ABSTRACT

QUANTITATIVE TRAIT LOCI (QTL) DETECTION USING THE ACCELERATED FAILURE TIME CURE MODEL

Devrim Bilgili, Ph.D.
Department of Mathematical Sciences
Northern Illinois University, 2009
Nader Ebrahimi, Director

Many important problems in evolutionary biology begin with observations of phenotypic variation. For example, variation in survival time is partly due to genetic differences among individuals and partly due to environmental factors. Developing statistical methodologies relating the unknown genotype of QTL (Quantitative Trait Loci) to the survival time of an individual is important in identifying the disease causing genes. In some genetic studies with survival end points the population under study consists of susceptible and nonsusceptible individuals. All susceptible subjects would eventually experience the event in the absence of censoring, while nonsusceptible subjects are not at risk of developing such events and can be regarded as "cured." In this thesis we will discuss a procedure to detect the QTL locations for survival data using Accelerated Failure Time Cure Model.

This thesis is organized as follows. In Chapter 1, we give a brief review of the literature and introduce several terminologies. In Chapter 2, we concentrate on detection of QTL using parametric Accelerated Failure Time Cure Model. For that we propose two methods, namely EM-algorithm and imputation methods to estimate the unknown parameters. In Chapter 3, we work on identifying the QTL

using nonparametric Accelerated Failure Time Cure model. In Chapters 4 and 5, we study the asymptotic behavior of LOD scores and compare them with empirical LOD scores obtained by permuting the phenotypes of survival times of the data. Finally in Chapter 6, we describe several problems that we plan to work in the future.

NORTHERN ILLINOIS UNIVERSITY
DE KALB, ILLINOIS

JANUARY 2009

QUANTITATIVE TRAIT LOCI (QTL) DETECTION USING THE ACCELERATED FAILURE TIME CURE MODEL

BY

DEVRIM BILGILI, PH.D.

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE
DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICS

Doctoral Director:
Nader Ebrahimi

ACKNOWLEDGMENTS

This thesis has emerged as a result of the help and encouragement of many people. I am especially grateful to my mentor, Professor Nader Ebrahimi in the Department of Mathematics, statistics division at Northern Illinois University. Throughout my graduate studies, Dr. Ebrahimi offered me great advise. I would like to thank all my dissertation committee members–Dr. Alan Polansky, Dr. Sudhir Gupta, Dr. Rama Lingham and Dr. Syed Kirmani–for their valuable opinions and time. I am grateful to Professor Peter Waterman and Vadim Sokolov for their helpful discussions in Mathematica. Of course without graduate students the graduate school would be nothing. I am so lucky to have the friendships of Katsuo Hisano, Yarong Yang, and Mohan Thapa for so many years. Also, my thanks to staff members Donna Lynn and Sue Altenburg for smoothing things in graduate school. There is no person like my mom. I would not be where I am now without her generous and unconditional support. Unfortunately, my dad did not live through to see this day. But he was the architect of this Ph.D and the man who I am today. So dad, this is for you!

TABLE OF CONTENTS

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

In 1865, Mendel published ground-breaking experiments demonstrating evidence for a single-locus inheritance in plants; see WEILING (1991) for more details. To understand Mendel's view, consider his experiments which followed seven traits of the common garden pea. In one experiment Mendel crossed a pure-breeding line with yellow peas to a pure-breeding line with green peas. Let P_1 and P_2 denote these two parental populations. The cross $P_1 \times P_2$ is called the *first filial*, or F_1, population. In the F_1, Mendel observed that all of the peas were yellow! Crossing members of F_1, i.e., $F_1 \times F_1$ gives the second filial or F_2 population. The results from the F_2 were shocking:1/4 of the plants had green peas, 3/4 had yellow peas. *This outbreak of variation, recovering both green and yellow from yellow parents, was a breakthrough.* Mendel also observed that P_1(all yellow), F_1(all yellow) and F_2(3/4 is yellow), behaved very differently when crossed to P_2(pure breeding green). Below is the summary of all these crosses.

What is the explanation of these rather complex-looking results? *Genes are discrete particles, with each parent passing one copy to its offspring.* Furthermore, the F_1 are all yellow and are heterozygous, which means that both the YY and Yg genotypes map to the yellow pea phenotype. Likewise, the gg genotype maps to the green pea phenotype. Since the Yg heterozygote has the same phenotype as the YY homozygote, we say that the Y allele is *dominant* to g or that g is recessive to

Y. With this model of inheritance in hand, we can now revisit the above crosses. Consider the results in the F_2 cross. Here *both parents are heterozygotes*. Now the question is: what are the probabilities of the three possible genotypes in their offspring?

$$\begin{aligned} \mathrm{P}(YY) &= \mathrm{P}(Y \text{ from dad})P(Y \text{ from mom}) = (1/2)(1/2) = 1/4 \\ \mathrm{P}(gg) &= \mathrm{P}(g \text{ from dad})P(g \text{ from mom}) = (1/2)(1/2) = 1/4 \\ \mathrm{P}(Yg) &= 1 - P(YY) - P(gg) = 1/2 \end{aligned}$$

Hence, Prob(yellow phenotype)=P(YY)+P(Yg) = 3/4 as Mendel observed. See Table 1.1 for details.

Table 1.1: Mendel's Crosses for Peas.

Cross	Offspring
P_1	Yellow Peas
P_2	Green Peas
$F_1 = P_1 \times P_2$	Yellow Peas
$F_2 = F_1 \times F_1$	3/4 Yellow Peas, 1/4 Green Peas
$F_1 \times P_2$	1/2 Yellow Peas, 1/2 Green Peas
$F_2 \times P_2$	2/3 Yellow Peas, 1/3 Green Peas

For the seven traits that Mendel followed, he observed *independent assortment* of the genetic factors at different loci(genes), with the genotype at one locus being independent of the genotype at the second. See Table 1.1. To understand this, consider a cross involving two traits: round vs. wrinkled seeds and green vs. yellow peas. The genotype to phenotype mapping for these traits is (RR, Rr=round seeds), rr wrinkled seeds, and (YY, Yg=yellow), gg=green. Consider the cross of a pure round green($RRgg$)line×pure wrinkled yellow($rrYY$)line. In the F_1 cross, all the offsprings are $RrYg$ or round and yellow. What happens in F_2 cross? A quick way to figure this out is to use the notation $R-$ to denote both the RR and Rr genotypes.

Hence, round peas have genotype $R-$. Likewise, yellow peas have genotype $Y-$. In the F_2, the probability of getting an $R-$ genotype is just

$$P(R\_|F_2) = P(RR|F_2) + P(Rr|F_2) = 1/4 + 1/2 = 3/4.$$

Since the genotypes at different loci are independently inherited, the probability of seeing a round, yellow F_2 is

$$P(R - Y-) = P(R-)P(Y-) = (3/4)(3/4) = 9/16.$$

Likewise,

$$\begin{aligned} P(\text{yellow, wrinkled}) &= P(rrY-) = P(rr)P(Y-) = (1/4)(1/4 + 1/2) = 3/16 \\ P(\text{green, round}) &= P(ggR-) = P(gg)P(R-) = (1/4)(3/4) = 3/16 \\ P(\text{green, wrinkled}) &= P(rrgg) = P(rr)P(gg) = (1/4)(1/4) = 1/16. \end{aligned}$$

Hence, the four possible genotypes are seen in a 9:3:3:1 ratio.

Shortly after the rediscovery of Mendel, Bateson and Punnett (1911) looked at a cross in peas involving a flower color locus (with the purple P allele dominant over the red p allele) and a pollen shape locus (with the long allele L dominant over the round allele l). They examined the F_2 from a pure breeding purple long($PPLL$) and red round($ppll$) cross. The resulting genotypes, and their actual and expected numbers under independent assortment were as shown in Table 1.2.

Table 1.2: Observed and Expected Counts of Flower Color Locus.

Phenotype	Genotype	Observed	Expected
Purple long	$P - L-$	284	215
Purple round	$P - ll$	21	71
Red long	$ppL-$	21	71
Red round	$ppll$	55	24

This was a significant departure from independent assortment based on the Karl Pearson's χ^2 test with an excess of PL and pl gametes over Pl and pL and evidence that the P and L genes are *linked*, physically associated on the same chromosome. In fact, it is now known that if genes are located on different chromosomes they (with few exceptions) show independent assortment. Indeed, peas have only 7 chromosomes and the traits were selected by Mendel were in seven different chromosomes. Mendel simply could have observed an independent assortment because of simple luck.

Current studies move beyond Mendel's single-locus models to consider complex genetic traits governed by combinations of multiple genetic and environmental factors. For example, several investigations have identified genomic regions related to obesity in mice; see STYLIANOU et al. (2006) and WUSCHKE et al. (2007), while environmental factors such as high-fat diet CHEN et al. (2007), and exposure to high-frequency vibration RUBIN et al. (2007), may also play a role. In general, experimental genetics allows the study of phenotypes of interest using specified matings in a controlled environment. Once a cross experiment is performed, statistical analysis is used to identify regions of the genome related to the trait of interest. Such analysis is an important step toward understanding the molecular bases of phenotypic variation. For example, the identification of the genes underlying a disease can be a crucial step toward understanding disease pathways, providing tools for public health screening and molecular targets for drug development.

In this dissertation, our goal is to develop statistical methodologies that can be used to explain the variation of a specific trait, for example survival time, in terms of variation in one gene or multiple genes in Quantitative Trait Locus or loci respectively. We call both of them QTL. To achieve our goal we first provide an

overview and some basic terminology of QTL mapping and techniques that serve as a foundation for the work we present in the body of the thesis.

1.1 Quantitative Trait Loci

A quantitative trait locus is a genetic locus, where we consider measurement of a continuous trait–for example, blood pressure. When we say *loci*, we mean there are multiple locations (genes) that have an effect on the trait versus *locus* which means there is only one location (gene) that impacts the trait. We usually observe the variation across the population. Some of this variation can be explained by genetic factors. The goal is to model the overall contribution to phenotype, and the genetic contribution at each location compared to the overall variance. QTL can apply similarly to ordinal or binary traits. Any given trait may be governed by one or more genetic loci, each contributing a portion to overall phenotypic variation. An underlying genetic locus can lie in a protein-coding gene, or in a non-coding regulatory region. A protein-coding gene consists of a promoter followed by the coding sequence for the protein and then a terminator. The promoter is a base-pair sequence that specifies where transcription begins. The coding sequence is a base-pair sequence that includes coding information for the polypeptide chain specified by the gene. The terminator is a sequence that specifies the end of the mRNA transcript. Alternatively, the genetic locus of interest could be a combination of neighboring genes or regulatory regions within a small physical distance. The definition of genetic locus is kept deliberately broad to cover the range of possible relationships between genotypes and phenotypic outcomes.

