm{pq}} & \text{for } \mathrm{d} = 0 \text{ (additive)} \\[2ex]
\dfrac{\mathrm{h}^2\sigma_{\mathrm{e}}^2}{(1-\mathrm{h}^2)2\mathrm{pq}(1-\mathrm{p}+\mathrm{q})^2 - 4\mathrm{h}^2\mathrm{p}^2\mathrm{q}^2} & \text{for } \mathrm{d} = \mathrm{a} \text{ (dominant)} \\[2ex]
\dfrac{\mathrm{h}^2\sigma_{\mathrm{e}}^2}{(1-\mathrm{h}^2)2\mathrm{pq}(1+\mathrm{p}-\mathrm{q})^2 - 4\mathrm{h}^2\mathrm{p}^2\mathrm{q}^2} & \text{for } \mathrm{d} = -\mathrm{a} \text{ (recessive).}
\end{cases}
\tag{3.5}
$$

However, the given allele frequencies and heritability combinations result in a non-positive denominator when inheritance is recessive. Upon setting p=0.1 and p=0.01, we were required to set h^2 at less than 0.18 and 0.0198, respectively. Table 4.1 shows the genetic models considered in our simulation. For each model, we can calculated the genotype mean effects, a and d. Thus given p=0.1 and 0.01, there is no recessive model that would generate heritability as high as 0.2.

Let X_1 and X_2 be the trait values of a sibling pair. The overall mean and the overall variance of the trait values are computed in eq. (1.3.1) and (1.3.2) as

$$E(X_1) = E(X_2) = \mu_x = p^2a + 2pqd + q^2(-a) \qquad (1.3.1)$$

$$Var(X_1) = Var(X_2) = \sigma_x^2 = 2pq\big(a - d(p - q)\big)^2 + 4p^2q^2d^2 + \sigma_e^2. \qquad (1.3.2)$$

The conditional covariance between two siblings given their genotypes is computed as $\pi \cdot 2pq\big(a - d(p - q)\big)^2 + \Delta \cdot 4p^2q^2d^2 + \rho_e\sigma_e^2$, where π is the proportion of alleles IBD and Δ is the probability of sharing 2 alleles IBD at the trait locus. Then the marginal covariance due to genotype effect between two siblings is denoted as

$$Cov(X_1, X_2) = \rho\sigma_x^2 = pq\big(a - d(p - q)\big)^2 + p^2q^2d^2 + \rho_e\sigma_e^2 \qquad (3.6)$$

since the expected values of π and Δ are 0.5 and 0.25, respectively. One should note that the (polygenic) environmental correlation between siblings is zero at this point, i.e. all correlation between siblings is a result of the major quantitative trait locus.

Next, we can compute the standardized trait random variables Z_1 and Z_2 as follows

$$Z_1 = \frac{X_1 - E(X_1)}{\sqrt{Var(X_1)}} \text{ and } Z_2 = \frac{X_2 - E(X_2)}{\sqrt{Var(X_2)}} \tag{3.7}$$

so that

$$E(Z_1) = E(Z_2) = 0 \tag{3.8}$$

$$Var(Z_1) = Var(Z_2) = 1 \tag{3.9}$$

$$Cov(Z_1, Z_2) = \rho = \frac{pq\big(a - d(p-q)\big)^2 + p^2q^2d^2}{2pq\big(a - d(p-q)\big)^2 + 4p^2q^2d^2 + \sigma_e^2}. \tag{3.10}$$

In order to obtain additional correlation, ρ_e, so that the correlation between siblings Z_1 and Z_2 becomes a fixed value of ρ, Z_1 and Z_2into $Z_1^{'}$ and $Z_2^{'}$ can be transformed as follows

$$Z_1^{'} = Z_1 \tag{3.11}$$

$$Z_2^{'} = \frac{Z_2 + cZ_1}{\sqrt{Var(Z_2) + Var(cZ_1) + 2Cov(Z_2, cZ_1)}} = \frac{Z_2 + cZ_1}{\sqrt{1 + c^2 + 2c\rho_g}} \tag{3.12}$$

where $c = \sqrt{\frac{1-\rho_g^2}{\frac{1}{\rho^2}-1}} - \rho_g$ is a constant (Appendix A), used for this transformation for a given ρ, along with known $\rho_g = \frac{pq\left(a-d(p-q)\right)^2+p^2q^2d^2}{2pq\left(a-d(p-q)\right)^2+4p^2q^2d^2+\sigma_e^2}$ of the equation above, so that

$$E(Z_1^{'}) = E(Z_2^{'}) = 0 \tag{3.13}$$

$$Var(Z_1^{'}) = Var(Z_2^{'}) = 1 \tag{3.14}$$

$$Cov(Z_1^{'}, Z_2^{'}) = \rho. \tag{3.15}$$

Chapter 4 Results

In this section, we provide details about the genetic models we consider, as well as our findings on the power of the model-free methods with various parameter estimators.

Table 4.1 and Figure 4.1 show the details of each genetic model and the probability density function before and after standardization. Two allele frequencies (0.1 and 0.01), two heritabilities (0.2 and 0.4), and two inheritance types (Additive and Dominant) were considered as trait model generating parameter values. Based on these parameters, the genotype mean effect, variance and correlation, due to a single locus were computed. The trait value is the genotype mean effect, plus an environmental effect that follows a standard normal distribution. In Figure 4.1, the black line shows the probability density function under each model (Table 4.1). The blue dashed line shows the probability density function after standardization. The vertical red line is the threshold value used for truncate samples. We set the threshold at 1.28 in all cases. (The 90th percentile of standard normal equals to 1.28.) One should note that 1.28 is not the actual threshold for the top 10% of actual distribution of each model of mixture normals. However, in using truncate sampling for the top 10%, investors would assume a single normal distribution and hence take the threshold of single normal which is the approximate 1.28 standard deviation from mean. Finally, additional environmental correlation between siblings was added, so that all cases resulted in a correlation between siblings equal to 0.25. This approach was used in order to make our findings comparable to those reported by Cuenco et al. (2003) and Bhattacharjee et al. (2008).

Figures 4.2, 4.3 and 4.4 show that when allele frequency is rare (in this 0.01, Models 3, 4, 7 and 8), the models have higher correlations between the magnitude of trait differences and the magnitude of genotype differences.

Figure 4.2 shows the distribution of samples in one replicate under each model. This helps us to understand how the trait values of truncate samples and the proportion of alleles IBD will look like under each model. For rare allele frequency models (Models 3, 4, 7 and 8), the concordant pair samples (where trait values for both sibs in a pair are extremely high, in this case) have a higher proportion of alleles IBD (colored in purple) than other samples. Thus, when both siblings have high trait values, they tend to share more alleles IBD. This tendency is less obvious for models in which the allele frequency is higher, in this case 0.1 (Models 1, 2, 5 and 6). In that case, siblings sharing 0.25, 0.5 and 0.75 proportion of alleles IBD have relatively uniform trait values, at least in a replicate shown here.

Figures 4.3 and 4.4 show the same tendency for siblings to share more alleles IBD, when the magnitude of the siblings’ trait difference decreases, or when the sum of their values increases. Especially when the allele frequency is 0.01 (Models 3, 4, 7, and 8), the slope is steeper than when an allele frequency is 0.1 (Models 1, 2, 5, and 6). We can also see from Figures 4.3 and 4.4 that steeper slopes are associated with higher heritability, comparing Models 1, 3, 5, and 7 with Models 2, 4, 6, and 8, respectively.

Table 4.1. Description of Models 1-8

	Model 1	Model 2	Model 3	Model 4
Allele frequency	0.1	0.1	0.01	0.01
Heritability	0.2	0.4	0.2	0.4
Inheritance Type	Additive	Additive	Additive	Additive
Genotype Mean Effect	-1.78, 0, 1.78	-1.92, 0, 1.92	-3.55, 0, 3.55	-5.80, 0, 5.80
Variance due to Genotype	1.25	1.67	1.25	1.67
Correlation due to Genotype	0.1	0.2	0.1	0.2
After Standardization with Additional Environmental Correlation				
Overall Mean	0.0	0.0	0.0	0.0
Overall Variance	1.0	1.0	1.0	1.0
Overall Correlation	0.25	0.25	0.25	0.25
	Model 5	Model 6	Model 7	Model 8
Allele frequency	0.1	0.1	0.01	0.01
Heritability	0.2	0.4	0.2	0.4
Inheritance Type	Dominant	Dominant	Dominant	Dominant
Genotype Mean Effect	-0.66, 0.66, 0.66	-1.09, 1.09, 1.09	-1.80,1.80, 1.80	-2.94, 2.94, 2.94
Variance due to Genotype	1.27	1.73	1.25	1.67
Correlation due to Genotype	0.1	0.21	0.1	0.2
After Standardization with Additional Environmental Correlation				
Overall Mean	0.0	0.0	0.0	0.0
Overall Variance	1.0	1.0	1.0	1.0
Overall Correlation	0.25	0.25	0.25	0.25

Figure 4.1. Probability Density Function of a Quantitative Trait under Models 1-8. The trait value is the genotype mean effect, plus an environmental effect that follows a standard normal distribution. The black line shows the original probability density function under each model (Table 4.1). The blue dashed line shows the probability density function after standardization. The vertical red line is the threshold value used. We set the threshold at 1.28 in all cases. (The 90th percentile of standard normal equals to 1.28.) Although 1.28 is not the actual threshold for the top 10% of each distribution, we did this with the idea that in using truncate sampling for the top 10%, investors would assume a single normal distribution and hence, take approximately 1.28 standard deviations from the mean.

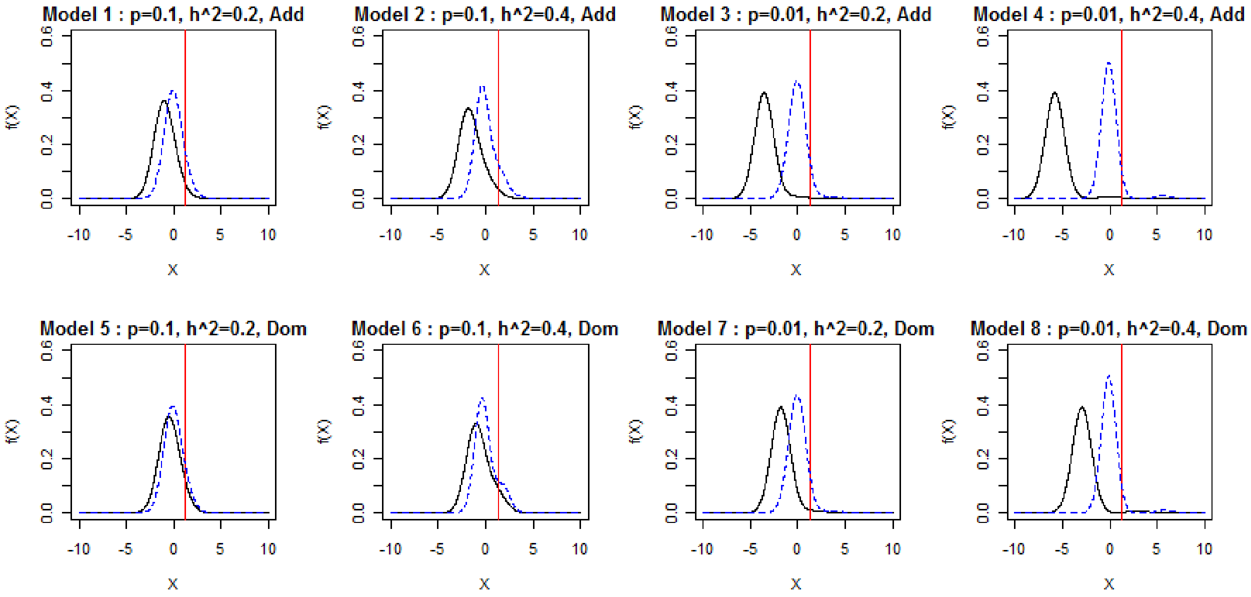


Figure 4.2. Trait Values vs. Proportion of Alleles IBD for Truncate Samples under Models 1-8 (1 Replicate, Sample Size=500). Each sample is represented by a point. The green, turquoise, blue, purple, and red colors of the point represent 0.0, 0.25, 0.5, 0.75, and 1.0 proportion of alleles IBD, respectively.