1.2 An Experimental Cross

Inbreeding plays an important role in identification of QTLs. Breeding experiments start with backcross and intercross. The other important source of genetic mapping is recombinant inbreds. Such strains can be obtained by repeated mating or selfing, at which point inbred strains develop to be homozygous at all sites on any given chromosome; see FESTING (1981) and SILVER (1995) for more details. Here, all offspring from two individuals of the same strain are genetically identical to their parents and each other, and any intra-strain phenotypic variation can be explained by environmental or subject-specific variation. In order to understand genetic factors, we have to select at least two different strains, which show different phenotypes under the same environmental conditions. Throughout this chapter the experimental crosses will be based on mice because of easy design and great control of the experiments with a minimum environmental effect on the identification of QTLs. Once we have identified two phenotypically distinct strains of mice, the next step is to perform an experimental cross to locate particular loci contributing to the difference in phenotypes. Two most popular experimental cross designs used for this purpose are the backcross and the intercross. Consider a genetic locus where two strains differ. Let's say that the genotype of one of the inbred strains is aa and the genotype of the other inbred strain is AA at a particular locus. After we cross these two different strains, it follows that the genotype of the offspring must be Aa, because offspring will inherit one of the alleles from each parent. The resulting offspring are called the F1 generation. A backcross is obtained by mating the F1 offspring with one of the original inbred strains. In a backcross (Figure 1), each of the F1 offspring is genetically identical to the others, with genotypes Aa across the genome indicating that the F1 offspring have one allele from each of the parental

strains at every position across the genome.

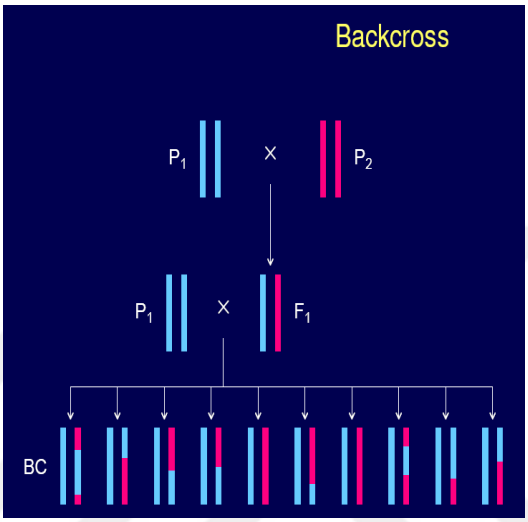

Figure 1.1: Backcross

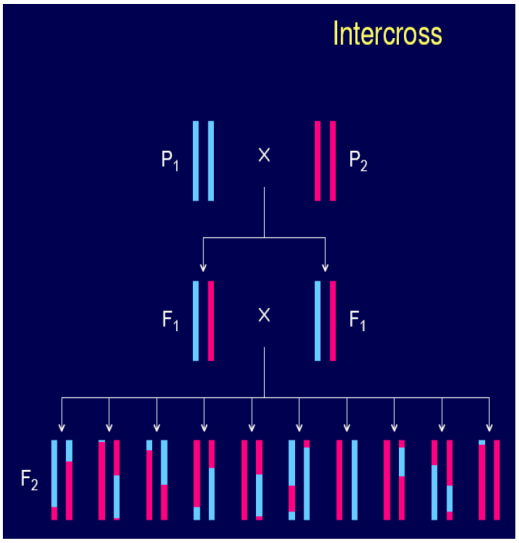

Figure 1.2: Intercross

Generating an intercross proceeds similarly (Figure 2). As with a backcross, we begin by creating the F1 line. The intercross is a result of the mating of an F1 male with an F1 female with possible genotypes AA, Aa and aa with probabilities $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$, respectively.

It is important to note that the Aa genotype which shows up in both the backcross and the intercross only corresponds to a heterozygous genotype at positions where the parent strains differ. Hence, the investigator should choose genetic markers which are informative based on the genotypes of the parental lines.

When it comes to choosing a cross type, there are some important differences between backcrosses and intercrosses according to the researchers' choice of experimental design. Specifically, backcross individuals have only two possible genotypes at each locus, while intercross individuals have three possible genotypes. This dif-

ference shows that analysis of a backcross will be simpler compared to that of an intercross. Because of the smaller number of possible genotypes in the backcross, the researcher will have fewer parameters to estimate; hence a backcross will have greater power than an intercross in identifying a QTL, given the same marker spacing, see SEN et al. (2007). On the other hand, the intercross provides the possibility to learn about both homozygote genotypes, while we are missing one of the homozygous genotypes in the backcross design.

1.3 Meiosis and Recombination

The cell process by which genetic material is passed from parent to offspring, meiosis, is at the top of all gene-mapping efforts. Sexual reproduction requires meiosis, a process in which a parent cell divides into four cells, each with half the genetic material of the parent. A diploid parent cell divides to make four haploid cells. In sexual reproduction, haploid gametes from two individuals then combine to produce a diploid zygote. An offspring resulting from sexual reproduction is genetically different from both parents. For example, if the parent's genotype is Aa at the given location, then it will pass either allele A or a with equal probabilities. If the parent's genotype is AA or aa it will certainly pass on the allele A or a respectively.

During this process, chromosomes undergo a process of recombination in which homologous chromosomes cross over and exchange sections of DNA. It is thought that this process of genetic recombination is very important to evolution, because this mechanism provides new combinations of haploid DNA which can emerge throughout this process; see WRIGHT (1931). As a result of recombination, every individual's genome is a mix of maternal and paternal genetic material. Because

the chromosomes are passed on from parent to offspring as relatively large segments of DNA, this means the inheritance of genetic material at a particular site on the chromosome is also related to the pattern of inheritance in close regions. That is, there is a tendency for two or more genes to be inherited together, and hence for related individuals with the disease phenotype to share alleles at some nearby marker locus. This is an important idea in mapping QTL to specific regions of the genome. Recombination information for QTL mapping requires a model to describe the process. Throughout this dissertation, we model recombination events to relate the rate of genetic recombination to genetic distance on a chromosome. Assuming no crossover interference induces a Poisson model on the sites of recombination corresponding to Haldane's map function; see HALDANE (1919). Genetic mapping is based on the idea that recombination is more likely between genes that are far apart on chromosomes than between genes that are close. Haldane's map function is a mathematical relationship between recombination fractions and map distances measured in centimorgans or Morgans. Many other map functions exist in the literature, and can be understood through stationary renewal processes; see ZHAO and SPEED (1996) for more details. Generally speaking, every map function explains genetic distance, as the expected number of crossovers within that region, in terms of the recombination fraction. Genetic distance is usually quantified in Morgans, where exactly one crossover is expected within a region of length 1 Morgan = 100 centiMorgans (cM), while a recombination fraction is quantified as a probability.

1.4 Simple Analysis

After the cross experiment is conducted, phenotype and genetic marker data are collected on each of the individuals. The main idea is to look at genotypes at a

particular marker of individuals to see if a significant part of phenotypic variance can be explained by genetic markers. The very simple way to do this is through a t-test or ANOVA.

A genome scan is then performed by least-squares regression of phenotypes on each marker, one at a time, with a corresponding test statistic to summarize evidence for genotype-phenotype association at each such position; see PETER ALMGREN et al. (2003) for more details. The idea is to find the marker which can explain the phenotypic variation. Then the neighboring region can be considered strong evidence which includes a QTL. However, genome scan one marker at a time has problems in that testing one location at a time for genetic loci is restricted to genetic markers, and it cannot be reliable in the case of small sample sizes or missing marker data. It is clear that multiple neighboring markers with a strong signal provide better evidence in identification of a QTL. However, using Hidden Markov Models (HMMs), see RABINER (1989), we can combine genotypes at individual markers as multipoint marker data. See also BROMAN (2006) for more details. One or two misplaced genotyping errors can have a big impact on the identification of QTL. Although many errors can be detected by rereading, it is a good idea to have a genotyping error model. We can model the errors in genotyping through HMMs (Hidden Markov Models) that a Hidden Markov Model generates the genotypes for all individuals. However, we can only observe the full genotypes across a genome, with some error, at a certain number of marker locations. HMMs provides a model for the unobservable true genotypes in terms of observed genotypes, emission probabilities, transmission probabilities, and initiation probabilities. Emission probabilities denote the probability of observing a particular genotype given the underlying true genotype at that locus, or the genotyping error rates. The transmission

probabilities indicate the genotype probabilities at a particular locus, conditional on the genotype at a neighboring locus, and are directly related to recombination fractions specified in a linkage map. The initiation probabilities specify the genotypes to start off the Markov chain at the beginning of the chromosome, and are taken as the marginal genotype probabilities known according to the structure of the experimental cross. It is worth noticing that the hidden Markov chain could be initiated with the same marginal genotype probabilities at any point on the chromosome, and the same results would follow. This is because the these initiation probabilities are an equilibrium point under the specified transmission probabilities. See MANICHAUKUL (2007) for more details.

HMMs not only can be used to model the joint probability of the observed multipoint marker data at any location, but also can be reversed through the Bayes rule to calculate the probability of a particular genotype at a locus of interest. This is done by conditioning on the observed set of genotypes for an individual. We are going to see that these conditional genotype probabilities are the main tool to do the genome scan by allowing putative QTL at positions in between markers. We will explain this procedure in more details in the next chapters.

1.5 Linkage Analysis

Linkage analysis is an important tool which can be used not only in experimental crosses but also in human pedigrees. Linkage analysis is a statistical technique used to find the approximate chromosomal locations of disease genes relative to a map of other genes with known locations. The idea is to look for evidence of co-segregation between the disease genes and markers whose locations are already known. By co-segregation, we mean a tendency for two or more genes to be inherited together,

and hence for related individuals with the disease phenotype to share alleles at some nearby marker locus.

The simplest linkage analysis is the *two-locus* linkage analysis which is between two genes, usually, but not necessarily, a disease gene and a marker gene. *The parameter of interest is the recombination fraction* r. Two genes perfectly linked to each other will always be transmitted together during meiosis, corresponding to $r = 0$, whereas unlinked genes, e.g. genes located on different chromosomes, are transmitted independently, corresponding to $r = 0.5$. The co-segregation of disease and marker alleles in a pedigree can be summarized in the likelihood function which measures the support, given by the data, for different r values.

The value of r that maximizes likelihood $L(r)$, or in other words, the value that best fits the observed data, is the relative frequency of recombinants. Once r has been estimated, we proceed by testing the null hypothesis of no linkage ($r = 0.5$), that is, whether the observed deviation from 50% recombination is statistically significant at some predefined level for each putative location in the genome with some usually fixed distance apart. For this purpose we could use the likelihood ratio test statistic,

$$LOD = \log_{10}\frac{L(r)}{L(0.5)},$$

known as the LOD score (for log-odds). The LOD score is the single most important concept in linkage analysis; it was introduced for human studies by HALDANE (1919). Positive LOD scores indicate evidence in favor of linkage.

A LOD score above 3 is generally accepted as significant evidence of linkage. A LOD score of 3 at $r = \hat{r}$ means that the observed data is 1000 times (10^3) more likely when $r = \hat{r}$ than under the null hypothesis $r = 0.5$, or equivalently that the odds for linkage are 1000:1. For linkage to a locus on the X chromosome, a LOD

score of 2 is usually considered sufficient for significant linkage. It should be noted that one of the nice properties of the LOD score function is that it is additive over independent families.

To see how the LOD score is calculated for a pedigree with n individuals, the likelihood for a pedigree with n individuals is defined as the probability of observing the phenotypes $y = (y_1, y_2, \cdots, y_n)$ given the model parameter r and the penetrance parameters f_0, f_1, f_2. Suppose there are two possible alleles A and a at this locus. Usually A denotes the disease susceptibility allele and a the normal allele, respectively; then,

$$f_0 = P(\text{affected}|aa)$$

$$f_1 = P(\text{affected}|Aa)$$

$$f_2 = P(\text{affected}|AA).$$

For instance, for a monogenic disease the penetrance parameters reveal if the disease is dominant (one disease allele of the disease genotype is sufficient for becoming affected), or recessive (both disease alleles are needed). The penetrance of the disease is the conditional probability that an individual is affected given the genotype.