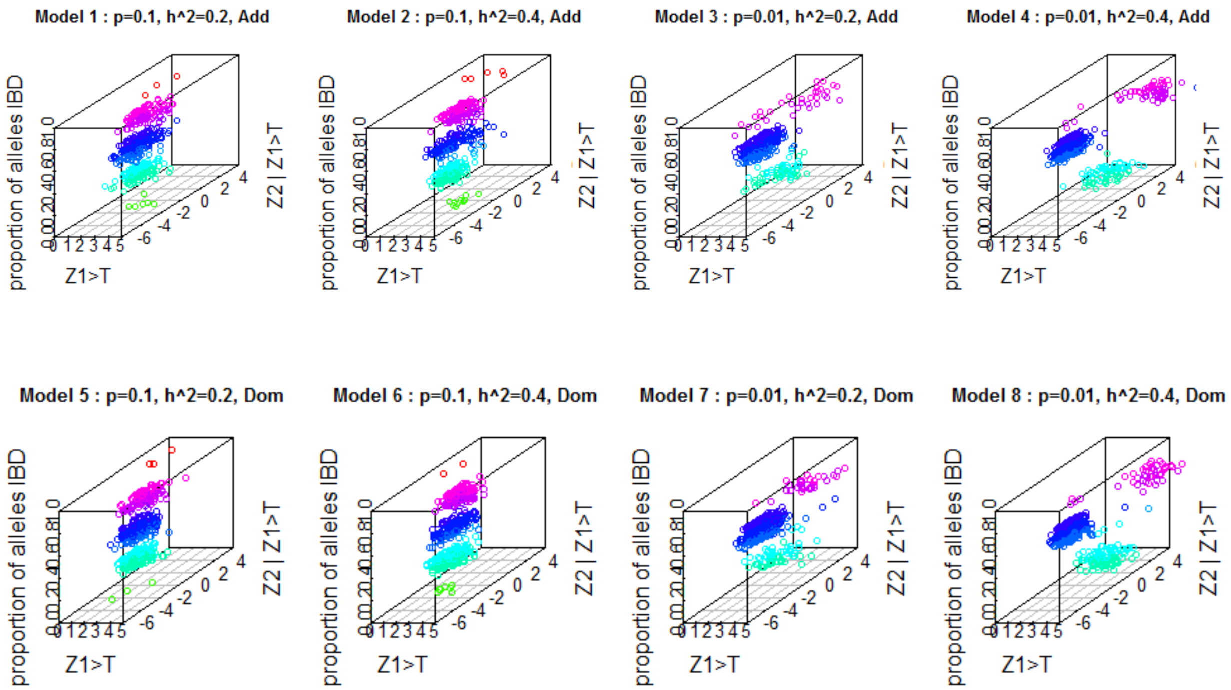


Figure 4.3. Squared Trait Differences vs. Proportion of Alleles IBD for Truncate Samples under Models 1-8 (1 Replicate, Sample Size=500). The simple regression line of squared trait differences on the proportion of alleles IBD, is shown as a red dashed line. The green, turquoise, blue, purple, and red points represent 0.0, 0.25, 0.5, 0.75, and 1.0 proportion of alleles IBD, respectively.

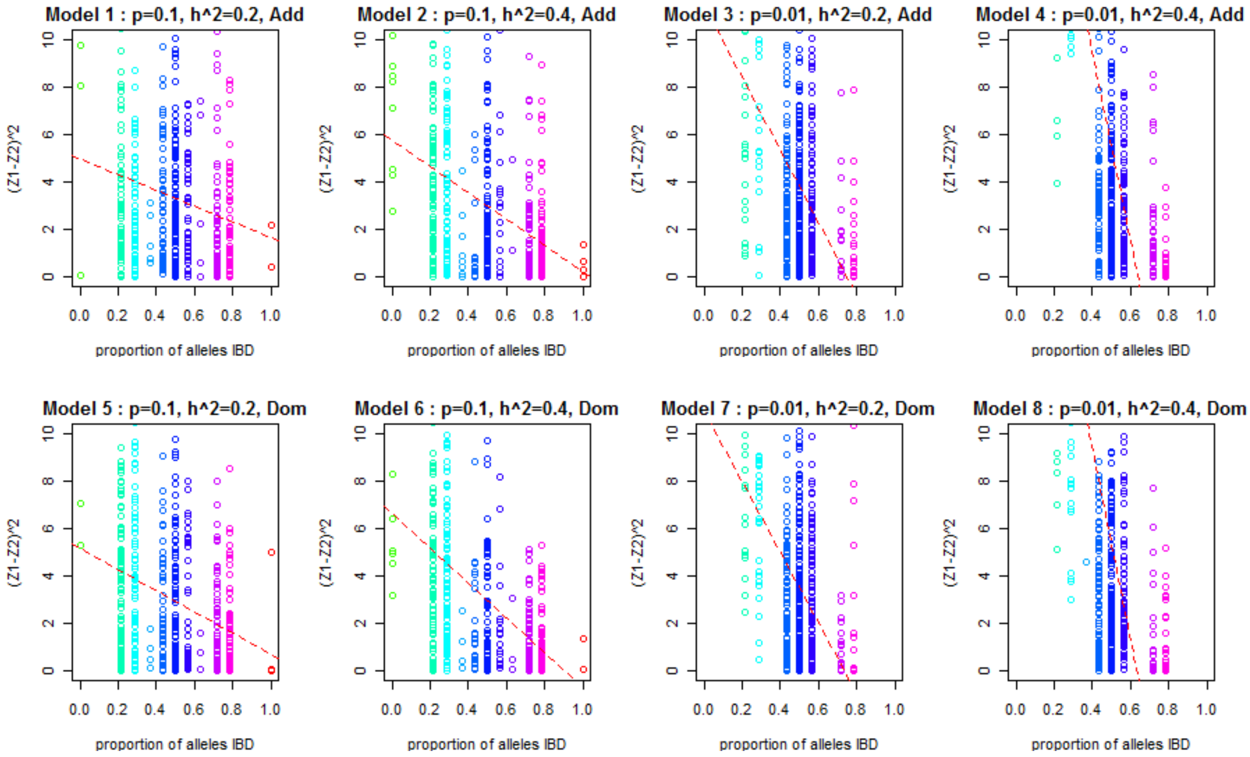
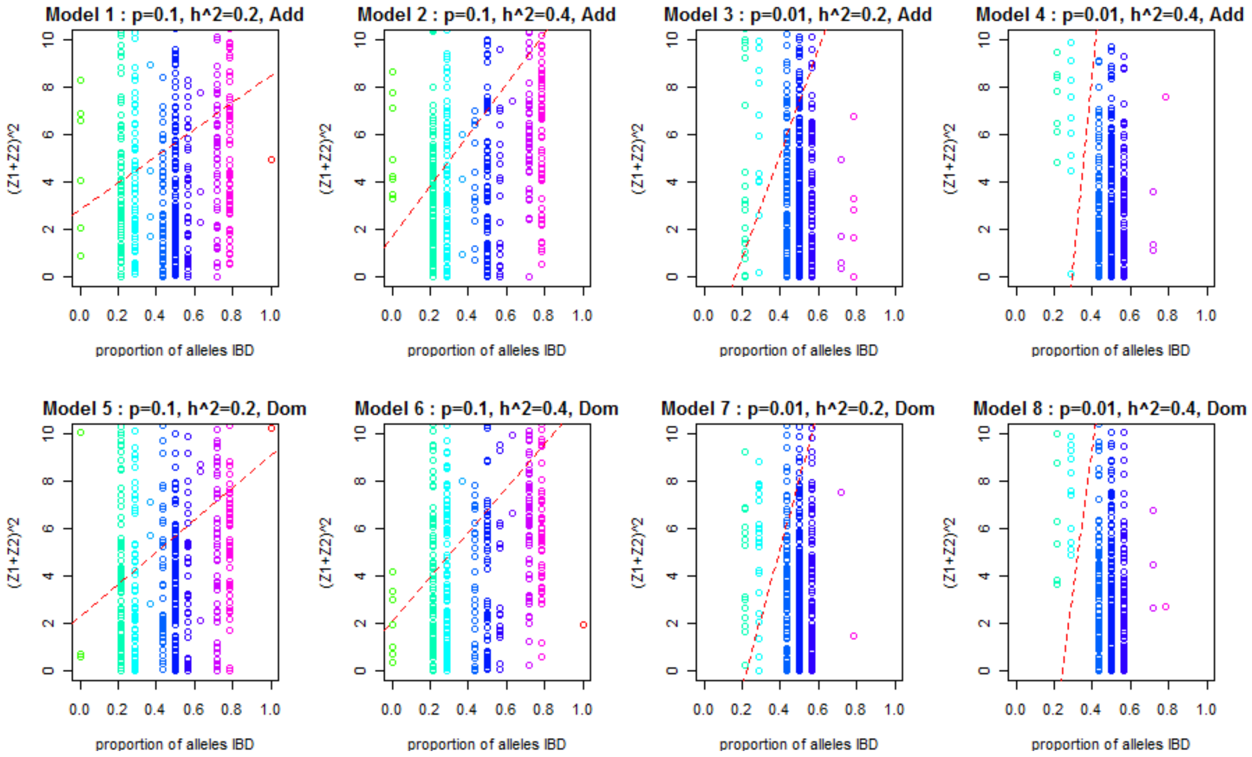


Figure 4.4. Squared Mean Corrected Trait Sum vs. Proportion of Alleles IBD for Truncate Samples under Models 1-8 (1 Replicate, Sample Size=500). The simple regression line of squared trait differences on the proportion of alleles IBD, is shown as a red dashed line. The green, turquoise, blue, purple, and red points represent 0.0, 0.25, 0.5, 0.75, and 1.0 proportion of alleles IBD, respectively.



In Table 4.2, we show the five different estimated parameter values for truncate samples, under each genetic model with different sample sizes. The best estimate is defined as the one having the smallest absolute value of difference with the true parameter value, and is marked in bold. (However, the correlation estimate should have both nonnegative and the minimum difference value, to be considered as the best estimate.) When two estimated values are not significantly different (paired t-test, $\alpha = 0.01$,) both are marked in bold.

Our estimators (Shin RM or Shin CA) give better mean estimates than other three estimators we consider, with only a few exceptions: Model 2 with a sample size of 200, Model 6 with a sample size of 200 and 500, Model 4 and Model 8. The Shin CA gives better variance estimates under Models 3, 4, 7 and 8. However, Peng's CMLEs give the best estimates under Model 2 with sample sizes of 200 and 500, and under Model 6. It should be noted that the sample moments give better mean estimates than other estimators, under Models 4 and 8. Our suggested estimators and Peng's CMLEs give the same correlation estimates because we used the regression slope between sibling trait values, as the unbiased correlation estimator (Section 2.2.4) and Peng's likelihood function is based on the conditional mean (Section 2.1.2).

Our major interest is evaluating different approaches for estimating trait parameter values in truncate samples that can be used in model-free linkage analysis. Before evaluating estimators, we will first address which of the model-free linkage analysis methods gives the greatest power. Figure 4.5 shows the power of five model-free methods by Haseman and Elston (1972, Original H & E), Xu et al. (2000, Xu), Sham and Purcell (2001, S & P), Tang and Siegmund (2001, T & S) and Cuenco et al. (2003

Cuenco). These methods were introduced in Section 1.5. Five estimators were introduced in Chapter 2, and they include two existing estimators (sample moments and Peng’s CMLEs), and three suggested approaches derived from truncate likelihood function: Shin’s application to Rao and Mendell (Shin RM), Shin’s application to Cohen and Aboueissa (Shin CA), and Shin’s application to the Newton-Raphson numerical algorithm (Shin Nu).