These intra-pedigree phenotypes are typically dependent for genetic and/or environmental reasons, but let us assume that the dependency is purely genetic and that it can be completely accounted for by our model for shared genotypes. Thus, we assume that an individual's phenotypes y_i are independent conditional on their joint marker-disease genotypes, $g_i = (m_i, d_i)$.

$$P(y|g) = \prod_{i=1}^{n} P(y_i|g_i),$$

where $m_i = (m_{i1}, m_{i2})$ is the marker genotype and $d_i = (d_{i1}, d_{i2})$ is the disease genotype. The unconditional probability of the phenotypes $y = (y_1, y_2 \cdots, y_n)$ is

the pedigree likelihood:

$$L(r) = P(y|r) = \sum_g P(y, g|r) = \sum_g P(y|g)P(g|r),$$

where the summation is taken over all joint marker-disease genotypes (including phase) compatible with the observed data. The first factor in the summand depends on the $P(y|g)$, which depends on the penetrance parameters $f = (f_0, f_1, f_2)$, whereas the second factor $P(g|r)$ parameter depends on the recombination fraction, the disease allele frequency p and the disease marker frequency p_M. Let us first note that $P(g) = P(g_1 \cap g_2 \cap ... \cap g_n)$, which can be expressed as a product of conditional probabilities,

$$P(g) = P(g_1)P(g_2|g_1)P(g_3|g_1, g_2) \cdots P(g_n|g_1, \cdots, g_{n-1}).$$

This expression can be simplified further because for non-founders genotypes are independent conditional on the genotypes of the parents. For founders, genotypes are assumed to be independent and hence only depend on the population allele frequencies. Let us therefore divide the n pedigree members into two groups–founders(F) and non-founders(NF)–, then we can write

$$P(g) = \prod_{i \in F} P(g_i) \prod_{j \in NF} P(g_i|g_{Fj}, g_{Mj}),$$

where F_j and M_j denote the father and mother of non-founder j, respectively. The first product depends on the marker allele frequency p_M and the disease allele frequency p, whereas the second depends on the recombination fraction r. For founders it is common to assume that all pairs of alleles at the two genes (disease and marker) are in *linkage equilibrium* (in that the proportion of haplotypes (M_i, D_i) is equal to the proportions of M_i alleles at the marker locus times the proportion of D_i alleles at the disease locus). Then under this assumption the probability of a joint

marker-disease genotype can be written as

$$P(g_i) = P((m_i, d_i); p_M, p) = P(m_i; p_M)P(d_i; p).$$

Furthermore we assume that each of the two genes are at Hardy-Weinberg equilibrium. For non-founders, we have

$$P(g_i|g_{F_i}, g_{M_i}) = P(m_i, d_i|m_{F_i}, m_{M_i}, d_{F_i}, d_{M_i}; r).$$

Therefore, the pedigree likelihood is:

$$L(r) = \sum_g \prod_{i=1}^{n} P(y_i|g_i; f) \prod_{i \in F} P(g_i) \prod_{j \in NF} P(g_j|g_{F_j}, g_{M_j}) \tag{1.1}$$

Equation (1.1) can also be written as

$$L(r) = \sum_{g_1} \sum_{g_2} \cdots \sum_{g_n} P(y|g)P(g|r) \tag{1.2}$$

where g_i, $i = 1, \cdots, n$ is the set of possible joint marker-disease genotypes for individual number i in the pedigree. Equation (1.2) can be used to compute LOD score. The number of terms in the sum grows exponentially with the size of the pedigree, n, but fortunately it is possible to carry out the calculations sequentially in a way that the amount of computing rises only linearly with pedigree size. The idea is to break down the pedigree into nuclear families and peel the result from each nuclear-family calculation onto the individual linking that particular nuclear family to the rest of the pedigree. The procedure is known as the Elston-Stewart algorithm. See ELSTON (1971).

Example 1: Consider the pedigree given in Figure 1.3. This family has four members, mother and father and two children. The mother (2) and the second daughter (4) are affected with a certain disease (dark symbols). On the other hand, the rest of the family members are unaffected (open symbols). Let us assume that

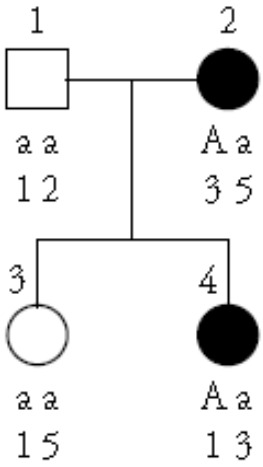

Figure 1.3: Pedigree.

the disease is dominantly inherited (one disease allele will be enough to get the person affected), which is penetrant. Also, all the family members are typed at one informative location. The genotypes at the disease locus and the marker locus are shown above for each individual. It is clear that the mother has transmitted the haplotype (a5) to her unaffected daughter (3) and (A3) to her affected daughter (4); however, it is impossible to determine whether a haplotype is recombinant or not, because we do not have any information about the phase of the mother. We consider two cases.

***Case.* 1** : The phase is $P_1 = (A3|a5)$.

***Case.* 2** : The phase is $P_2 = (A5|a3)$.

Under Case 1 the daughters are nonrecombinant. Under Case 2 both daughters are recombinant. The likelihood of this pedigree considering two phases can now be

written as

$$\begin{aligned} L(r) &= \sum_{i=1}^{2} L(r|P_i)P(P_i) \\ &= 0.5(L(r|P_1) + L(r|P_2)) \\ &= 0.5((1-r)^2 + r^2) \end{aligned}$$

In the above equation, 0.5 represents the fact that we don't have any information on both phases. Also, $(1-r)^2$ represents the two non-recombinants and r^2 represents the two recombinants respectively. The likelihood attains its maximum value at $r = 0$ where $r \in [0, 0.5]$. Thus the LOD score is

$$\begin{aligned} LOD &= log_{10}\frac{L(0)}{L(0.5)} \\ &= log_{10}\frac{0.5(1^2 + 0^2)}{0.5(0.5^2 + 0.5^2)} \\ &= 0.301 \end{aligned}$$

By looking at the LOD score above, we have a score of 0.301. If the denominator is significantly larger than the numerator then this is an indication that there is no linkage. In this case, since 0.301 is a significant departure from 0 this tells us that we have evidence to have a QTL at that particular location. Because the LOD scores are additive over the families, we can compute the linkage signal at this particular location among approximately ten families to get the desired threshold of 3; see LANDER and BOTSTEIN for details (1989), to consider the location as a putative QTL.

1.6 Interval Mapping

The term interval mapping is used for estimating the position of a QTL within two markers and was proposed by LANDER and BOTSTEIN (1989) to extend

the idea of testing genetic markers one at a time, with the interest in obtaining information for QTL in between markers. Interval mapping is originally based on maximum likelihood, but there are also very good approximations possible with simple regression to relate phenotypes to genotypes at marker locations. For a backcross with genotypes AA and AB, a regression model can be written as: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where X_i is a 0/1 indicator of the AB/AA genotype for the i^{th} individual respectively, and Y_i is the phenotype of the i^{th} individual. For individuals with the AA genotype, β_0 is the phenotype mean associated with the AA genotype which ignores QTL. The parameter β_1 is the increase in phenotype mean for the AB genotype compared to the AA genotype, and ϵ_i the normally distributed error term with common variance σ^2 across individuals.

Evidence in identifying QTL is summarized by a LOD score, defined as the log10-likelihood ratio comparing the alternative hypothesis of a QTL at the position of interest with the null hypothesis of no QTL.

In interval mapping, each position in the genome, one at a time, is taken as the putative QTL. Let $g = 0/1$ if the (unobserved) QTL genotype is AB/AA. Under the assumption that $Y \sim N(\mu_g, \sigma^2)$, given genotypes at linked markers, Y can be modeled as mixture of normals with mixing proportion $\Pr(g = 1|\text{marker data})$. For n individuals in the sample, the likelihood is

$$L(p_{ij}, \mu, \beta, \sigma^2|y, m) = \prod_{i=1}^{n} \sum_{j=0,1} p_{ij} N(\mu_{ij}, \sigma^2),$$

where $N(\mu_{ij}, \sigma^2)$ denotes the normal density function with mean μ_{ij} and variance σ^2 and

$$\mu_{i0} = \beta_0 + \beta_1/2 \text{ and } \mu_{i1} = \beta_0 - \beta_1/2,$$

are genotypic values of AA and AB. In estimation, this normal mixture problem

can be treated as an incomplete data problem, and the EM algorithm can be implemented to maximize the likelihood and obtain the MLEs of $\widehat{\mu_{i0}}$, $\widehat{\mu_{i1}}$, $\widehat{\sigma^2}$.

Now one can also consider a multiple QTL model, say p QTLs for the multiple interval mapping. See KAO (1999) for more details. Other methods have been proposed which also allow analysis to be presented as a continuum of evidence at both marker and non-marker locations. For example, Haley-Knott regression is a computationally simpler formulation of interval mapping, yielding nearly identical results in cases of little or no missing genotype data. See HALEY and KNOTT (1992). Multiple imputation– see SEN and CHURCHILL, (2001)– is an alternative approach which uses pseudomarkers to fill in unobserved marker data at non-marker locations, and averages over several sets of imputed genotypes to provide the marginal likelihood, as opposed to the profile likelihood produced by standard interval mapping.

We should mention that each of the methods presented so far relies on a normality assumption. That is, we assume phenotypes are normally distributed conditional on genotypes at the QTL positions. Still, many analysts go ahead and just check for normality in the overall phenotype distribution in practice, perhaps because there are too many possible genetic markers to check the conditional phenotype distributions in advance. Further, it can be difficult to check the normality assumption at positions in between markers, where we cannot readily produce residual plots since the actual genotypes are unknown. To address issues about distributional assumptions on phenotypes, several methods have been developed which are robust to non-normality, or are tailored for other distributional assumptions. For a completely non-parametric analysis, see KRUGLYAK and LANDER (1995). We also refer to XU and ATCHLEY (1996) for interval mapping for binary traits.

1.7 Significance Thresholds

Since a genome scan always involves looking for evidence of QTL at many putative positions, we must consider issues of multiple testing in reporting the level of significance for any LOD peak. Toward controlling genome-wide statistical significance, LANDER and BOTSTEIN (1989) proposed analytical approximations based on a limiting scenario of dense markers. LANDER and KRUGLYAK (1995) further argued that the dense marker approximation should be applied even to sparse genome scans, since investigators routinely add a dense set of markers to genomic regions identified by the preliminary sparse scan. By using the dense approximation a LOD threshold of 3.3 applies for a backcross, or 4.3 for an intercross. See LANDER and KRUGLYAK (1995). The investigator ensures a genome-wide false positive rate of 5%, without having to update significance thresholds after additional testing in follow-up. Empirical methods such as simulation and permutation provide an alternative to analytical approximation, with reported significance thresholds tailored to the particular data set at hand. A useful simulation technique known as gene-dropping preserves the relationship structure among individuals in the cross, generating new genotypes according to the cross structure while preserving the observed phenotypes. (see SAWCER [1997]). and the pattern of missing genotypes. The simulation yields an empirical distribution of the maximum LOD score under the null hypothesis of no genotype-phenotype association, conditional on cross structure and phenotypes. A LOD threshold at the desired level of significance may then be obtained as a quantile of this empirical distribution. CHURCHILL and DOERGE (1994) goes a step further and provides significance thresholds conditional on observed genotypes and phenotypes using permutation. Randomly shuffling genotypes relative to phenotypes preserves the observed set of genotypes and missing

data pattern, while eliminating any genotype-phenotype association. The resulting permutation distribution may be used to obtain a LOD threshold at the desired level of significance. This strategy works particularly well for the case of experimental crosses, in which all individuals may be treated as siblings, and are therefore exchangeable. Currently, the permutation thresholds of CHURCHILL and DOERGE (1994) are quite widely used in analysis of backcross and intercross data. For more complex pedigree structures, in which individuals are not exchangeable, we may use gene-dropping simulation, but not permutation.