Table 4.2. Mean and Mean Squared Error of Estimated Trait Parameter Values under Models 1-8 (1000 Replicates, Sample Size=100, 200 and 500). The best estimates are printed in bold .

a. Model 1: $\mathrm{p} = 0.1, \mathrm{h}^2 = 0.2$, Additive

True Values	**$\mu = 0.0$**	**$\sigma^2 = 1.0$**	**$\rho = 0.25$**
Sample Size=100	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.07 (1.15)	**1.10 (0.02)**	-0.40 (0.42)
Peng’s Conditional MLEs	-0.86 (11.11)	1.18 (1.43)	**0.31 (0.06)**
Shin RM	**-0.13 (0.05)**	1.22 (0.12)	**0.31 (0.06)**
Shin CA	**-0.13 (0.03)**	**1.10 (0.02)**	**0.31 (0.06)**
Shin Nu	-2.64 (48.59)	2.47 (13.08)	**0.31 (0.06)**
Sample Size=200	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.07 (1.15)	1.10 (0.01)	-0.39 (0.42)
Peng’s Conditional MLEs	-0.40 (0.51)	**1.03 (0.03)**	**0.30 (0.03)**
Shin RM	-0.14 (0.04)	1.23 (0.09)	**0.30 (0.03)**
Shin CA	**-0.13 (0.02)**	1.10 (0.01)	**0.30 (0.03)**
Shin Nu	-1.69 (22.20)	2.01 (6.28)	**0.30 (0.03)**
Sample Size=500	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.07 (1.14)	1.10 (0.01)	-0.39 (0.41)
Peng’s Conditional MLEs	-0.34 (0.22)	**1.01 (0.01)**	**0.30 (0.01)**
Shin RM	-0.15 (0.03)	1.24 (0.07)	**0.30 (0.01)**
Shin CA	**-0.13 (0.02)**	1.10 (0.01)	**0.30 (0.01)**
Shin Nu	-0.99 (5.21)	1.65 (1.57)	**0.30 (0.01)**

b. Model 2: $\mathrm{p} = 0.1, \mathrm{h}^2 = 0.4$, Additive

True Values	**$\mu = 0.0$**	**$\sigma^2 = 1.0$**	**$\rho = 0.25$**
Sample Size=100	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.14 (1.30)	**1.24 (0.07)**	-0.31 (0.32)
Peng’s Conditional MLEs	-0.80 (13.07)	1.38 (1.45)	**0.30 (0.05)**
Shin RM	**-0.33 (0.15)**	1.61 (0.49)	**0.30 (0.05)**
Shin CA	-0.35 (0.14)	1.27 (0.08)	**0.30 (0.05)**
Shin Nu	-2.46 (48.58)	2.84 (18.10)	**0.30 (0.05)**
Sample Size=200	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.14 (1.30)	**1.25 (0.07)**	-0.31 (0.31)
Peng’s Conditional MLEs	**-0.36 (0.59)**	**1.24 (0.12)**	**0.30 (0.02)**
Shin RM	**-0.35 (0.14)**	1.62 (0.44)	**0.30 (0.02)**
Shin CA	**-0.35 (0.13)**	1.27 (0.08)	**0.30 (0.02)**
Shin Nu	-1.37 (14.01)	2.22 (5.78)	**0.30 (0.02)**
Sample Size=500	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.14 (1.30)	1.25 (0.06)	-0.30 (0.31)
Peng’s Conditional MLEs	**-0.29 (0.18)**	**1.21 (0.06)**	**0.30 (0.01)**
Shin RM	-0.36 (0.14)	1.63 (0.42)	**0.30 (0.01)**
Shin CA	-0.35 (0.12)	1.27 (0.08)	**0.30 (0.01)**
Shin Nu	-0.96 (5.30)	1.97 (2.44)	**0.30 (0.01)**

c. Model 3: $p = 0.01, h^2 = 0.2,$ Additive

True Values	$\boldsymbol{\mu = 0.0}$		$\boldsymbol{\sigma^2 = 1.0}$		$\boldsymbol{\rho = 0.25}$	
Sample Size=100	$\bar{\hat{\mu}}$ (MSE)		$\overline{\hat{\sigma}^2}$ (MSE)		$\bar{\hat{\rho}}$ (MSE)	
Sample Moments	1.16	(1.35)	1.65	(0.47)	-0.13	(0.15)
Peng's Conditional MLEs	-5.43	(458.96)	3.33	(79.45)	**0.55**	**(0.13)**
Shin RM	-1.33	(1.90)	3.58	(7.42)	**0.55**	**(0.13)**
Shin CA	**-0.74**	**(0.59)**	**1.56**	**(0.35)**	**0.55**	**(0.13)**
Shin Nu	-32.74	(1136.95)	24.52	(610.18)	**0.55**	**(0.13)**
Sample Size=200	$\bar{\hat{\mu}}$ (MSE)		$\overline{\hat{\sigma}^2}$ (MSE)		$\bar{\hat{\rho}}$ (MSE)	
Sample Moments	1.16	(1.35)	1.65	(0.45)	-0.12	(0.14)
Peng's Conditional MLEs	-2.58	(55.69)	2.20	(11.68)	**0.55**	**(0.11)**
Shin RM	-1.34	(1.84)	3.55	(6.88)	**0.55**	**(0.11)**
Shin CA	**-0.73**	**(0.56)**	**1.56**	**(0.33)**	**0.55**	**(0.11)**
Shin Nu	-33.02	(1134.42)	24.46	(582.59)	**0.55**	**(0.11)**
Sample Size=500	$\bar{\hat{\mu}}$ (MSE)		$\overline{\hat{\sigma}^2}$ (MSE)		$\bar{\hat{\rho}}$ (MSE)	
Sample Moments	1.16	(1.35)	1.66	(0.44)	-0.11	(0.31)
Peng's Conditional MLEs	-1.81	(4.10)	1.86	(0.91)	**0.55**	**(0.10)**
Shin RM	-1.35	(1.86)	3.58	(6.81)	**0.55**	**(0.10)**
Shin CA	**-0.74**	**(0.55)**	**1.56**	**(0.32)**	**0.55**	**(0.10)**
Shin Nu	-33.12	(1137.74)	24.44	(575.19)	**0.55**	**(0.10)**

d. Model 4: $p = 0.01, h^2 = 0.4,$ Additive

True Values	$\boldsymbol{\mu = 0.0}$		$\boldsymbol{\sigma^2 = 1.0}$		$\boldsymbol{\rho = 0.25}$	
Sample Size=100	$\bar{\hat{\mu}}$ (MSE)		$\overline{\hat{\sigma}^2}$ (MSE)		$\bar{\hat{\rho}}$ (MSE)	
Sample Moments	**1.44**	**(2.09)**	3.19	(4.95)	**0.00**	**(0.07)**
Peng's Conditional MLEs	-2.41	(60.94)	3.23	(30.36)	0.52	(0.09)
Shin RM	-3.44	(11.98)	11.24	(107.89)	0.52	(0.09)
Shin CA	-2.19	(4.91)	**2.68**	**(2.90)**	0.52	(0.09)
Shin Nu	-56.73	(3370.83)	69.93	(5190.21)	0.52	(0.09)
Sample Size=200	$\bar{\hat{\mu}}$ (MSE)		$\overline{\hat{\sigma}^2}$ (MSE)		$\bar{\hat{\rho}}$ (MSE)	
Sample Moments	**1.43**	**(2.05)**	3.18	(4.82)	**0.00**	**(0.06)**
Peng's Conditional MLEs	-1.81	(4.20)	2.85	(3.96)	0.51	(0.08)
Shin RM	-3.46	(12.04)	11.23	(106.06)	0.51	(0.08)
Shin CA	-2.16	(4.74)	**2.66**	**(2.79)**	0.51	(0.08)
Shin Nu	-56.02	(3242.04)	67.48	(4647.69)	0.51	(0.08)
Sample Size=500	$\bar{\hat{\mu}}$ (MSE)		$\overline{\hat{\sigma}^2}$ (MSE)		$\bar{\hat{\rho}}$ (MSE)	
Sample Moments	**1.43**	**(2.06)**	3.19	(4.83)	**0.00**	**(0.06)**
Peng's Conditional MLEs	-1.67	(3.03)	2.80	(3.40)	0.51	(0.07)
Shin RM	-3.49	(12.18)	11.32	(107.14)	0.51	(0.07)
Shin CA	-2.17	(4.75)	**2.67**	**(2.80)**	0.51	(0.07)
Shin Nu	-56.41	(3294.44)	67.9	(4679.71)	0.51	(0.07)

e. Model 5: $p = 0.1, h^2 = 0.2, \mathrm{Dominant}$

True Values	$\boldsymbol{\mu = 0.0}$	$\boldsymbol{\sigma^2 = 1.0}$	$\boldsymbol{\rho = 0.25}$
Sample Size=100	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.07 (1.14)	1.07 (0.02)	-0.42 (0.46)
Peng's Conditional MLEs	-0.51 (8.76)	**1.04 (0.65)**	**0.26 (0.05)**
Shin RM	**-0.07 (0.03)**	1.14 (0.07)	**0.26 (0.05)**
Shin CA	-0.10 (0.02)	1.08 (0.01)	**0.26 (0.05)**
Shin Nu	-0.92 (15.62)	1.56 (3.93)	**0.26 (0.05)**
Sample Size=200	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.06 (1.14)	1.08 (0.01)	-0.42 (0.45)
Peng's Conditional MLEs	-0.26 (0.35)	**0.98 (0.03)**	**0.25 (0.02)**
Shin RM	**-0.07 (0.02)**	1.14 (0.04)	**0.25 (0.02)**
Shin CA	-0.10 (0.02)	1.08 (0.01)	**0.25 (0.02)**
Shin Nu	-0.49 (6.76)	1.34 (1.81)	**0.25 (0.02)**
Sample Size=500	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.07 (1.14)	1.07 (0.01)	-0.41 (0.44)
Peng's Conditional MLEs	-0.19 (0.11)	**0.96 (0.01)**	**0.26 (0.01)**
Shin RM	**-0.08 (0.01)**	1.15 (0.03)	**0.26 (0.01)**
Shin CA	-0.10 (0.01)	1.08 (0.01)	**0.26 (0.01)**
Shin Nu	-0.17 (2.38)	1.20 (0.86)	**0.26 (0.01)**

f. Model 6: $p = 0.1, h^2 = 0.4, \mathrm{Dominant}$

True Values	$\boldsymbol{\mu = 0.0}$	$\boldsymbol{\sigma^2 = 1.0}$	$\boldsymbol{\rho = 0.25}$
Sample Size=100	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.13 (1.27)	**1.18 (0.05)**	-0.36 (0.38)
Peng's Conditional MLEs	-0.18 (2.87)	1.19 (0.44)	**0.18 (0.06)**
Shin RM	**-0.09 (0.03)**	1.24 (0.11)	**0.18 (0.06)**
Shin CA	-0.25 (0.08)	1.20 (0.05)	**0.18 (0.06)**
Shin Nu	0.76 (2.16)	0.84 (0.48)	**0.18 (0.06)**
Sample Size=200	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.13 (1.27)	1.17 (0.04)	-0.36 (0.38)
Peng's Conditional MLEs	**0.00 (0.17)**	**1.12 (0.03)**	**0.18 (0.03)**
Shin RM	-0.10 (0.02)	1.26 (0.09)	**0.18 (0.03)**
Shin CA	-0.25 (0.07)	1.20 (0.04)	**0.18 (0.03)**
Shin Nu	0.84 (1.02)	0.80 (0.14)	**0.18 (0.03)**
Sample Size=500	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.13 (1.27)	1.18 (0.03)	-0.36 (0.38)
Peng's Conditional MLEs	**0.04 (0.06)**	**1.10 (0.02)**	**0.18 (0.02)**
Shin RM	-0.10 (0.01)	1.25 (0.07)	**0.18 (0.02)**
Shin CA	-0.25 (0.06)	1.20 (0.04)	**0.18 (0.02)**
Shin Nu	0.92 (0.93)	0.75 (0.09)	**0.18 (0.02)**

g. Model 7: $\mathrm{p} = 0.01, \mathrm{h}^2 = 0.2,$ Dominant

True Values	$\boldsymbol{\mu = 0.0}$	$\boldsymbol{\sigma^2 = 1.0}$	$\boldsymbol{\rho = 0.25}$
Sample Size=100	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.16 (1.35)	1.66 (0.47)	-0.13 (0.15)
Peng's Conditional MLEs	-6.18 (421.07)	3.63 (75.7)	**0.56 (0.14)**
Shin RM	-1.30 (1.82)	3.52 (7.06)	**0.56 (0.14)**
Shin CA	**-0.73 (0.59)**	**1.56 (0.34)**	**0.56 (0.14)**
Shin Nu	-32.61 (1135.99)	24.46 (612.17)	**0.56 (0.14)**
Sample Size=200	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.16 (1.36)	1.66 (0.45)	-0.12 (0.14)
Peng's Conditional MLEs	-2.09 (10.58)	1.99 (2.6)	**0.55 (0.11)**
Shin RM	-1.35 (1.87)	3.59 (7.05)	**0.55 (0.11)**
Shin CA	**-0.74 (0.58)**	**1.57 (0.34)**	**0.55 (0.11)**
Shin Nu	-33.56 (1172.47)	25.02 (614.26)	**0.55 (0.11)**
Sample Size=500	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	1.16 (1.35)	1.66 (0.44)	-0.11 (0.13)
Peng's Conditional MLEs	-1.76 (3.74)	1.84 (0.83)	**0.55 (0.1)**
Shin RM	-1.36 (1.86)	3.59 (6.86)	**0.55 (0.1)**
Shin CA	**-0.74 (0.56)**	**1.57 (0.33)**	**0.55 (0.1)**
Shin Nu	-33.25 (1147.46)	24.63 (585.14)	**0.55 (0.1)**

h. Model 8: $\mathrm{p} = 0.01, \mathrm{h}^2 = 0.4,$ Dominant

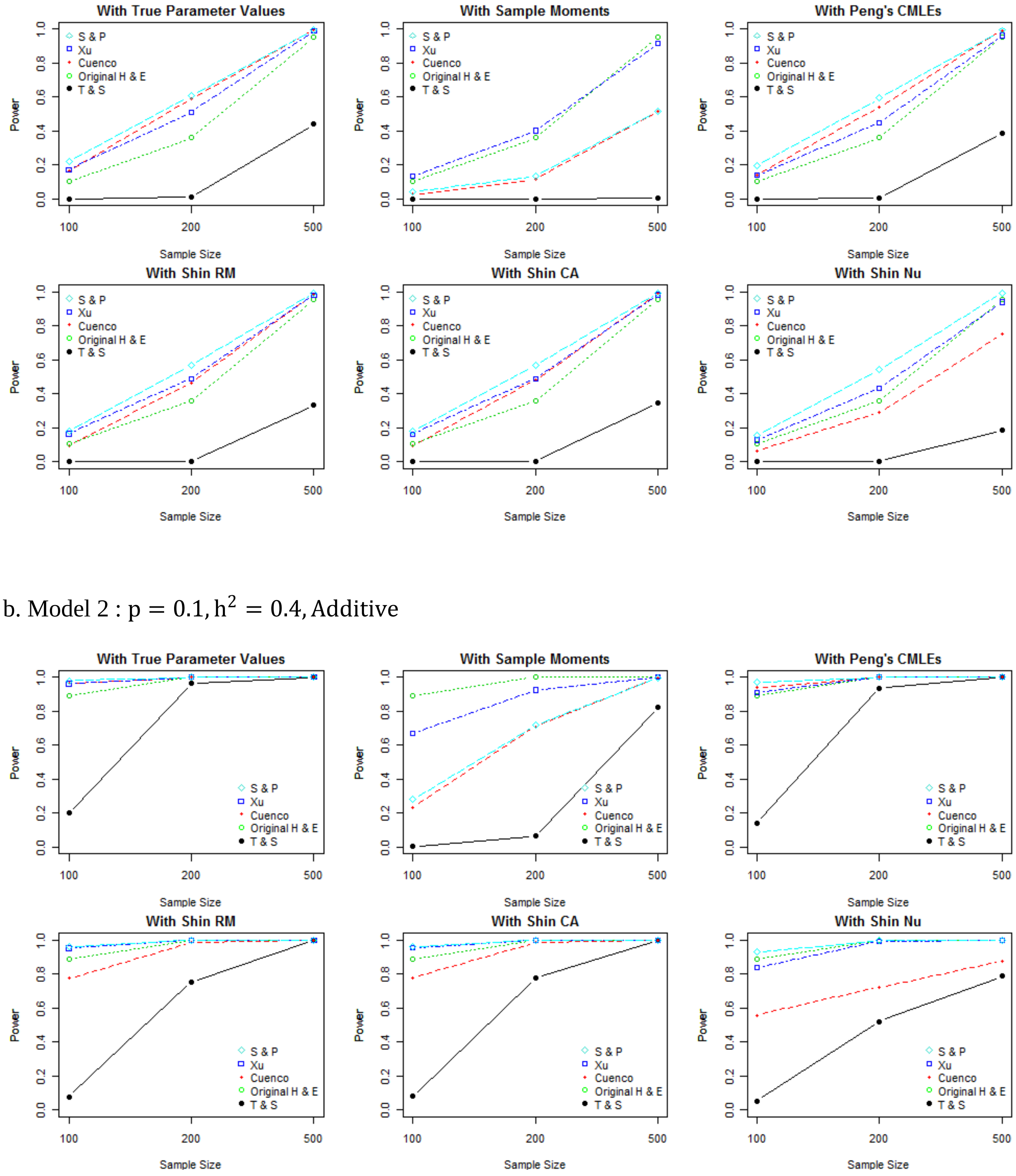
True Values	$\boldsymbol{\mu = 0.0}$	$\boldsymbol{\sigma^2 = 1.0}$	$\boldsymbol{\rho = 0.25}$
Sample Size=100	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	**1.44 (2.08)**	3.20 (4.98)	**0.00 (0.07)**
Peng's Conditional MLEs	-2.47 (49.97)	3.24 (28.2)	0.52 (0.09)
Shin RM	-3.47 (12.14)	11.34 (109.18)	0.52 (0.09)
Shin CA	-2.20 (4.95)	**2.69 (2.92)**	0.52 (0.09)
Shin Nu	-56.6 (3346.67)	69.76 (5157.91)	0.52 (0.09)
Sample Size=200	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	**1.44 (2.08)**	3.21 (4.93)	**0.00 (0.06)**
Peng's Conditional MLEs	-1.81 (4.18)	2.87 (4.07)	0.51 (0.08)
Shin RM	-3.48 (12.16)	11.35 (108.23)	0.51 (0.08)
Shin CA	-2.20 (4.88)	**2.69 (2.88)**	0.51 (0.08)
Shin Nu	-56.93 (3358.31)	69.47 (4961.45)	0.51 (0.08)
Sample Size=500	$\bar{\hat{\mu}}$ (MSE)	$\overline{\hat{\sigma}^2}$ (MSE)	$\bar{\hat{\rho}}$ (MSE)
Sample Moments	**1.44 (2.08)**	3.21 (4.92)	**0.00 (0.06)**
Peng's Conditional MLEs	-1.64 (2.94)	2.77 (3.29)	0.51 (0.07)
Shin RM	-3.50 (12.29)	11.44 (109.51)	0.51 (0.07)
Shin CA	-2.21 (4.91)	**2.70 (2.89)**	0.51 (0.07)
Shin Nu	-57.7 (3432.47)	70.29 (4993.07)	0.51 (0.07)

Figure 4.5 presents the additive models with an allele frequency of 0.1. When true parameter values, Peng’s CMLEs, Shin RM, and Shin CA values are used, three methods (S & P, Cuenco, and Xu) work better than the other two (Original H & E and T & S), as seen in Figure 4.5.a and 4.5.b. When the sample moments are used, the original H & E works better than S & P and Cuenco. With the Shin Nu estimator, the original H & E method works better than the Cuenco method. Among the best three methods (Xu, S & P, and Cuenco), the Cuenco method is more sensitive than the others to the biased parameter values used, such as sample moments and Shin Nu estimates. Overall, the T & S method shows the least power regardless of estimating methods. Thus this one was dropped from further study. Also, the original H & E method will not be considered in our evaluation study because it was insensitive to the estimating methods. Figure 4.5 also shows that power increases, as heritability increases from 0.2 to 0.4 (Figure 4.5.b). Most model-free methods (except T & S) have power close to 0.8, even with sample sizes of 100 if one of the three estimators (Peng’s CMLEs, Shin RM and Shin CA) or the true parameter values are used.

Figures 4.5.c and 4.5.d show power under Models 3 and 4 that is greater than that of Models 1 and 2, respectively, when allele frequency is reduced from 0.1 to 0.01. The increase in heritability from 0.2 to 0.4 also results in greater power. In these rare allele frequency models, the Xu, S & P, and Original H & E methods work better than the other methods. The Cuenco method does not work well when the Shin Nu, Shin RM, and Shin CA estimators are used.

Based on Figures 4.5.e, f, g and h, we chose to use the Xu, S & P, and Cuenco methods, in our evaluation of the effects of different estimators on power under dominant models. The reason that the Xu, S & P, and Cuenco methods are chosen is because T & S shows the lowest power, and the original H & E is not affected by the trait parameter values used, even though the original H & E method is one of the estimators giving greater power, especially with very rare frequencies.

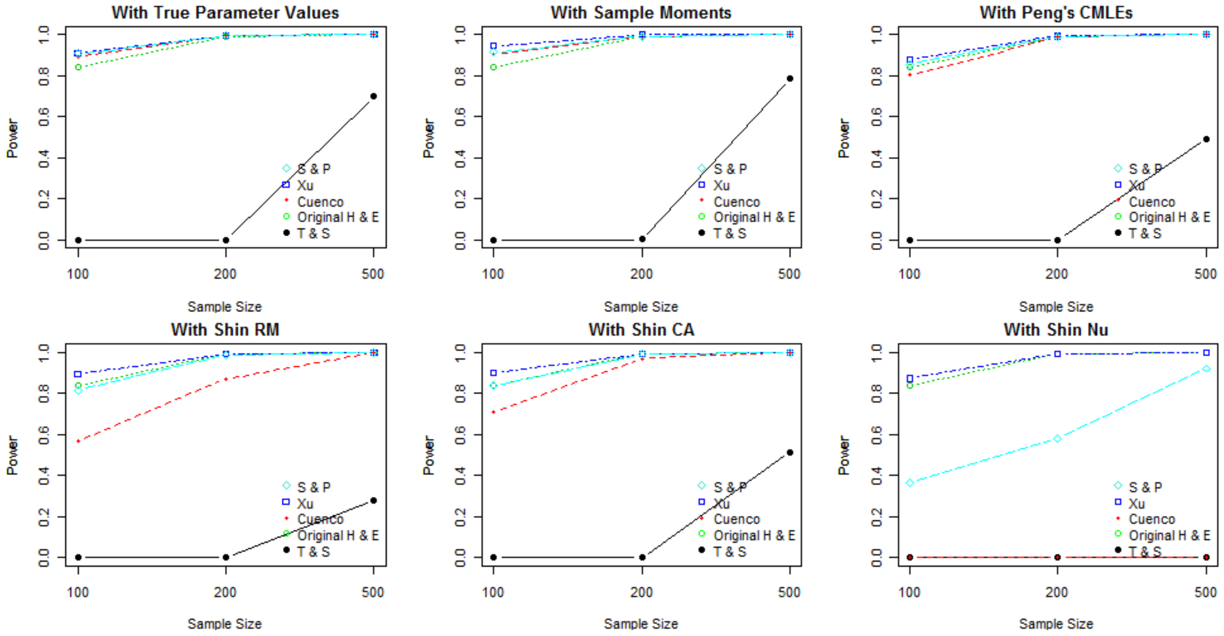
Figure 4.5. Comparison of Model-Free Methods using Five Different Estimators as well as True Parameter Values, for Truncate Samples under Models 1-8 (1000 Replicates, Sample Size=100, 200 and 500). Power is calculated for $\alpha = 0.0001$, as the proportion of replicates having a LOD score greater than 3.0 out of 1000 replicates.