As with any statistical analysis, the standard tools available are limited by particular assumptions, and are designed to operate within a fixed set of scenarios. Each new data set presents a new set of challenges, stimulating new areas of statistical methodology. In this dissertation, we present work we have done to improve basic analysis for QTL mapping in three different directions. Chapter 2 shows work on detection of QTL using a parametric Accelerated Failure Time Model. Chapter 3 extends parametric efforts in identifying the QTL into a nonparametric approach. Chapter 4 will describe the permutation tests for the threshold values for the LOD scores and bootstrap for the coverage probabilities of the QTL location. Finally, Chapter 5 presents asymptotic approximations in order to understand the significance threshold values. These five chapters cover a broad spectrum of methodological development, reflecting a model of data-driven research, with efforts shifting according to challenges emerging from real data analysis.

CHAPTER 2

INTERVAL MAPPING WITH A PARAMETRIC ACCELERATED FAILURE TIME CURE MODEL

Time-to-event data (traits) present themselves in different ways, which create special problems in analyzing such data. They are often non-normally distributed and show a right-skewed distribution of trait values across all individuals. Additionally, time-dependent traits often include censored observations, which occur when the true time of the event is unknown. End-of-study censoring arises when the event of interest has not occurred by the end of the study period. Within-study censoring arises if an individual is lost to follow-up during the course of study. The loss of information due to censoring results in lower statistical power. Typically, the greater the proportion of censoring the lower the statistical power. Some of this power can be recovered by modeling censored individuals in the statistical analysis; however, standard QTL-mapping techniques typically do not account for this. The field of survival analysis utilizes special methods to make better use of the information provided by censored observations. Below, we describe several models frequently used to analyze survival data.

2.1 Cox Proportional Hazard Model

COX (1972) proposed a semiparametric proportional hazards model that can be used to model survival data without prespecifying the distribution of the baseline hazard. The Cox proportional hazards (PH) model offers a method for exploring the

association of covariates with the failure time variable often seen in medical studies. The observed data consists of independent observations on the triple $(Y, \delta, \mathbf{z})$, where Y is the minimum of a failure and censoring time pair $Y = \min(X, C)$, $\delta = I(X \leq C)$ is the indicator of the event that a failure has been observed, and $\mathbf{z}' = (z_1, \cdots, z_p)$ is a p vector of covariates. The random variables X and C denote the failure and censoring times, respectively. The general layout for survival data is shown in Table 2.1. Assume that we have a survival data set with n patients. The first column identifies patients from 1 to n. The second column gives the information of survival time y_i for the i^{th} individual. The third column gives the information of censored status. The fourth column provides the information on covariates.

Table 2.1: Survival Data Layout.

Individual	Y	Δ	Covariates
1	y_1	δ_1	$z_{11} \cdots z_{1p}$
2	y_2	δ_2	$z_{21} \cdots z_{2p}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	y_n	δ_n	$z_{n1} \cdots z_{np}$

The emphasis here is on studying the effect of covariates on failure time. The effect of the covariate vector $\mathbf{z}$ on the survival probability for an individual is given by the conditional survival function $S(x|\mathbf{z}) = P(X > x|\mathbf{z})$. The conditional hazard function gives the instantaneous failure rate at $X = x$ given the covariates and survival up to time x and is defined by

$$h(x|\mathbf{z}) = \lim_{(\Delta x \to 0)} \frac{P(x \leq X \leq x + \Delta x | X \geq x, \mathbf{z})}{\Delta x},$$

which the distribution of X since $h(x|\mathbf{z}) = -\frac{S'(x|\mathbf{z})}{S(x|\mathbf{z})}$, where

$$S(x|\mathbf{z}) = \exp\left(-\int_0^x h(u|\mathbf{z})du\right).$$

The cumulative hazard function which measures the total hazard accrued up to time x is given by $H(x|\mathbf{z}) = \int_0^x h(u|\mathbf{z})du = -\log S(x|\mathbf{z})$. The requirement between X and C of independent censoring implies that

$$P(x \leq X \leq x + \Delta x | X \geq x, C \geq x, \mathbf{z}) = P(x \leq X \leq x + \Delta x | X \geq x, \mathbf{z}).$$

For more details see KALBFLEISCH and PRENTICE (2002).

For a given vector of possibly time-dependent covariates $\mathbf{z}$, the hazard function at time x is modeled as

$$h(x|\mathbf{z}) = h_o(x)g(\mathbf{z}), \tag{2.1}$$

where $h_o(x)$ is the conditional hazard function given $\mathbf{z}' = (0, \cdots, 0)_p$ and it is referred to as the baseline hazard function, and $g(\mathbf{z})$ is a nonnegative function of the covariate vector $\mathbf{z}$. Both $h_o(x)$ and $g(\mathbf{z})$ may involve unknown parameters. An interesting feature of the model is that it is semiparametric in the sense that it can be factored into a parametric part, $g(\mathbf{z})$, consisting of a vector of regression parameters associated with the covariates and a nonparametric part, $h_o(x)$, which can be left completely unspecified. The model given in Equation (2.1) is referred to as the Cox PH model.

The hazard ratio corresponding to two individuals with covariate vectors $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ is

$$\frac{h(x|\boldsymbol{z_1})}{h(x|\boldsymbol{z_2})} = \frac{g(\boldsymbol{z_1})}{g(\boldsymbol{z_2})}, \tag{2.2}$$

that is, the conditional hazard functions have a fixed ratio over time and plot as parallel lines. The model given in Equation (2.1) implies a constant relative risk as given by Equation (2.2) and the covariates affect survival by changing scale in

which the hazard is measured. Hence, the Cox PH model does not allow crossing of hazard curves. Alternative forms of the model include

$$S(x|\mathbf{z}) = \{S_o(x)\}^{g(\mathbf{z})} \tag{2.3}$$

and

$$H(x|\mathbf{z}) = \{H_o(x)\}^{g(\mathbf{z})}, \tag{2.4}$$

where $S_o(x)$ and $H_o(x)$ are the baseline survival and cumulative hazard functions respectively corresponding to $\boldsymbol{z}' = (0, \cdots, 0)_p$. The most commonly used form of the Cox PH model is

$$h(x|\mathbf{z}) = h_o(x)\exp(\boldsymbol{\theta}'\mathbf{z}) \tag{2.5}$$

where $\boldsymbol{\theta}' = (\theta_1, \cdots, \theta_p)$ is a p vector of regression coefficients. The focus is on the inference based on the sample in Table 2.1 for $\boldsymbol{\theta}$ with the baseline hazard function, $h_o(t)$, the nonparametric part, left completely unspecified.

The model given in Equation (2.5) is widely used and has been shown to be both robust and powerful. Parametric proportional hazards models also exist, which assume that the survival times follow a given distribution (for example, the Weibull distribution). Under the correct baseline hazard distribution, parametric models are more powerful than the equivalent semiparametric models. However, when using a real data, the true underlying distribution of the baseline hazard is unknown. For this reason the Cox proportional hazards model remains the method of choice for the most simple survival analysis.

2.2 Accelerated Failure-Time Models

Another widely used model linking survival time to covariates is the accelerated failure time model (AFT), which does not require some of the assumptions that proportional hazards model do.

The accelerated failure time model (AFT) states that the survival function of an individual with covariates $\mathbf{z}$ at time x is the same as the survival function of an individual with a baseline survival function at a time $x\exp(\boldsymbol{\theta}'\mathbf{z})$ where $\boldsymbol{\theta}' = (\theta_1, \cdots, \theta_p)$ is a vector of regression coefficients. In other words, the accelerated failure time model is defined by the relationship.

$$S(x|\mathbf{z}) = S_0[\exp(\boldsymbol{\theta}'\mathbf{z})x],$$

for all x. The factor $\exp(\boldsymbol{\theta}'\mathbf{z})$ is called an *acceleration factor* and tells the investigator how a change in covariate values changes the time scale from the baseline time scale. One implication of this model is that the hazard rate for an individual with covariate $\mathbf{z}$ is related to the baseline hazard rate by

$$h(x|\mathbf{z}) = \exp(\boldsymbol{\theta}'\mathbf{z})h_0[\exp(\boldsymbol{\theta}'\mathbf{z})x]$$

for all x.

The second representation of of the relationship between covariate values and survival is the linear relationship between log time and the covariate values. Here we assume the usual linear model for log time and the covariate values. That is,

$$U = \ln X = \mu + \boldsymbol{\theta}'\boldsymbol{z} + \sigma W, \tag{2.6}$$

where $\boldsymbol{\theta}' = (\theta_1, \cdots, \theta_p)$ is a vector of regression coefficients, σ is a scale parameter, μ is the effect of the gene after having removed the covariate effect and W is the error distribution. Here the distribution of error term is known. The goal is to estimate

μ, σ and $\boldsymbol{\theta}'$. The regression parameters have an interpretation similar to those in standard regression theory.

A variety of models can be used for W or, equivalently for S_0; for example, having a distribution of extreme value yields a Weibull distribution for X. Other common choices for error distribution include the standard normal distribution for W which gives a log normal regression model for X or a logistic distribution for W which yields a log logistic regression model for X. The Weibull model is very flexible because it has a hazard rate that can be either increasing, decreasing or constant. It has the unique property that, along with the accelerated failure time representation, it is the only parametric distribution that also has a proportional hazard representation.

2.3 Cure Model

A cure model is a survival model where a fraction of the population is not exposed to the hazard of interest. For example, a treatment for a fatal disease prevents death (from that disease) in a proportion of those treated (the cured fraction). It may or may not delay death in the remaining individuals (the vulnerable fraction).

More specifically, consider n individuals in the study. Let ${X_i}^*$ denote the potential time-to-event for the i^{th} subject. Then ${X_i}^*$ can be written as

$$ {X_i}^* = \eta_i X_i + (1 - \eta_i)\infty, \tag{2.7} $$

where $X_i < \infty$ denotes the failure time if the subject is susceptible and η_i indicates, by the value 1 or 0, whether the i^{th} subject is susceptible or not respectively. Throughout this chapter we will use the model given in Equation (2.7) and assume that X_i satisfies Equation (2.6). We refer to this as the parametric AFT cure model.

Notice that $\delta_i = 1$ implies $\eta_i = 1$ and otherwise, η_i is unobservable when $\delta_i = 0$. It should be noted that one can also use the Cox PH model given in Equation (2.5) for X_i in Equation (2.7). For more details about the Cox PH cure model, see LIU et al. (2006).