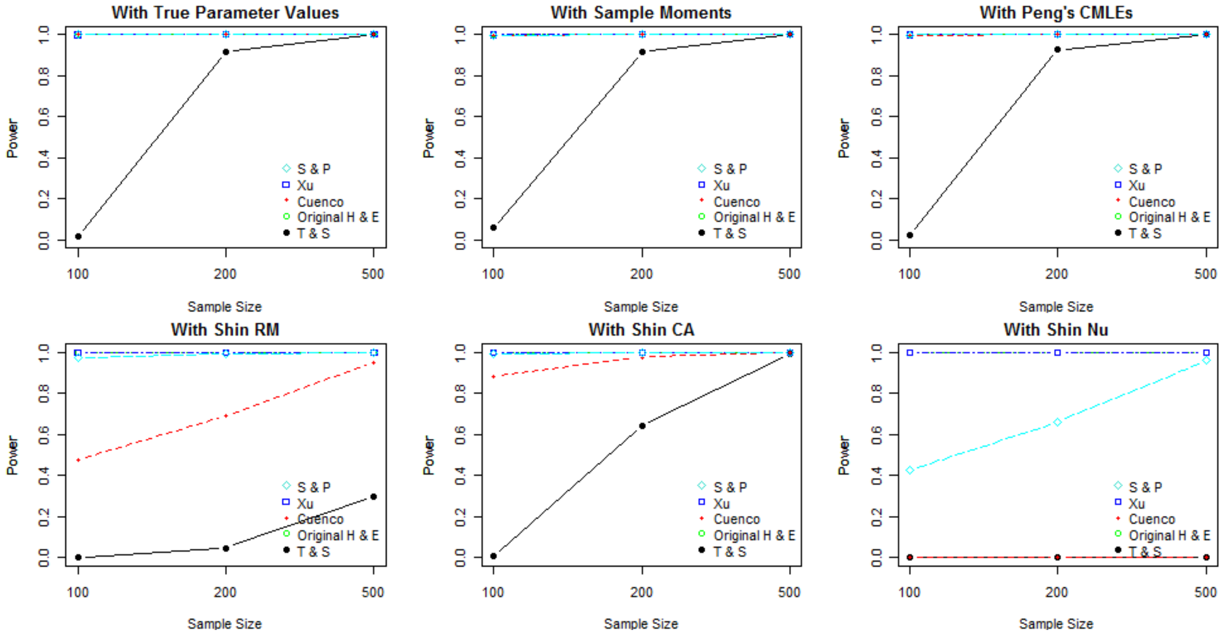
a. Model 1: $p = 0.1, h^2 = 0.2$, Additive

b. Model 2 : $p = 0.1, h^2 = 0.4$, Additive

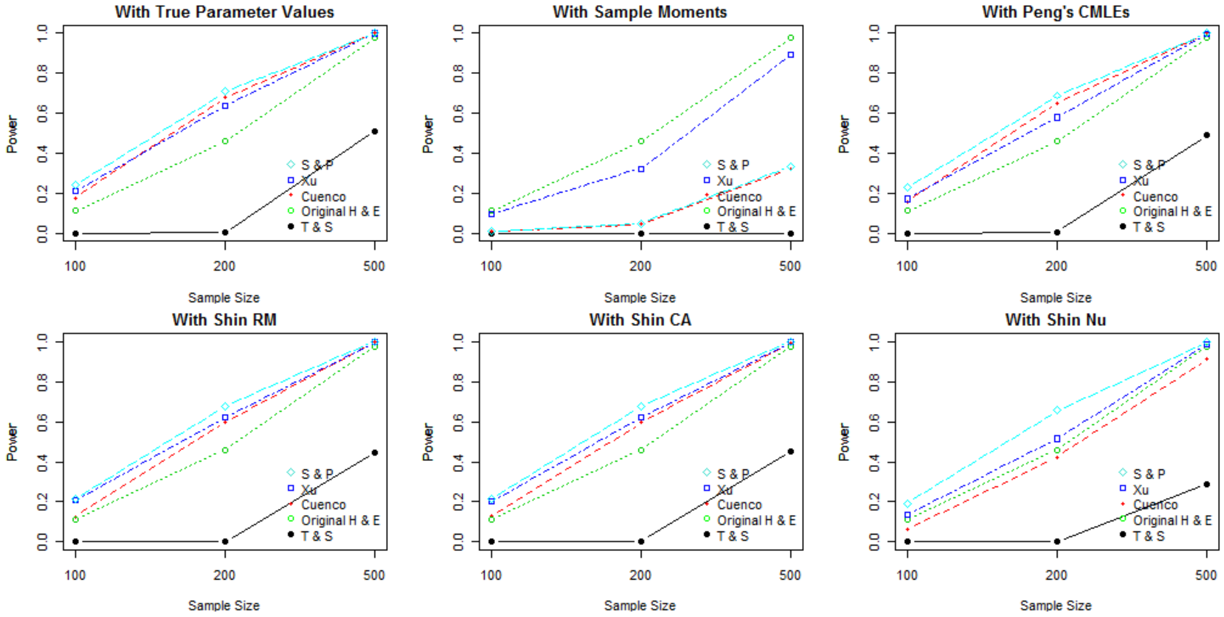
c. Model 3: $p = 0.01, h^2 = 0.2$, Additive



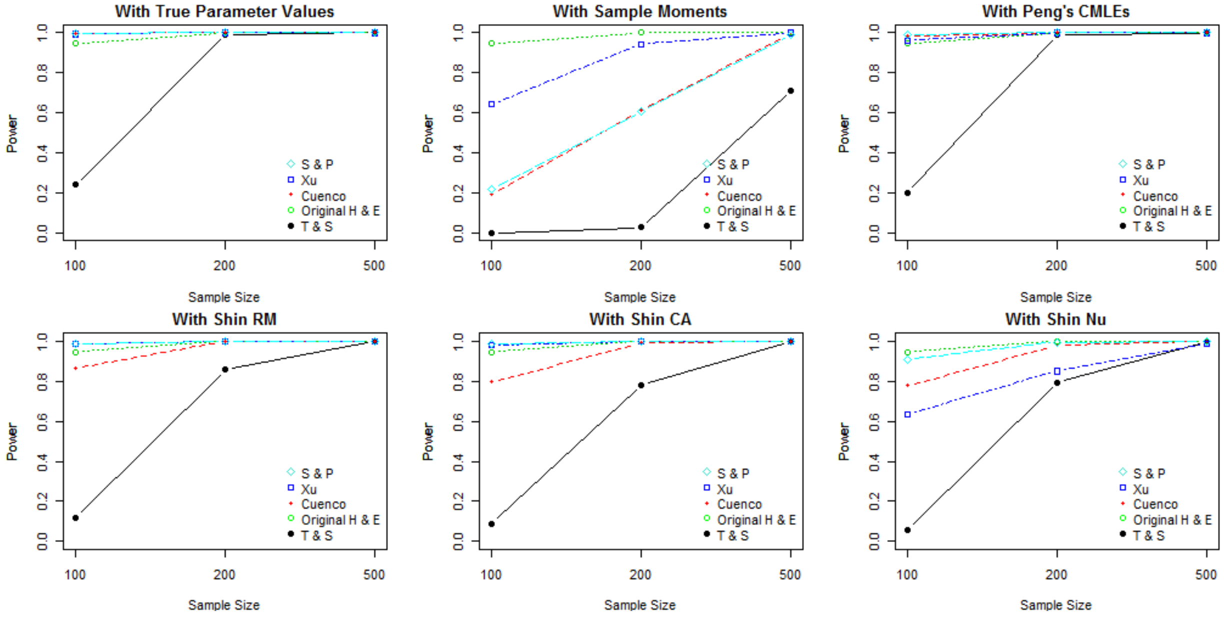
d. Model 4: $p = 0.01, h^2 = 0.4$, Additive



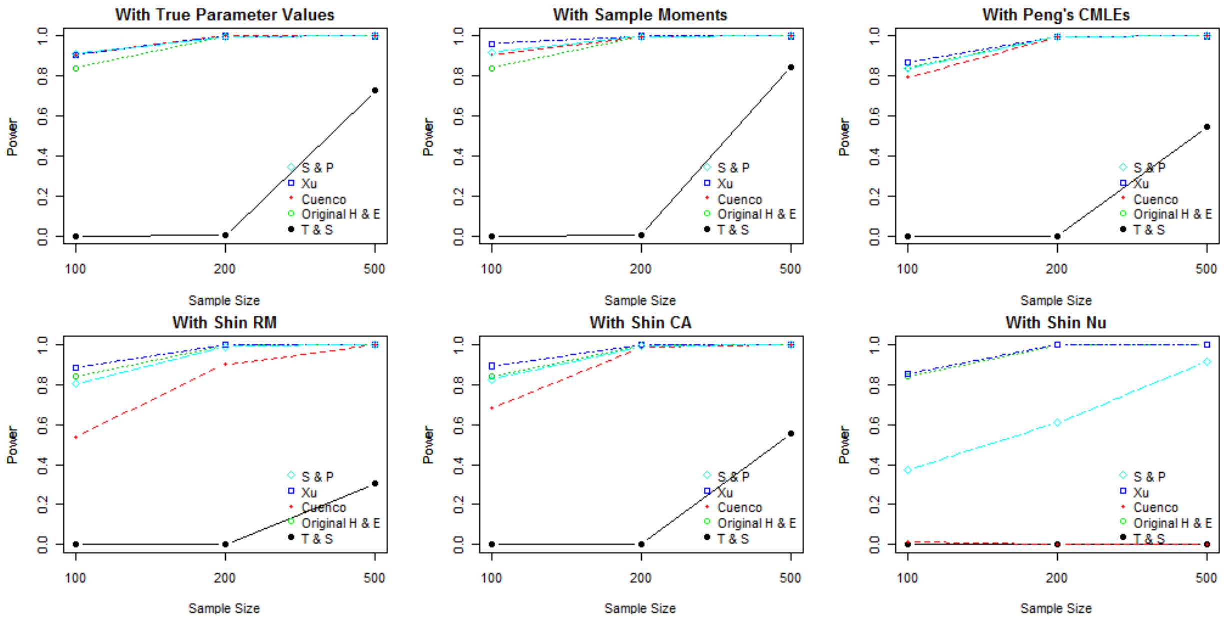
e. Model 5: $p = 0.1, h^2 = 0.2$, Dominant



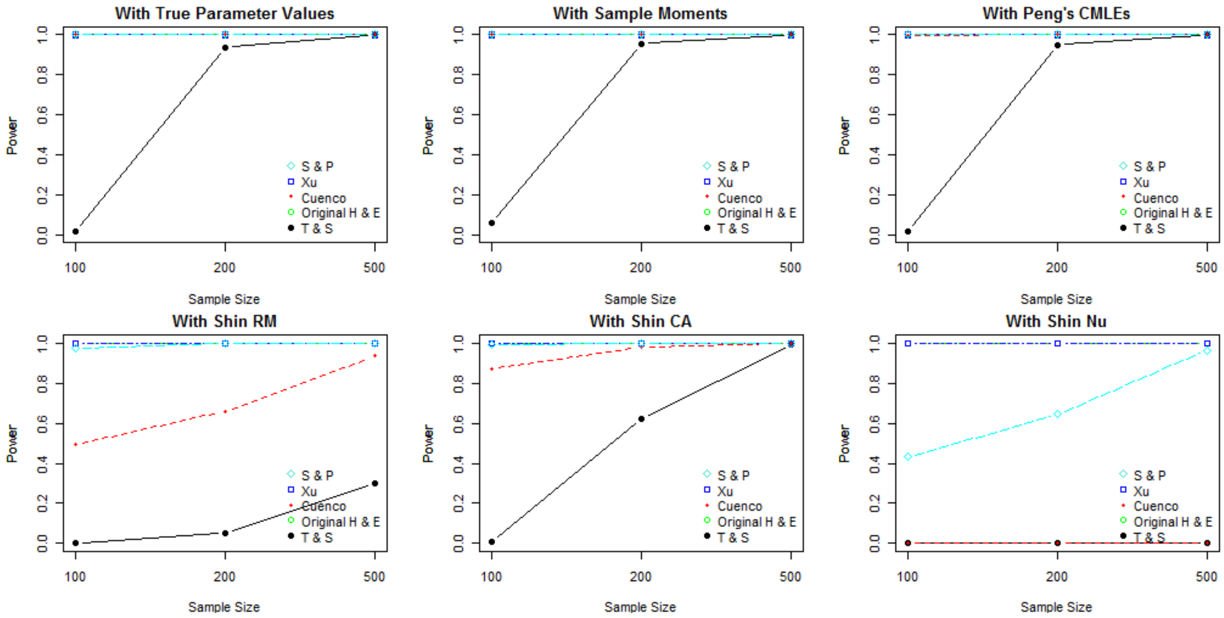
f. Model 6: $p = 0.1, h^2 = 0.4$, Dominant



g. Model 7: $p = 0.01, h^2 = 0.2$, Dominant



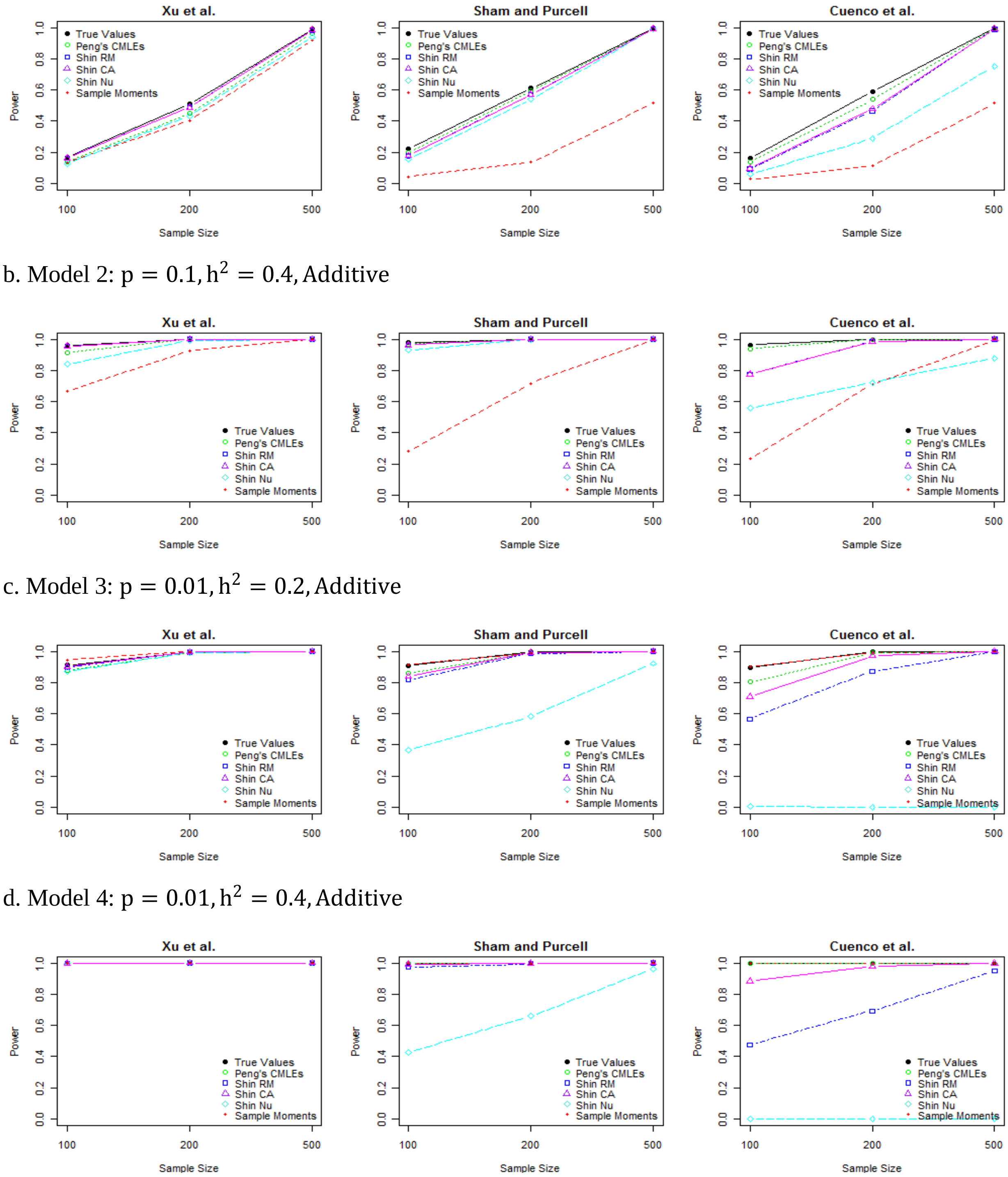
h. Model 8: $p = 0.01, h^2 = 0.4$, Dominant



We now compare the five estimators on the Xu, S & P, and Cuenco model-free linkage analysis methods. As seen in Figure 4.6, power using sample moments is lower than power with other estimators under the models with an allele frequency of 0.1 (Models 1, 2, 5, and 6 are shown in Figures 4.6.a, b, e, and f, respectively) except for application of Cuenco's method with a sample size of 500 where sample moments work better than Shin-Nu. Under the rare frequency models (Models 3, 4, 7 and 8) the Shin Nu estimator has less power than the other estimators. This is particularly the case for the Cuenco method. The other estimators (Peng's CMLEs, Shin RM, and Shin CA) work very well under all the models considered, especially with sample sizes greater than 200, except when the Shin RM estimator was applied to the Cuenco method under Models 4 and 8. One should note that the sample moments work as well as the true parameter values under the rare frequency models. The numerical values shown in Figure 4.6 are shown in Appendix B.

In order to further investigate differences among the estimating approaches, we performed a paired t-test for LOD scores and a McNemar's test for the significance of changes in the LOD > 3.0 between Peng's CMLEs and each of our two estimators (Shin RM and Shin CA). The results show that our estimators give greater power than Peng's estimator for Xu's method, under all models except Model 1 with a sample size of 500 and Model 3 with a sample size of 200. Also our estimators give greater LODs for Xu's method, except for Models 2 and 6 with sample sizes of 500, and Models 4 and 8. However, if the sample size equals 500 or if the model is based on a rare allele frequency, the power of all three methods (Xu, S & P, and Cuenco) reaches almost 100%, except when the Shin RM estimator is applied to Cuenco’s method. The details are shown in Appendices C and D.

Figure 4.6. Comparison of Five Different Trait Parameter Estimators as well as True Parameter Values on Three Model-Free Methods, for Truncate Samples under Models 1-8 (1000 Replicates, Sample Size=100, 200 and 500). Power is calculated for $\alpha = 0.0001$.



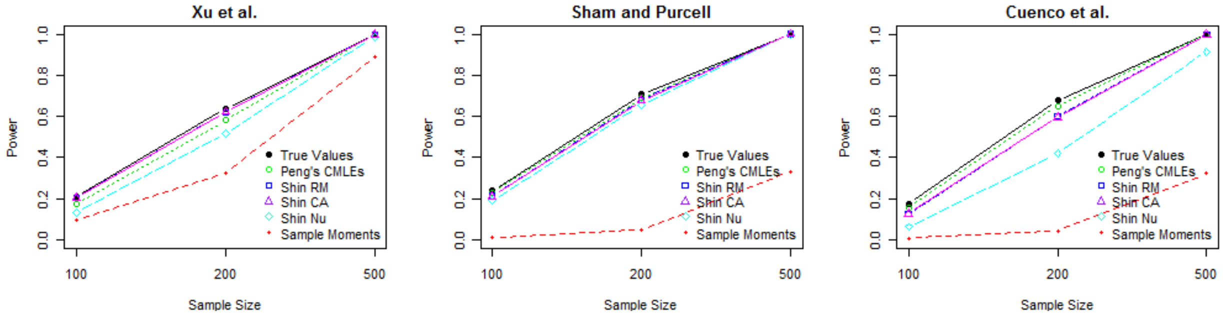
a. Model 1: $p = 0.1, h^2 = 0.2$, Additive

b. Model 2: $p = 0.1, h^2 = 0.4$, Additive

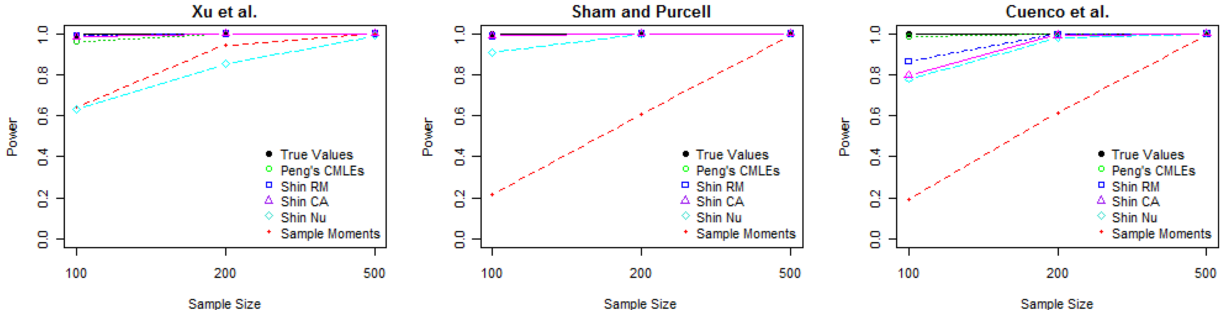
c. Model 3: $p = 0.01, h^2 = 0.2$, Additive

d. Model 4: $p = 0.01, h^2 = 0.4$, Additive

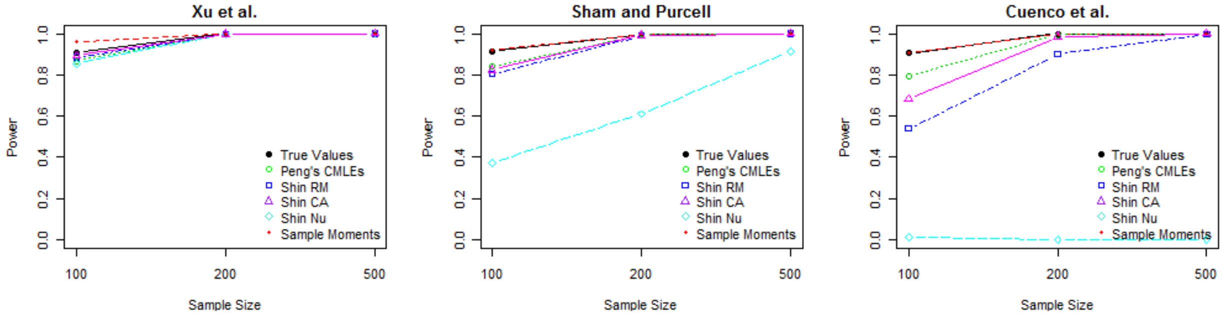
e. Model 5: $p = 0.1, h^2 = 0.2$, Dominant



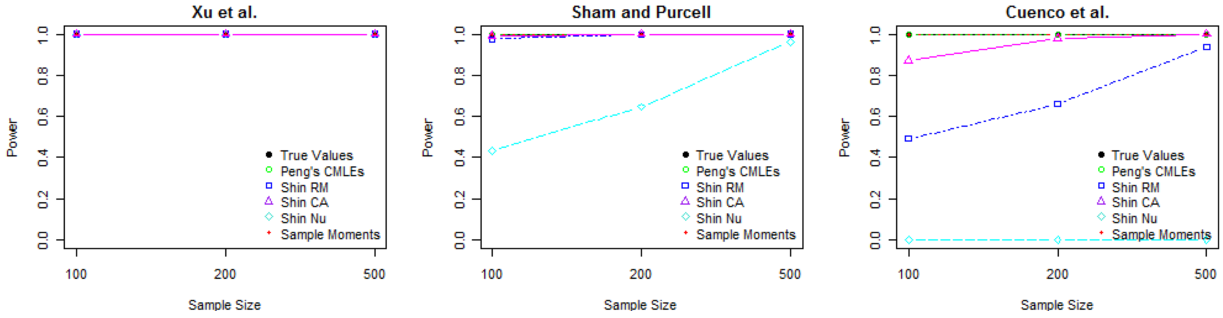
f. Model 6: $p = 0.1, h^2 = 0.4$, Dominant



g. Model 7: $p = 0.01, h^2 = 0.2$, Dominant



h. Model 8: $p = 0.01, h^2 = 0.4$, Dominant



Chapter 5 Discussion

We carried out a study of the power of model-free linkage analysis methods in truncate samples using three different estimators of trait parameter values that were derived under the assumption of normality. The simulation studies by Cuenco et al. (2003) first showed that selected samples are more affected than random samples by mis-specification of trait parameter values. The few additional studies on this topic were limited to the study of only one linkage analysis method with two estimators (Peng and Siegmund, 2006), or to the study of several linkage analysis methods with randomly chosen parameter values (Cuenco et al., 2003; Bhattacharjee et al., 2008). Ironically, Cuenco et al. (2003) showed that the one method considered by Peng and Siegmund was one of the weakest in power. Thus, we applied Peng's estimator to three different linkage analysis methods (especially, more powerful ones) and we also developed three new estimators that can be applied to model-free linkage analysis methods for truncate samples.