2.4 Using a Cure Model for Interval Mapping

In our situation, let G_i denote the genotypic information of the putative QTL and $M_i = (M_{i1}, \cdots, M_{ik})$ denote the genotypes of observed markers of a target interval for the i^{th} individual where $i = 1 \cdots n$. Assume that genotype probabilities depend on the neighboring markers. In the backcross design, $G_i = 1/0$ stands for the genotypes AA/AB, respectively, and M_i takes 2^k possible combinations. In addition let $\boldsymbol{z_i}$ denote the covariates of interest, such as environmental factors which are assumed to be independent of G_i.

A logistic model is assumed for the susceptibility indicator η_i. That is,

$$\pi(G, \boldsymbol{z_i}^*) \equiv P(\eta_i = 1 | G_i, \boldsymbol{z_i}^*) = \frac{e^{\boldsymbol{\gamma}' \boldsymbol{z_i}^* + \gamma_g{}' G_i}}{1 + e^{\boldsymbol{\gamma}' \boldsymbol{z_i}^* + \boldsymbol{\gamma_g}' G_i}}, \tag{2.8}$$

where $i = 1, \cdots, n$ and $\boldsymbol{z_i}^* = (1, \boldsymbol{z_i}')'$ so that $\boldsymbol{\gamma}$ contains the intercept term. An Accelerated failure time model is used for the failure time X_i of a susceptible subject. It is clear that, in this situation,

$$U_i = \log X_i = \mu + \boldsymbol{\beta}' \boldsymbol{z_i} + \beta_g{}' G_i + \sigma W_i. \tag{2.9}$$

From Equation (2.9) one can easily show that survival function $S_{U_i}(u)$ and probability density function $f_{U_i}(u)$ are

$$\begin{aligned} S_{U_i}(u) = P(U_i > u) &= P(\mu + \boldsymbol{\beta}'\boldsymbol{z_i} + \beta_g{}'G + \sigma W_i > u) \\ &= P\left(W_i > \frac{u - \boldsymbol{\beta}'\boldsymbol{z_i} - \beta_g{}'G_i - \mu}{\sigma}\right) \\ &= S_0\left(\frac{u - \boldsymbol{\beta}'\boldsymbol{z_i} - \beta_g{}'G_i - \mu}{\sigma}\right), \end{aligned} \tag{2.10}$$

and,

$$f_{U_i}(u) = (1/\sigma) f_0\left(\frac{u - \boldsymbol{\beta}'\boldsymbol{z_i} - \beta_g{}'G_i - \mu}{\sigma}\right), \tag{2.11}$$

respectively, where $i = 1, \cdots, n$. It should be noted that the models given in Equations (2.8) and (2.9) can also handle multiple QTL.

Now our objective is to estimate the parameter vector $\boldsymbol{\alpha} = (\mu, \sigma, \boldsymbol{\gamma}, \gamma_g, \boldsymbol{\beta}, \beta_g, \theta)$ in Equations (2.8) and (2.9) using observed data $\{Y_i, \delta_i, \boldsymbol{z_i}, M_i\}$, $i = 1, \cdots, n$. For the simplicity of our presentation we take $Y_i^* = \log Y_i$, $i = 1, \cdots, n$. Assuming that models are correct, we use the likelihood procedure in order to estimate these parameters based on the observed data. The likelihood under the observed data $(Y_i^*, \delta_i, \boldsymbol{z_i}, M_i)$, $i = 1, \cdots, n$, is

$$L = \prod_{i=1}^{n} P(Y_i^*, \delta_i, M_i) = \prod_{i=1}^{n} \sum_{G_i, \eta_i} P(Y_i^*, G_i, \delta_i, M_i, \eta_i)$$

$$= \prod_{i=1}^{n} [P(Y_i^*, 1, \delta_i, M_i, 1) + P(Y_i^*, 1, \delta_i, M_i, 0) + P(Y_i^*, 0, \delta_i, M_i, 1) + P(Y_i^*, 0, \delta_i, M_i, 0)]$$

$$\begin{aligned} = \prod_{i=1}^{n} \{ & [f(y_i{}^*)^{\delta_i} (S(y_i{}^*))^{1-\delta_i}] P(\eta_i = 1 | G_i = 1) P(G_i = 1 | M_i) \\ & + P(\eta_i = 0 | G_i = 1) P(G_i = 1 | M_i)(1 - \delta_i) \\ & + [f(y_i{}^*)^{\delta_i} (S(y_i{}^*))^{1-\delta_i}] P(\eta_i = 1 | G_i = 0) P(G_i = 0 | M_i) \\ & + P(\eta_i = 0 | G_i = 0) P(G_i = 0 | M_i) \}, \end{aligned} \tag{2.12}$$

where f and S are given by Equations (2.10) and (2.11) respectively. We refer to Equation (2.12) as the incomplete (observed) likelihood function. Since it is difficult to get the maximum likelihood estimate of $\boldsymbol{\alpha}$ by maximizing Equation (2.12), one option is to use the EM algorithm which we describe in the next section to obtain the parameter estimates. It should be noted that the conditional probability of QTL genotype $P(G_i|M_i)$ in Equation (2.12) is obtained by assuming that this probability depends only on the nearest flanking markers. Specifically, the genotype of the putative QTL cannot be observed but can be inferred from the genotypes of the flanking markers. It is known that given the genotypes of the flanking markers, the conditional probabilities for the putative QTL to take genotype AA and AB, that is,

$$P(AA|M_1 = AA, M_2 = AA), P(AB|M_1 = AA, M_2 = AA);$$

$$P(AA|M_1 = AA, M_2 = AB), P(AB|M_1 = AA, M_2 = AB);$$

$$P(AA|M_1 = AB, M_2 = AA), P(AB|M_1 = AB, M_2 = AA);$$

$$P(AA|M_1 = AB, M_2 = AB), P(AB|M_1 = AB, M_2 = AB),$$

are given in Table 2.2. In Table 2.2, r is the recombination fraction between $M1$ and $M2$, r_R is the right recombination fraction between QTL and M_2 and r_L is the right recombination fraction between $M1$ and QTL.

Table 2.2: Conditional Probabilities of a Putative QTL Given Flanking Marker Genotypes for a Backcross Population.

M1	M2	AA	AB
AA	AA	$(1-r_L)(1-r_R)/(1-r)$	$r_L r_R/(1-r)$
AA	AB	$(1-r_L)r_R/r$	$r_L(1-r_R)/r$
AB	AA	$r_L(1-r_R)/r$	$(1-r_L)r_R/r$
AB	AB	$r_L r_R/(1-r)$	$(1-r_L)(1-r_R)/(1-r)$

2.5 Using an EM Algorithm for the Accelerated Failure Time Cure Model

In this section we obtain maximum likelihood estimates of the unknown parameters $\boldsymbol{\alpha} = (\mu, \sigma, \boldsymbol{\beta}, \beta_g, \boldsymbol{\gamma}, \theta)$ using the EM algorithm. The idea behind the EM algorithm is to replace the difficult likelihood maximization with a sequence of easier maximization whose limit is the answer of the original problem. We will consider two different likelihoods: complete and observed (incomplete) likelihoods. The likelihood we would like to maximize is the incomplete likelihood given in Equation (2.12) but we will actually solve this problem in the case of complete likelihood.

In this case, the likelihood function for the complete data $(Y_i^*, \delta_i, \eta_i, G_i, M_i, \mathbf{z_i})$ is

$$\begin{aligned} L_{\text{Complete}} &= \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\gamma}' \boldsymbol{z_i}^* + \gamma_g{}' G_i}}{1 + e^{\boldsymbol{\gamma}' \boldsymbol{z_i}^* + \gamma_g{}' G_i}} \right)^{\eta_i} \left(\frac{1}{1 + e^{\boldsymbol{\gamma}' \boldsymbol{z_i}^* + \gamma_g{}' G_i}} \right)^{(1-\eta_i)} \\ &\quad \times \prod_{i=1}^{n} (1/\sigma) f_0 \left(\frac{y_i{}^* - \boldsymbol{\beta}' \boldsymbol{z_i} - \beta_g{}' G_i - \mu}{\sigma} \right)^{\delta_i \eta_i} \\ &\quad \times S_0 \left(\frac{y_i{}^* - \boldsymbol{\beta}' \boldsymbol{z_i} - \beta_g{}' G_i - \mu}{\sigma} \right)^{(1-\delta_i)\eta_i} \\ &\quad \times \prod_{i=1}^{n} P(G_i | M_i). \end{aligned} \tag{2.13}$$

This likelihood can also be expressed in terms of hazard function $h(Y_i^*)$. That is,

$$\begin{aligned} L_{\text{Complete}} &= \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\gamma}' \boldsymbol{z_i}^* + \gamma_g{}' G_i}}{1 + e^{\boldsymbol{\gamma}' \boldsymbol{z_i}^* + \gamma_g{}' G_i}} \right)^{\eta_i} \left(\frac{1}{1 + e^{\boldsymbol{\gamma}' \boldsymbol{z_i}^* + \gamma_g{}' G_i}} \right)^{(1-\eta_i)} \\ &\quad \times \prod_{i=1}^{n} (1/\sigma) h_0 \left(\frac{y_i{}^* - \boldsymbol{\beta}' \boldsymbol{z_i} - \beta_g{}' G_i - \mu}{\sigma} \right)^{\delta_i \eta_i} S_0 \left(\frac{y_i{}^* - \boldsymbol{\beta}' \boldsymbol{z_i} - \beta_g{}' G_i - \mu}{\sigma} \right)^{\eta_i} \\ &\quad \times \prod_{i=1}^{n} P(G_i | M_i), \end{aligned} \tag{2.14}$$

where h_o is the hazard for W and $P(G_i|M_i)$ can be obtained from Table 2.2. From Equations (2.13) and (2.14) it is clear that $L_{Complete}$ is a function of unobserved quantities $\{\eta_i, G_i, \eta_i G_i\}$. The EM algorithm consists of two steps: E-step and M-step.

In the E-step we will calculate the conditional expectation of l_{complete}, with respect to unobserved quantities $\{\eta_i, G_i, \eta_i G_i\}$ given the current estimated parameter values and the observed data $O_i = \{Y_i^*, \delta_i, M_i, \boldsymbol{z_i}\}$. Note that $l_{\text{complete}} = \log(L_{Complete})$. In order to calculate $E(l_{complete})$, we need to calculate several conditional expectations. Here those conditional expectations are derived only for backcross design. However, the same procedure can be used for the other multiple crosses including intercross. In order to calculate $E(l_{complete})$ we need to find conditional expectations of unobserved quantities. Below we describe how to do it.

Let $P_i = \{P_{\eta_i}, P_{G_i}, P_{\eta_i G_i}\}, i = 1, \cdots, n$ denote the corresponding conditional expectations of $\{\eta_i, G_i, \eta_i G_i\}$ given the observed data. Also, let $P(\eta_i = 1|G_i = 1) = \pi_i(1)$, $P(\eta_i = 0|G_i = 1) = 1 - \pi_i(1)$, $P(\eta_i = 1|G_i = 0) = \pi_i(0)$, $P(\eta_i = 0|G_i = 0) = 1 - \pi_i(0)$ where $\pi_i(1)$ and $\pi_i(0)$ are defined in Equation (2.8) and $P(M_i, \boldsymbol{z_i}, \delta_i) = \zeta_i$.