In our studies, Xu, S & P, and Cuenco's linkage methods gave greater power than T & S, a finding that is consistent with the results of other researchers (Cuenco et al. 2003). Even though comparing the power of different model-free methods was not the primary focus of this study, our results recommended using S & P's method with Peng's CMLEs for most studies, and the original H & E for rare allele frequency model.

Cuenco et al. (2003) noted that Xu's method may be the most robust except in the case of selected samples. However, we showed that several of our estimators (Shin RM or Shin CA) or Peng’s CMLEs retain the robustness and the power of Xu’s method, even for selected samples. In most cases, two of our suggested estimators give greater power (and higher LOD scores) for Xu’s method than Peng’s CMLEs. The reason that our estimators have more power is that our estimators yield better mean estimates and Xu’s test statistic requires only the trait mean, not the trait variance or trait sibpair correlation.

However, when S & P’s and Cuenco’s methods are used, rather than Xu et al’s, Peng’s CMLEs have greater power than our estimators. As mentioned above, our estimators give better mean estimates than Peng’s CMLEs, and all of the estimators give the same correlation estimates. Thus, the variance estimator is the factor that could account for the differences in power of analyses based our estimators vs. Peng’s estimators. Cuenco et al. (2003) and Bhattacharjee et al. (2008) emphasized the sensitivity of power to misspecified mean and correlation. However, our results show that the trait variance estimator is as important as the trait mean estimator and the correlation estimator for maximizing power to misspecified mean and correlation.

Overall, S & P’s method seems to be more powerful than Xu and Cuenco’s methods and Xu’s method seems more robust to the estimators used (even with sample moments and Shin Nu). One should note that our estimators are derived from a truncated likelihood function, assuming the normal distribution and a known truncation percentile. Therefore, if information on the truncation scheme is limited to knowing just the truncation value, we suggest the use of S & P’s method with Peng’s CMLEs.

To our knowledge, rare allele frequency models ($p = 0.01$) have not been considered by other researchers. These models may be unrealistic, since the required spacing between trait distributions would have to be as large as 5 SD units (with an additive model) to result in a heritability of 0.2 or more. In a future study, an allele frequency of 0.05 may result in a more realistic effect size for a rare allele frequency model. Under rare frequency models, the power is great no matter what estimators are used, but the estimates are very biased. Surprisingly, sample moments give the best estimates and the greatest power for Xu, S & P, and Cuenco's methods. One should note that two of our estimators did not perform well when they are applied to Cuenco's method. We suggest that the major reason for the uniformly high power under rare allele frequency models, is that the greater spacing between trait distributions results in larger genotype variance and less overlap in the trait distributions of individuals with different genotypes. For rare allele frequency models, the original H & E method is the most robust and is essentially as powerful as other methods.

A very important issue to keep in mind is that all linkage methods, except for the original H & E, are derived under the assumption of a single normal distribution of a quantitative trait. However, by definition, a major quantitative trait locus will always generate a mixture distribution. When we treat the mixture of normal as a single normal, we will invariably have biased estimates for trait parameter values, especially in truncate samples.

Bhatttacharjee et al. (2008) examined the effects of including higher moments (skewness and kurtosis) on the power of linkage methods. They concluded that incorporating higher moments results in more power than ignoring higher moments, only under highly non-normal distributions.

A logical next step that might be more successful for truncate sampling at a known percentile, would be to simultaneously consider the “correct” likelihood; specifically either a two or three component mixture of normals. Recently, estimation of parameter values for the mixture models that result in the case of a major gene has been studied by Gianola et al. (2007). However, we need to be aware that in doing so we are incorporating genetic parameter values and thus the methods are no longer gene model-free.

References

Aboueissa A.E.A., Stoline M.R. (2004) Estimation of the mean and standard deviation from normally distributed singly-censored samples. *Environmetrics* 15:659-673

Amos C.I. (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics* 54:535-43

Bhattacharjee S., Kuo C., Mukhopadhyay N., Brock G.N., Weeks D.E., Feingold E. (2008) Robust score statistics for QTL linkage analysis. *American Journal of Human Genetics* 82:567-582

Carroll R.J., Ruppert D., Stefanski L.A., Crainiceanu C.M. (2006) Measurement error in nonlinear models: a modern perspective *Chapman & Hall/CRC*

Casella G., Berger R.L. (2002) Statistical inference *Duxbury*

Camp N.J., Cox A. (2002) Quantitative trait loci: methods and protocols. *New Jersey: Humana Press*

Clerget-Darpoux F., Elston R.C. (2007) Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. *Human Heredity* 64:91-96

Cohen A.C. (1959) Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics* 3:217-237

Cuenco K.T., Szatkiewicz J.P., Feingold E. (2003) Recent advances in human Quantitative-Trait-Locus mapping: Comparison of Methods for Selected Sibling Pairs. *American Journal of Human Genetics* 73:876-873

Drigalenko E. (1998) How sib pairs reveal linkage. *American Journal of Human Genetics* 63:1242-1245

Elston R.C., Buxbaum S., Jacobs K.B., Olson J.M. (2000) Haseman and Elston revisited. *Genetic Epidemiology* 19:1-17

Feingold E. (2001) Methods for linkage analysis of quantitative trait loci in humans. *Theoretical Population Biology* 60:167-180

Fisher R.A. (1918) The correlation between relatives on the supposition of mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh* 52:399-433

Gianola D., Boettcher P.J., Odegard J., Heringstad B. (2007) Mixture models in quantitative genetics and application to animal breeding

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

Huang C., Li K., Fleur R.S., Chang S., Choi S.H., Shen T., Shin S., Finch S.J., Mendell N.R. (2009) Family-based analysis of a myocardial infarction endophenotype: comparison of sampling designs. *BMC Genetics*

Laird N.M., Lange C. (2006) Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics* 7:385-394

Mendell N.R., Elston R.C. (1974) Multifactorial qualitative traits: genetic analysis and prediction of recurrence risks. *Biometrics* 30:41-57

Peng J. and Siegmund D. (2006) QTL mapping under ascertainment. *Annals of Human Genetics* 70:867-881

Sham P. (1998) Statistics in human genetics. *Great Britain: Arnold*

Sham P.C., Purcell S. (2001) Equivalence between Haseman-Elston and Variance-Components linkage analyses for sib pairs. *American Journal of Human Genetics* 68:1527-1532

Schnabel R.B., Koontz J.E., Weiss B.E. (1985) A modular system of algorithms for unconstrained minimization. *ACM Transactions on Mathematical Software*, 11:419-440

Tang H.K., Siegmund D. (2001) Mapping quantitative trait loci in oligogenic models, Biostatistics 2:147-162

Thomas D.C. (2004) Statistical methods in genetic epidemiology. *New York: Oxford University Press*

Xu X., Weiss S., Xu X., Wei L.J. (2000) A unified Haseman-Elston method for testing linkage with quantitative traits. *American Journal of Human Genetics* 67:1025-1028

Appendix A. The constant used in transformation, c, is derived as follows

$$\mathrm{Cov}(Z_1^{'}, Z_2^{'}) = \frac{\mathrm{Cov}(Z_1, Z_2 + cZ_1)}{\sqrt{1 + c^2 + 2c\rho_g}} = \frac{\rho_g + c}{\sqrt{1 + c^2 + 2c\rho_g}} \equiv \rho. \tag{A.1}$$

$$\rho_g + c = \rho\sqrt{1 + c^2 + 2c\rho_g} \tag{A.2}$$

$$\frac{\left(\rho_g + c\right)^2}{\rho^2} = 1 + c^2 + 2c\rho_g = \left(\rho_g + c\right)^2 + \left(1 - \rho_g^2\right) \tag{A.3}$$

$$\left(\frac{1}{\rho^2} - 1\right)\left(\rho_g + c\right)^2 = 1 - \rho_g^2 \tag{A.4}$$

$$\rho_g + c = \sqrt{\frac{1 - \rho_g^2}{\frac{1}{\rho^2} - 1}} \tag{A.5}$$

$$c = \sqrt{\frac{1 - \rho_g^2}{\frac{1}{\rho^2} - 1}} - \rho_g \tag{A.6}$$

Appendix B. Power of Five Model-Free Methods with Estimators, for Truncate Samples under Models 1-8 (1000 Replicates, Sample Size=100, 200 and 500, $\alpha = 0.0001$). Margin of Error at 95% Confidence ≈ 0.03.

a. Additive Models

	Model 1			Model 2		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
True Values	0.17	0.22	0.16	0.96	0.98	0.96
Peng's CMLEs	0.14	**0.20**	0.14	0.91	**0.97**	0.94
Shin RM	0.17	0.18	0.09	0.95	**0.96**	0.78
Shin CA	0.16	0.18	0.10	0.95	**0.96**	0.78
Shin Nu	0.13	0.16	0.06	0.84	0.93	0.56
Sample Moments	0.14	0.04	0.03	0.67	0.28	0.24
Sample Size=200						
True Values	0.51	0.61	0.59	1.00	1.00	1.00
Peng's CMLEs	0.45	**0.59**	0.54	1.00	1.00	1.00
Shin RM	0.49	0.57	0.47	1.00	1.00	0.98
Shin CA	0.49	0.57	0.48	1.00	1.00	0.99
Shin Nu	0.44	0.54	0.29	0.99	1.00	0.72
Sample Moments	0.40	0.14	0.12	0.93	0.72	0.71
Sample Size=500						
True Values	0.99	1.00	1.00	1.00	1.00	1.00
Peng's CMLEs	0.96	**0.99**	**0.99**	1.00	1.00	1.00
Shin RM	0.98	**0.99**	**0.99**	1.00	1.00	1.00
Shin CA	0.98	**0.99**	**0.99**	1.00	1.00	1.00
Shin Nu	0.94	0.99	0.75	1.00	1.00	0.88
Sample Moments	0.92	0.52	0.51	1.00	1.00	1.00
	Model 3			Model 4		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
True Values	0.91	0.91	0.89	1.00	1.00	1.00
Peng's CMLEs	0.88	0.86	0.80	**1.00**	**1.00**	**1.00**
Shin RM	0.89	0.82	0.57	**1.00**	0.97	0.48
Shin CA	**0.90**	0.84	0.71	**1.00**	0.99	0.89
Shin Nu	0.87	0.37	0.01	1.00	0.43	0.00
Sample Moments	0.95	0.91	0.90	1.00	1.00	1.00
Sample Size=200						
True Values	0.99	1.00	1.00	1.00	1.00	1.00
Peng's CMLEs	**0.99**	**0.99**	**0.99**	**1.00**	**1.00**	**1.00**
Shin RM	**0.99**	**0.99**	0.87	**1.00**	**1.00**	0.69
Shin CA	**0.99**	**0.99**	0.97	**1.00**	**1.00**	0.98
Shin Nu	0.99	0.58	0.00	1.00	0.66	0.00
Sample Moments	1.00	0.99	0.99	1.00	1.00	1.00
Sample Size=500						
True Values	1.00	1.00	1.00	1.00	1.00	1.00
Peng's CMLEs	1.00	1.00	1.00	**1.00**	**1.00**	**1.00**
Shin RM	1.00	1.00	1.00	**1.00**	**1.00**	0.95
Shin CA	1.00	1.00	1.00	**1.00**	**1.00**	**1.00**
Shin Nu	1.00	0.92	0.00	1.00	0.96	0.00
Sample Moments	1.00	1.00	1.00	1.00	1.00	1.00