If $\delta_i = 1$ then $\eta_i = 1$, then $P_{\eta_i} = E(\eta_i|\zeta_i, y_i)$ is 1. If $\delta_i = 0$, then

$$\begin{aligned} E(\eta_i|\zeta_i, {y_i}^*) &= P(\eta_i = 1|\zeta_i, {y_i}^*) + 0P(\eta_i = 0|\zeta_i, {y_i}^*) \\ &= P(\eta_i = 1, G_i = 1|\zeta_i, Y_i^*) + P(\eta_i = 1, G_i = 0|\zeta_i, Y_i^*) \\ &= \frac{P(\eta_i = 1, G_i = 1, \zeta_i, Y_i^*)}{P(\zeta_i, Y_i^*)} + \frac{P(\eta_i = 1, G_i = 0, \zeta_i, Y_i^*)}{P(\zeta_i, Y_i^*)} \end{aligned}$$

$$\begin{aligned} = & \frac{P({y_i}^*|\zeta_i, \eta_i = 1, G_i = 1)P(\eta_i = 1|\zeta_i, \delta_i, G_i = 1)P(G_i = 1|\zeta_i)P(\zeta_i)}{D_{i0}} \\ & + \frac{P({y_i}^*|\zeta_i, \eta_i = 1, G_i = 0)P(\eta_i = 1|\zeta_i, \delta_i, G_i = 0)P(G_i = 0|\zeta_i)P(\zeta_i)}{D_{i0}}, \end{aligned}$$

where

$$\begin{aligned} D_{i0} &= P(y_i^*|\zeta_i,\eta_i=1,G_i=1)P(\eta_i=1|\zeta_i,G_i=1)P(G_i=1|\zeta_i)P(\zeta_i) \\ &+ P(y_i^*|\zeta_i,\eta_i=0,G_i=1)P(\eta_i=0|\zeta_i,G_i=1)P(G_i=1|\zeta_i)P(\zeta_i) \\ &+ P(y_i^*|\zeta_i,\eta_i=1,G_i=0)P(\eta_i=1|\zeta_i,G_i=0)P(G_i=0|\zeta_i)P(\zeta_i) \\ &+ P(y_i^*|\zeta_i,\eta_i=0,G_i=0)P(\eta_i=0|\zeta_i,G_i=0)P(G_i=0|\zeta_i)P(\zeta_i) \end{aligned}$$

$$\begin{aligned} &= P(y_i^*|\zeta_i,\eta_i=1,G_i=1)\pi_i(1)P(G_i=1|\zeta_i)P(\zeta_i) \\ &+ P(y_i^*|\zeta_i,\eta_i=0,G_i=1)(1-\pi_i(1))P(G_i=1|\zeta_i)P(\zeta_i) \\ &+ P(y_i^*|\zeta_i,\eta_i=1,G_i=0)\pi_i(0)P(G_i=0|\zeta_i)P(\zeta_i) \\ &+ P(y_i^*|\zeta_i,\eta_i=0,G_i=0)(1-\pi_i(0))P(G_i=0|\zeta_i)P(\zeta_i) \\ &= \frac{\pi_i(1)p_iS_0(\frac{y_i^*-\boldsymbol{\beta}'\boldsymbol{z_i}-\beta_g{}'G_i-\mu}{\sigma})+\pi_i(0)(1-p_i)S_0(\frac{y_i^*-\boldsymbol{\beta}'\boldsymbol{z_i}-\mu}{\sigma})}{D_{i0}}. \end{aligned}$$

It is clear that $P_{G_i} = E(G_i|\zeta_i,y_i^*) = P(G_i=1|\zeta_i,y_i^*)+0.P(G_i=0|\zeta_i,y_i^*)$. Now if $\delta_i=1$ then $\eta_i=1$, and consequently,

$$\begin{aligned} P(G_i=1|\zeta_i,y_i^*) &= P(G_i=1,\eta_i=1|\zeta_i,y_i^*)+P(G_i=1,\eta_i=0|\zeta_i,y_i^*) \\ &= \frac{P(y_i^*,\zeta_i,\eta_i=1,G_i=1)}{P(\zeta_i,y_i^*)} \\ &= \frac{P(y_i^*|\zeta_i,\eta_i=1,G_i=1)P(\eta_i=1|\zeta_i,G_i=1)P(G_i=1|\zeta_i)P(\zeta_i)}{P(y_i^*,\zeta_i,G_i=1,\eta_i=1)+P(y_i^*,\zeta_i,G_i=0,\eta_i=1)} \\ &= \frac{P(y_i^*|\zeta_i,\eta_i=1,G_i=1)\pi_i(1)p_i}{D_{i1}}, \end{aligned}$$

where $p_i = P(G_i=1|\zeta_i)$ and

$$\begin{aligned} D_{i1} &= P(y_i^*|\zeta_i,\eta_i=1,G_i=1)\pi_i(1)p_i \\ &\quad +P(y_i^*|\zeta_i,\eta_i=1,G_i=0)\pi_i(0)(1-p_i) \end{aligned}$$

$$= \frac{f_0(\frac{y_i^*-\boldsymbol{\beta}'\boldsymbol{z_i}-\beta_g{}'G_i-\mu}{\sigma})\pi_i(1)p_i}{f_0(\frac{y_i^*-\boldsymbol{\beta}'z_i-\beta_g{}'-\mu}{\sigma})\pi_i(1)p_i+f_0(\frac{y_i^*-\boldsymbol{\beta}'\boldsymbol{z_i}-\mu}{\sigma})\pi_i(0)(1-p_i)}.$$

Now, if $\delta_i = 0$, then

$$\begin{aligned} P(G_i = 1|\zeta_i, {y_i}^*) &= \frac{P({y_i}^*, \zeta_i, \eta_i = 1, G_i = 1)}{P(\zeta_i, {y_i}^*)} + \frac{P({y_i}^*, \zeta_i, \eta_i = 0, G_i = 1)}{P(\zeta_i, {y_i}^*)} \\ &= \frac{P({y_i}^*|\zeta_i, \eta_i = 1, G_i = 1)P(\eta_i = 1|\zeta_i, G_i = 1)P(G_i = 1|\zeta_i)P(\zeta_i)}{D_{i0}} \\ &\quad + \frac{P({y_i}^*|\zeta_i, \eta_i = 0, G_i = 1)P(\eta_i = 0|\zeta_i, G_i = 1)P(G_i = 1|\zeta_i)P(\zeta_i)}{D_{i0}} \\ &= \frac{S_0(\frac{{y_i}^* - \boldsymbol{\beta}'\boldsymbol{z_i} - {\beta_g}'G_i - \mu}{\sigma})\pi_i(1)p_i}{D_{i0}} + \frac{(1 - \pi_i(1))p_i}{D_{i0}}. \end{aligned}$$

Finally, $P_{\eta_i G_i} = E(\eta_i G_i|\zeta_i, {y_i}^*) = 1P(\eta_i G_i = 1|\zeta_i, {y_i}^*) + 0P(\eta_i G_i = 1|\zeta_i, {y_i}^*)$. This requires $\eta_i = 1$ and $G_i = 1$. Hence, $E(\eta_i G_i|\zeta_i, {y_i}^*) = 1.P(G_i = 1, \eta_i = 1|\zeta_i, {y_i}^*)$. Now, if $\delta_i = 1$, then

$$E(\eta_i G_i|\zeta_i, {y_i}^*) = \frac{\pi_i(1)p_i S_0(\frac{{y_i}^* - \boldsymbol{\beta}'\boldsymbol{z_i} - {\beta_g}'G_i - \mu}{\sigma}) + \pi_i(0)(1 - p_i)S_0(\frac{{y_i}^* - \boldsymbol{\beta}'\boldsymbol{z_i} - \mu}{\sigma})}{D_{i0}}.$$

If $\delta_i = 0$, then

$$E(\eta_i G_i|\zeta_i, {y_i}^*) = \frac{S_0(\frac{{y_i}^* - \boldsymbol{\beta}'\boldsymbol{z_i} - {\beta_g}'G_i - \mu}{\sigma})\pi_i(1)p_i}{D_{i0}}.$$

We are ready to get the estimates of $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \beta_g, \mu, \sigma, \boldsymbol{\gamma}, \gamma_g, \theta)$ through the M(Maxizimation)- step of the EM algorithm. In order to achieve this goal we make an initial guess, say, $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha_0}$ and we maximize the $E(l_{complete})$ with the optimization software and get the new estimate of $\boldsymbol{\alpha}$. The new estimates replace the initial guess, which is called updating. We continue to maximize $E(l_{complete})$ and update our estimates until the estimates converge.

2.6 Illustration of EM Algorithm with no Environmental Covariates

Suppose $f_0(w) = e^w \exp(-e^w)$, $\sum_{i=1}^{n} \delta_i = n_1$ and $n_2 = n - n_1$, then from Equation (2.13) for E-step

$$E(l_{complete}) = A1 + A2 + A3 + A4, \tag{2.15}$$

where A_1 and A_3 corresponds to all y_i's such that $\delta_i = 1$ and A_2 and A_4 corresponds to all y_i's such that $\delta_i = 0$ and they are given by

$$\begin{aligned} A_1 &= -n_1\mu/\sigma - n_1\log/\sigma + (\sum_{i=1}^{n_1} y_i)/\sigma - \beta_g/\sigma \sum_{i=1}^{n_1} E(G_i) - \sum_{i=1}^{n_1} e^{(y_i-\beta_g-\mu)/\sigma} E(G_i) \\ &- \sum_{i=1}^{n_1} e^{(y_i-\beta_g-\mu)/\sigma}(1 - E(G_i)), \end{aligned}$$

$$A_2 = -\sum_{j=1}^{n_2} e^{(y_j-\beta_g-\mu)/\sigma} P(\eta_j = 1, G_j = 1) - \sum_{j=1}^{n_2} e^{(y_j-\mu)/\sigma} P(\eta_j = 1, G_j = 0),$$

$$A_3 = n_1\theta + \gamma_g \sum_{i=1}^{n_1} E(G_i) - [\sum_{i=1}^{n_1} \log(1 + e^{\theta+\gamma_g}) E(G_i) + \sum_{i=1}^{n_1} \log(1 + e^{\theta})(1 - E(G_i)))],$$

and

$$\begin{aligned} A_4 &= \theta \sum_{j=1}^{n_2} E(\eta_j) + \gamma_g \sum_{j=1}^{n_2} P(\eta_j = 1, G_j = 1) - [\sum_{j=1}^{n_2} \log(1 + e^{\theta+\gamma_g}) E(G_i) \\ &+ \sum_{j=1}^{n_2} \log(1 + e^{\theta})(1 - E(G_i)))]. \end{aligned}$$

Now, for the M-step we set our initial guess and we maximize Equation (2.15) with the optimization software to get the new estimates. These new estimates replace the initial guess (updating). We continue to maximize and update our parameters until the estimates converge. Any EM sequence is convergent, if the maximizer at the M-step is unique.