b. Dominant Models

	Model 5			Model 6		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
True Values	0.21	0.24	0.18	0.99	1.00	1.00
Peng’s CMLEs	0.17	**0.23**	0.16	0.96	**0.99**	0.98
Shin RM	0.21	0.21	0.13	**0.99**	**0.99**	0.86
Shin CA	0.20	0.21	0.13	0.98	**0.99**	0.80
Shin Nu	0.14	0.19	0.06	0.63	0.91	0.78
Sample Moments	0.10	0.01	0.01	0.64	0.22	0.19
Sample Size=200						
True Values	0.64	0.71	0.68	1.00	1.00	1.00
Peng’s CMLEs	0.58	**0.69**	0.65	1.00	1.00	1.00
Shin RM	0.62	0.68	0.60	1.00	1.00	1.00
Shin CA	0.62	0.68	0.60	1.00	1.00	0.99
Shin Nu	0.52	0.66	0.42	0.85	1.00	0.98
Sample Moments	0.32	0.05	0.05	0.94	0.61	0.61
Sample Size=500						
True Values	1.00	1.00	1.00	1.00	1.00	1.00
Peng’s CMLEs	1.00	1.00	1.00	1.00	1.00	1.00
Shin RM	1.00	1.00	1.00	1.00	1.00	1.00
Shin CA	1.00	1.00	1.00	1.00	1.00	1.00
Shin Nu	0.99	1.00	0.91	0.99	1.00	1.00
Sample Moments	0.89	0.33	0.32	1.00	0.99	0.99
	Model 7			Model 8		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
True Values	0.91	0.91	0.91	1.00	1.00	1.00
Peng’s CMLEs	0.87	0.84	0.79	**1.00**	**1.00**	**1.00**
Shin RM	**0.89**	0.80	0.54	**1.00**	0.98	0.49
Shin CA	**0.89**	0.83	0.68	**1.00**	0.99	0.87
Shin Nu	0.86	0.37	0.01	1.00	0.43	0.00
Sample Moments	0.96	0.92	0.91	1.00	1.00	1.00
Sample Size=200						
True Values	1.00	1.00	1.00	1.00	1.00	1.00
Peng’s CMLEs	1.00	1.00	0.99	1.00	1.00	1.00
Shin RM	1.00	0.99	0.90	1.00	1.00	0.66
Shin CA	1.00	0.99	0.98	1.00	1.00	0.98
Shin Nu	1.00	0.61	0.00	1.00	0.65	0.00
Sample Moments	1.00	1.00	1.00	1.00	1.00	1.00
Sample Size=500						
True Values	1.00	1.00	1.00	1.00	1.00	1.00
Peng’s CMLEs	1.00	1.00	1.00	1.00	1.00	1.00
Shin RM	1.00	1.00	1.00	1.00	1.00	0.94
Shin CA	1.00	1.00	1.00	1.00	1.00	1.00
Shin Nu	1.00	0.91	0.00	1.00	0.96	0.00
Sample Moments	1.00	1.00	1.00	1.00	1.00	1.00

Appendix C. P-value of McNemar’s test for the significance of changes in the LOD > 3 for the Three Model-Free Methods with Peng’s CMLEs and our estimators, for Truncate Samples under Models 1-8 (1000 Replicates, Sample Size=100, 200 and 500) When there is significant changes in the LOD > 3 for linkage methods with Peng’s CMLEs and with our estimators, the p-value is marked in bold (significance level=0.05)

a. Additive Models

	Model 1			Model 2		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
CMLEs vs. Shin RM	< 0.001	< 0.001	< 0.001	< 0.001	**0.11**	< 0.001
CMLEs vs. Shin CA	< 0.001	< 0.001	< 0.001	< 0.001	**0.11**	< 0.001
Sample Size=200						
CMLEs vs. Shin RM	< 0.001	< 0.001	< 0.001	NA	NA	NA
CMLEs vs. Shin CA	< 0.001	< 0.001	< 0.001	NA	NA	NA
Sample Size=500						
CMLEs vs. Shin RM	< 0.001	**1**	**0.34**	NA	NA	NA
CMLEs vs. Shin CA	< 0.001	**1**	**1**	NA	NA	NA
	Model 3			Model 4		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
CMLEs vs. Shin RM	0.002	< 0.001	< 0.001	NA	< 0.001	< 0.001
CMLEs vs. Shin CA	< 0.001	0.002	< 0.001	NA	0.013	< 0.001
Sample Size=200						
CMLEs vs. Shin RM	**1**	**0.07**	< 0.001	NA	NA	NA
CMLEs vs. Shin CA	**1**	**0.48**	< 0.001	NA	NA	NA
Sample Size=500						
CMLEs vs. Shin RM	NA	NA	NA	NA	NA	NA
CMLEs vs. Shin CA	NA	NA	NA	NA	NA	NA

b. Dominant Models

	Model 5			Model 6		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
CMLEs vs. Shin RM	< 0.001	< 0.001	< 0.001	< 0.001	**0.29**	< 0.001
CMLEs vs. Shin CA	< 0.001	< 0.001	< 0.001	< 0.001	**0.45**	< 0.001
Sample Size=200						
CMLEs vs. Shin RM	< 0.001	0.044	< 0.001	NA	NA	NA
CMLEs vs. Shin CA	< 0.001	0.022	< 0.001	NA	NA	NA
Sample Size=500						
CMLEs vs. Shin RM	**0.25**	NA	NA	NA	NA	NA
CMLEs vs. Shin CA	**0.25**	NA	NA	NA	NA	NA
	Model 7			Model 8		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
CMLEs vs. Shin RM	< 0.001	< 0.001	< 0.001	NA	NA	< 0.001
CMLEs vs. Shin CA	< 0.001	0.025	< 0.001	NA	NA	< 0.001
Sample Size=200						
CMLEs vs. Shin RM	NA	0.04	< 0.001	NA	NA	NA
CMLEs vs. Shin CA	NA	**0.13**	0.016	NA	NA	NA
Sample Size=500						
CMLEs vs. Shin RM	NA	NA	NA	NA	NA	NA
CMLEs vs. Shin CA	NA	NA	NA	NA	NA	NA

Appendix D. 95% Confidence Interval for the LOD differences: (LOD with Peng’s CMLEs) – (LOD with our estimators) for the Three Model-Free Methods (Paired t-test). The limits of confidence interval are marked in bold, when they are negative, i.e. when our estimator give higher power than Peng’s CMLEs. The limits of confidence interval are colored with green, when the interval includes zero, i.e. when there is no significant difference in LODs with our estimators and those with Peng’s CMLEs. Truncate Samples Under Models 1-4 (1000 Replicates, Sample Size=100, 200 and 500)

a. Additive Models

	Model 1			Model 2		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
CMLEs vs. Shin RM	**-0.16, -0.10**	0.06, 0.08	0.22, 0.29	**-0.36, -0.25**	0.26, 0.32	0.89, 1.05
CMLEs vs. Shin CA	**-0.16, -0.10**	0.06, 0.08	0.21, 0.27	**-0.32, -0.22**	0.25, 0.32	0.85, 1.02
Sample Size=200						
CMLEs vs. Shin RM	**-0.21, -0.15**	0.07, 0.10	0.23, 0.31	**-0.26, -0.15**	0.34, 0.42	1.43, 1.68
CMLEs vs. Shin CA	**-0.21, -0.15**	0.07, 0.09	0.20, 0.28	**-0.24, -0.13**	0.33, 0.41	1.25, 1.49
Sample Size=500						
CMLEs vs. Shin RM	**-0.38, -0.30**	0.10, 0.13	0.26, 0.35	0.05, 0.17	0.35, 0.44	2.42, 2.84
CMLEs vs. Shin CA	**-0.41, -0.34**	0.10, 0.13	0.22, 0.32	0.05, 0.17	0.34, 0.43	1.76, 2.14
	Model 3			Model 4		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
CMLEs vs. Shin RM	**-0.32, -0.26**	0.46, 0.58	1.26, 1.45	1.16, 1.33	5.63, 6.48	7.57, 8.02
CMLEs vs. Shin CA	**-0.22, -0.20**	0.20, 0.29	0.43, 0.57	0.36, 0.52	3.38, 3.93	2.82, 3.20
Sample Size=200						
CMLEs vs. Shin RM	**-0.14, -0.06**	0.73, 0.91	2.24, 2.61	2.39, 2.60	9.78, 11.13	16.68, 17.36
CMLEs vs. Shin CA	**-0.59, -0.50**	0.23, 0.35	0.46, 0.68	0.83, 1.03	4.86, 5.67	4.24, 4.88
Sample Size=500						
CMLEs vs. Shin RM	**-0.41, -0.30**	0.60, 0.77	3.51, 4.11	6.14, 6.49	22.45, 24.92	45.34, 46.45
CMLEs vs. Shin CA	**-1.58, -1.45**	**-0.17, -0.01**	**-0.49, -0.17**	2.31, 2.62	8.17, 9.46	8.36, 9.50

b. Dominant Models

	Model 5			Model 6		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
CMLEs vs. Shin RM	**-0.21, -0.14**	0.04, 0.06	0.18, 0.25	**-0.42, -0.25**	0.11, 0.18	0.87, 1.03
CMLEs vs. Shin CA	**-0.19, -0.12**	0.04, 0.06	0.18, 0.25	**-0.06, 0.11**	0.08, 0.15	1.22, 1.41
Sample Size=200						
CMLEs vs. Shin RM	**-0.26, -0.18**	0.07, 0.09	0.25, 0.33	**-0.09, 0.13**	0.14, 0.21	1.50, 1.76
CMLEs vs. Shin CA	**-0.22, -0.15**	0.07, 0.09	0.25, 0.33	0.56, 0.77	0.09, 0.17	2.19, 2.50
Sample Size=500						
CMLEs vs. Shin RM	**-0.43, -0.32**	0.09, 0.11	0.27, 0.35	0.92, 1.24	0.02, 0.11	2.99, 3.46
CMLEs vs. Shin CA	**-0.38, -0.27**	0.08, 0.11	0.25, 0.34	2.54, 2.87	**-0.11, 0.00**	5.04, 5.62
	Model 7			Model 8		
	Xu	S & P	Cuenco	Xu	S & P	Cuenco
Sample Size=100						
CMLEs vs. Shin RM	**-0.13, -0.07**	0.46, 0.58	1.26, 1.47	1.15, 1.31	5.62, 6.48	7.58, 8.03
CMLEs vs. Shin CA	**-0.35, -0.28**	0.21, 0.30	0.48, 0.63	0.31, 0.48	3.38, 3.93	2.79, 3.18
Sample Size=200						
CMLEs vs. Shin RM	**-0.17, -0.10**	0.65, 0.81	2.12, 2.47	2.48, 2.69	10.63, 12.03	16.91, 17.61
CMLEs vs. Shin CA	**-0.65, -0.57**	0.17, 0.28	0.37, 0.58	0.86, 1.07	5.30, 6.15	4.60, 5.26
Sample Size=500						
CMLEs vs. Shin RM	**-0.37, -0.25**	0.66, 0.86	3.79, 4.43	6.64, 6.95	25.01, 27.68	46.48, 47.56
CMLEs vs. Shin CA	**-1.55, -1.43**	**-0.17, -0.01**	**-0.46, -0.15**	2.65, 2.97	9.63, 11.12	9.47, 10.70