2.7 Estimating the Parameters Using the Imputation Method

Another option to get the MLE of unknown parameters is through imputation. Missing marker data are common in QTL experiments sometime due to the difficulty of genotyping. Here we employ multiple imputation to take the missing data into

account. It has advantages over the EM algorithm because it does not have to calculate all the expected values. The essence of the imputed genotypes is Hidden Markov Models (HMM); see BAUM (1970). The crucial point is, HMM also gives us the allowance of possibility of genotyping errors, whereas in Table 2.2 we accept the assumption that no genotyping error is made. QTL genotype probabilities, conditional on the marker data, admit Markovian property through the nearest flanking markers. Generally speaking, a Markov chain is a collection of random variables, $\{G_1, G_2, \cdots, G_m\}$, satisfying the Markov property, $\mathrm{P}(G_{i+1}|G_i, \cdots, G_1) = \mathrm{P}(G_{i+1}|G_i)$ for all i. A Hidden Markov model consists of unobservable underlying Markov Chain, $\{G_i\}$, and observable random variables $\{O_1, \cdots, O_m\}$, where each O_i depends only on the G_i, $i = 1, \cdots, m$. The hidden states, $\{G_i\}$, take values on a finite set Ω. The joint distribution of the O_i and G_i can be explained in terms of three sets of probabilities called *initiation*, *transition* and *emission* probabilities. The initial probabilities define the distribution of the initial hidden state, $\pi(g) = \mathrm{P}(G_1 = g)$. Transition probabilities are defined as $t_i(g, g') = \mathrm{P}(G_{i+1} = g'|G_i = g)$ for $i = 1, \cdots, m-1$. The emission probabilities relate the conditional distribution of the observed states given the hidden states: $e_i(g, o) = \mathrm{P}(O_i = o|G_i = g)$ for $i = 1, \cdots, m$. In our situation, we will assume that the emission probabilities are homogeneous, with $e_i(g, o) = e_i(o, g)$ for all i, g and o.

Let G_i denote the true underlying genotypes for the individual at a set of m ordered loci, and let O_i denote the observed marker "phenotype" at locus i. The genotypes are often assumed to be phase-known genotypes where we have information about which alleles are transmitted to the offspring. Under the assumption of no crossover interference in meiosis, G_i form a Markov chain.

The initiation probabilities correspond to a segregation model at a single lo-

cus(e.g. allele pairs generally segregating in precisely equal numbers); the transition probabilities are a function of the recombination fractions r_i, between adjacent markers. The set Υ corresponds to the set of observed realizations of $O_1, \cdots, O_m$. The emission probabilities involve a model for errors in genotyping, which are assumed to be common across the markers but in practice this is not really true.

Consider a backcross individual obtained from two inbred strains, A and B, where the F_1 parent crossed back to the A strain. Let $\Omega = \{AA, AB\}$, the possible genotypes at a locus. The possible marker genotypes are $\Upsilon = \{A, H, -\}$, where A is the observed phenotype for homozygous underlying genotype, H is the observed phenotype for the underlying heterozygous genotype and $-$ represents the missing value. Note the distinction in the notation in regards to distinguishing the observed marker phenotypes and their underlying genotypes. In other words, suppose we observe AA, ideally $AA = A$. However, because of the possible genotyping error this may not be the case. This problem will be accounted for by the emission probabilities as explained below. The initiation probabilities, assuming Mendel's rules, are $\pi(AA) = \pi(AB) = 1/2$. The transition probabilities are $t_i(AA, AB) = t_i(AB, AA) = r_i$, where r_i is denotes the recombination fraction between loci i and $i + 1$. Then $t_i(AA, AA) = t_i(AB, AB) = 1 - r_i$.

For emission probabilities, we assume a constant error rate in genotyping, ϵ, so that $e(AA, A) = e(AB, H) = 1 - \epsilon$ and $e(AA, H) = e(AB, A) = \epsilon$. Conditioning on the observed pattern of missing data we also have $e(AA, -) = e(AB, -) = 1$. We can think as $- = \{A \text{ or } H\}$, so that $e(AA, -) = e(AA, A) + e(AA, H) = 1$.

We will look in the calculation of the conditional QTL genotype probabilities given multipoint marker data, which are very important in understanding the interval mapping in a single QTL model. We simply look for $\mathrm{P}(G_i = g|\mathbf{O})$, where

$\mathbf{O} = (O_1, \cdots, O_m)$. This probability can be calculated summing over all possible genotypes at the other loci.

$$\begin{aligned} \mathrm{P}(G_i = g|\mathbf{O}) &= \sum_{g_1} \cdots \sum_{g_{i-1}} \sum_{g_{i+1}} \cdots \sum_{g_m} \mathrm{P}(G_1 = g_1, \cdots, G_i = g \\ &\quad , \quad G_{i+1} = g_{i+1}, \cdots, G_m = g_m|\mathbf{O}) \\ &\propto \sum_{g_1} \cdots \sum_{g_{i-1}} \sum_{g_{i+1}} \cdots \sum_{g_m} \pi(g_1) \prod_{j=1}^{m-1} t_j(g_j, g_{j+1}) \\ &\times \prod_{j=1}^{m} e(g_j, O_j). \end{aligned} \tag{2.16}$$

In a backcross with two possible genotypes, this sum will include 2^{m-1} terms. Since this summation can get complicated there are algorithms to handle this difficulty in experimental crosses. We will start by introducing two sets of probabilities in the calculation of (2.16). They are,

$$\rho_i(g) = \mathrm{P}(O_1, \cdots, O_i, G_i = g),$$

and

$$\xi_i(g) = \mathrm{P}(O_{i+1}, \cdots, O_m|G_i = g).$$

Once we compute the ρ and ξ, the probability we are interested in will follow immediately. That is,

$$\begin{aligned} \mathrm{P}(G_i = g|\mathbf{O}) &= \frac{\mathrm{P}(G_i = g, \mathbf{O})}{\mathrm{P}(\mathbf{O})} \\ &= \frac{\rho_i(g)\xi(g)}{\sum_{g'} \rho_i(g')\xi_i(g')}. \end{aligned}$$

The ρ and ξ are calculated inductively, so called by *forward* and *backward* equations respectively. Let us begin with forward equations. First, note that,

$\rho_1(g) = \mathrm{P}(O_1, G_1 = g) = \pi(g)e(g, O_1)$. Now, suppose that we have calculated $\rho_1(g)$ for every $g \in \Omega$. Then we have,

$$\begin{aligned}
\rho_{i+1}(g) &= \mathrm{P}(O_1, \cdots, O_i, G_{i+1} = g) \\
&= \sum_{g'} \mathrm{P}(O_1, \cdots O_i, O_{i+1}, G_i = g', G_{i+1} = g) \\
&= \sum_{g'} \mathrm{P}(O_1, \cdots, O_i, G_i = g')\mathrm{Pr}(G_{i+1} = g|G_i = g')\mathrm{Pr}(O_{i+1} = g|G_{i+1} = g) \\
&= e(g, O_{i+1}) \sum_{g'} \rho_i(g')t_i(g', g).
\end{aligned}$$

In the third line above, conditional independence structure of the HMM,

$$\mathrm{Pr}(G_{i+1} = g|G_i = g', O_1, \cdots, O_i) = \mathrm{Pr}(G_{i+1} = g|G_i = g'),$$

and

$$\mathrm{Pr}(O_{i+1}|G_{i+1} = g, G_i = g', O_1, \cdots O_i) = \mathrm{Pr}(O_{i+1}|G_{i+1} = g),$$

are used.

Calculation of the ξ's can be done in a similar way. Specifically, let $\xi_n(g) = 1$ for all $g \in \Omega$ and suppose we have already calculated $\xi_i(g)$ for all g. Then, one can easily show that

$$\begin{aligned}
\xi_{i-1}(g) &= \mathrm{Pr}(O_i, \cdots, O_m|G_{i-1} = g) \\
&= \sum_{g'} \mathrm{Pr}(O_i, \cdots, O_m, G_i = g'|G_{i-1} = g) \\
&= \sum_{g'} \mathrm{Pr}(O_{i+1}, \cdots, O_m|G_i = g')\mathrm{Pr}(G_i = g'|G_{i-1} = g)\mathrm{Pr}(O_i = g|G_i = g') \\
&= \sum_{g'} \xi_i(g')t_{i-1}(g, g')e(g', O_i).
\end{aligned}$$

Again, in the third line we made use of the conditional independence structure of the HMM; see BROMAN (2006).

Now, we are ready to talk about the simulation of QTL genotypes through their joint distribution on the observed multipoint marker data, which is the central idea of multiple imputation. Consider a single chromosome and a single individual at a time. This simulation algorithm makes use of the ξ's; therefore we have to perform the backward equations described above. Backward and forward equations depend on the recombination fractions through transition probabilities. We can calculate recombination fractions through Haldane's map function Equation (2.17) given below, which is a model to connect the genetic distance to a recombination fraction in order to explain the crossover process. The Haldane's map function is

$$r = 1/2(1 - \exp(-2d)), \tag{2.17}$$

where d is the average number of crossovers per meiotic product measured by Morgan between two points (marker, qtl) or (marker, marker). See HALDANE (1919) for more details. In order to impute the genotypes, first we draw ${g_1}^*$ from the distribution

$$\Pr(G_1 = g|\mathbf{O}) = \frac{\rho_1(g)\xi_1(g)}{\sum_{g'} \rho_1(g')\xi_1(g')}. \tag{2.18}$$

Genotypes for further loci are drawn iteratively: having drawn ${g_1}^*, \cdots, {g_i}^*$, one draws g^*_{i+1} from the distribution

$$\begin{aligned} \Pr(G_{i+1} = g|\mathbf{O}, G_i = {g_i}^*) &= \frac{\Pr(G_{i+1} = g, G_i = {g_i}^*|\mathbf{O})}{\Pr(G_i = {g_i}^*|\mathbf{O})} \\ &= \frac{\rho_i({g_i}^*)t_i({g_i}^*, g)e(g, O_{i+1})\xi_{i+1}(g)}{\rho_i({g_i}^*)\xi_i({g_i}^*)} \\ &= \frac{t_i({g_i}^*, g)e(g, O_{i+1})\xi_{i+1}(g)}{\xi_i({g_i}^*)}. \end{aligned}$$

Since ρ's canceled out we don't have to perform forward equations. To summarize, for each individual, one first uses the backward equations to calculate the ξ's and

then simulates the chain from left to right, using the above equations. We can also use the forward equations to calculate the $\rho's$, and then simulate the chain from right to left, using the formulas analogous to above.

Once we complete imputing the genotypes including QTL locations, we can consider the likelihood based on the imputed genotypes. At this time we observe $\boldsymbol{O_{imp}} = (Y_i^*, \delta_i, G_i, M_i)\ i = 1, \cdots, n$. However, for our procedure after imputation we do not need M_i's. The likelihood based on the complete data $(Y_i^*, \boldsymbol{z_i}, \delta_i, G_i, \eta_i)$ and observed data $(Y_i^*, \boldsymbol{z_i}, \delta_i, G_i)$ are (2.19) and (2.20) respectively and are given below.

$$\begin{aligned} L_{\text{Completeimp}} &= \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\gamma}' \boldsymbol{z_i}^* + \gamma_g{}' G_i}}{1 + e^{\boldsymbol{\gamma}' \boldsymbol{z_i}^* + \gamma_g{}' G_i}} \right)^{\eta_i} \left(\frac{1}{1 + e^{\boldsymbol{\gamma}' \boldsymbol{z_i}^* + \gamma_g{}' G_i}} \right)^{(1-\eta_i)} \\ &\quad \times \prod_{i=1}^{n} (1/\sigma) f_0 \left(\frac{y_i{}^* - \boldsymbol{\beta}' \boldsymbol{z_i} - \beta_g{}' G_i - \mu}{\sigma} \right)^{\delta_i \eta_i} \\ &\quad \times S_0 \left(\frac{y_i{}^* - \boldsymbol{\beta}' \boldsymbol{z_i} - \beta_g{}' G_i - \mu}{\sigma} \right)^{(1-\delta_i)\eta_i}, \end{aligned} \tag{2.19}$$

$$\begin{aligned} L_{impobserved} = \prod_{i=1}^{n} P(Y_i^*, G_i, \delta_i, M_i) &= \prod_{i=1}^{n} \sum_{\eta_i} P(Y_i^*, G_i, \delta_i, \eta_i) \\ &= \prod_{i=1}^{n} [P(Y_i, G_i, \delta_i, 1) + P(Y_i, G_i, \delta_i, 0)] \\ &= \prod_{i=1}^{n} [f(y_i)]^{\delta_i} [S(y_i)]^{1-\delta_i} P(\eta_i = 1 | G_i) \\ &\quad + (1 - \delta_i) P(\eta_i = 0 | G_i) \\ &= \prod_{i=1}^{n} [f(y_i)]^{\delta_i} [S(y_i)]^{1-\delta_i} \frac{e^{\theta + \gamma_g G_i + \boldsymbol{\beta z_i}}}{1 + e^{\theta + \gamma_g G_i + \boldsymbol{\beta z_i}}} \\ &\quad + (1 - \delta_i) \frac{1}{1 + e^{\theta + \gamma_g G_i + \boldsymbol{\beta z_i}}}. \end{aligned} \tag{2.20}$$

As in the previous section, it is difficult to maximize the $L_{impobserved}$ directly; hence we use the EM algorithm once again and maximize $E(l_{completeimp})$ where $l_{completeimp} = \log(L_{completeimp})$. As we mentioned before, the only advantage is that we do not need to find $E(G_i)$. However, to implement our approach we still need $E(\eta_i)$, which is given below.

$$E(\eta_i) = \delta_i + \frac{(1-\delta_i)\pi(G_i, \boldsymbol{z_i}^*)S_0(y_i{}^* - \beta \boldsymbol{z_i}^* - \beta_g G_i)}{[1-\pi(G_i, \boldsymbol{z_i}^*) + \pi(G_i, \boldsymbol{z_i}^*)S_0(y_i{}^* - \beta \boldsymbol{z_i}^* - \beta_g G_i)]}. \tag{2.21}$$

Now, we start with initial values for our parameters, compute $E(\eta_i)$ in (2.21) and then maximize $E(l_{completeimp})$ with respect to unknown parameters for the E-step of the EM algorithm. Use updated values of parameters to compute $E(\eta_i)$ and then maximize $E(l_{completeimp})$ in the M-step with respect to unknown parameters again. Continue this procedure until the algorithm converges.

2.8 Illustration of Our Method Using the Listeria Data

In this section, we apply our methods described in Sections 2.6 and 2.7 to the Listeria monocytogenes data set by BOYARTCHUK, BROMAN, MOSHER, D'ORAZIO, STARNBACH (2001) and BROMAN (2003).

The data set consists of 120 age-matched (9 weeks of age) female $BALB/cByJ \times C57BL/6ByJ$ intercross (CB6F2/ByJ) mice that were infected by intravenous injection with 300 μl of 3×10^4 cfu *Listeria monocytogenes* (strain 1040s), and were monitored to discover their time of death to within eight hours. All moribund animals were recorded as dead. Animals surviving past 264 hours ($\delta_i = 0$) were considered recovered from the disease. There are a total of 133 genetic markers. Four mice were excluded from the analysis for the technical reasons, such as under or overdosing of the animals. Genomic DNA was extracted from all $BALB/cByJ \times C57BL/6ByJ$ animals. MapPairs primer sets for genetic mapping were selected based on the allelic

difference of at least 8bp. The initial genetic markers on which the CB6F2/ByJ individuals were genotyped covered all autosomes with a spacing of 20 cM. Additional markers were added to regions showing linkage in preliminary analysis. The inbred mouse strains BALB/cByJ and C57BL/6ByJ show differences in their susceptibility to *Listeria monocytogenes* infection. Intravenous infection with $2 - 5 \times 10^4$ cfu of *Listeria* leads invariably to death within 72 hours for BALB/cByJ animals, while similarly infected C57BL/6ByJ mice survive indefinitely. Among 116 F2 intercross progeny infected with 3×10^4 cfu *Listeria*, 70% died within 264 hours, with an average (SD) time to death of 106 (33) hours. The other 30% recovered from the infection. Traditional interval mapping, which relies on normally distributed traits, is thus not appropriate. Right-censored observations, in which time-to-death for the surviving mice in this cross, are not really right-censored: the observed times-to-death were not truncated due to the limited length of the study, but rather the mice recovered from the infection to die considerably later from other causes. Thus, this cross presents a novel quantitative trait requiring the development of new methods for QTL mapping.

For our use we convert intercross into backcross by combining the two homozygous genotypes in an intercross and use it as homozygous backcross genotype and keep the heterozygous genotype of intercross the way it is. We looked at the chromosomes 1,5,13. Because chromosome 13 showed promising results from the other's investigations as well as our investigation, we decided to analyze only chromosome 13 where $m = 29$, $n = 116$ and $k = 12$. We investigated the position of a QTL using the EM algorithm for accelerated failure time model using extreme value distribution for the error term on chromosome 13, which has 29 locations with 12 markers on that chromosome. We used *Mathematica 5.2* and *NMaximize* function for the

maximization. Our parameters of interest are β_g and γ_g. The results are given in Tables 2.3, 2.4, 2.5 and 2.6. It should be noted that we have also analyzed this data using *MATLAB*. We have used the *fminsearch* function for the maximization.

Table 2.3: Locations with Observed Loglikelihood Values for EM Method.

Location	Loglikelihood
1	95.8916
2	−95.5331
3	−95.463
4	−95.3412
5	−95.2299
6	−95.1757
7	−95.1846
8	−95.1916
9	−94.6208
10	−94.2808
11	−94.2807
12	−93.9575
13	−93.3808
14	−93.0996
15	−93.0873
16	−92.721
17	−92.4764
18	−91.4167
19	−86.4101
20	−84.6713
21	−82.7725
22	−82.6634
23	−84.3528
24	−85.2296
25	−85.2256
26	−85.9162
27	−87.2473
28	−88.7776
29	−90.2764

Table 2.4: Locations with Observed Loglikelihood Values for Imputation Method.

Location	Loglikelihood
1	−95.6766
2	−95.0744
3	−94.2148
4	−95.2939
5	−93.7977
6	−95.5965
7	−94.4743
8	−94.8724
9	−94.0983
10	−93.6042
11	−93.6042
12	−94.4427
13	−95.1569
14	−94.2352
15	−93.0881
16	−92.476
17	−92.476
18	−91.4689
19	−89.5705
20	−84.6723
21	−83.0931
22	−81.8923
23	−85.0924
24	−85.0924
25	−85.0924
26	−88.2173
27	−88.6223
28	−89.562
29	−90.2201

Table 2.5: Estimates of Parameters and Standard Errors for EM Method.

Parameters	Estimate	Bias	SSE
β	-0.273	0.011	0.0047
σ	0.269	-0.06	0.0009
μ	4.930	-0.006	0.003
γ	1.278	0.075	0.022
θ	0.235	-0.022	0.012

Table 2.6: Estimates of Parameters and Standard Errors for Imputation Method.

Parameters	Estimate	Bias	SSE
β	-0.284	0.0001	0.005
σ	0.267	-0.007	0.0014
μ	4.934	-0.024	0.0042
γ	1.330	0.019	0.029
θ	0.218	-0.005	0.016

From Tables 2.3 and 2.4, it is clear that the location 22 has the highest log-likelihood using both EM and imputation methods. This means that location 22 is the possible QTL. Therefore, we report the parameter estimates for location 22 in Tables 2.5 and 2.6 using the EM algorithm and Imputation method respectively. From both tables one can interpret β_g such that mice with the heterozygous genotype AB at the QTL tend to have more late event times than those with homozygous genotypes AA at the QTL. When it comes to interpreting parameter γ_g, the odds that homozygous mice are not cured compared to heterozygous mice is higher. Also it is clear that both methods give almost the same estimates. It should be noted that the same conclusions were obtained by LIU et al. (2006) where they fitted the Cox cure model to this data. In Tables 2.4 and 2.5, SSE and BIAS for our estimates are obtained using the bootstrap method.

CHAPTER 3

THE NONPARAMETRIC ACCELERATED FAILURE TIME CURE MODEL FOR DETECTING QTL

In this chapter, we use the Nonparametric Accelerated Failure Time Cure model for QTL detection.

3.1 A Nonparametric Accelerated Failure Time Cure Model and its Application for QTL Detection

We start with the model,

$$U^* = \ln X = \mu + \boldsymbol{\beta}'\boldsymbol{z} + \beta_g{}'G + \epsilon. \tag{3.1}$$

Notice that in Equation (3.1) we replaced σW in Equation (2.6) with ϵ. In this model we make no assumptions on the distribution of the error term. The model in Equation (3.1) is often called the nonparametric accelerated failure time model. We now combine Equation (3.1) and Equation (2.7) and refer to that as the nonparametric AFT cure model. The drawback in Equation (3.1) is that G, genotypes of QTLs, is not observable. However, this problem can be solved by imputing the genotypes for individuals as we did in Section 2.7. It should be noted that, in the previous chapter, we used both EM algorithm and imputation methods to get the parameter estimates. Since these two approaches gave very similar results, we were comfortable using the imputation method to obtain the genotypes. Now we are ready to proceed with the regular nonparametric methods in order to get the parameter estimates of $\boldsymbol{\alpha} = (\mu, \boldsymbol{\beta}, \beta_g, \boldsymbol{\gamma}, \gamma_g)$ in Equation (3.1) and Equation (2.8).

Let f_0 be the density probability function for ϵ and S_0 be the corresponding survival function. The conditional survival function of Y^*, given the patient is not cured, is $S_0(y_i{}^* - \mu - \boldsymbol{\beta'}\mathbf{z_i} - \beta_g{}'G_i)$. Let $O = (y_i{}^*, \delta_i, \mathbf{z_i}, G_i)$ denote the observed data for the i^{th} individual, $i = 1, \cdots, n$, where $y_i{}^*$ is the observed survival time for the i^{th} individual. The contribution to the likelihood from the i^{th} individual is $\pi(G_i, \boldsymbol{z_i}) f_0(y_i{}^* - \mu - \boldsymbol{\beta}\mathbf{z_i} - \beta_g G_i)$ when $\delta_i = 1$ and $1 - \pi(G_i, \boldsymbol{z_i}) + \pi(G_i, \boldsymbol{z_i}) S_0(y_i{}^* - \mu - \boldsymbol{\beta}\mathbf{z_i} - \beta_g G_i)$ when $\delta_i = 0$. Thus the observed likelihood function can be written as

$$L_{obs}(\mu, \boldsymbol{\gamma}, \gamma_g, \boldsymbol{\beta}, \beta_g, S_0; O) \propto \prod_{i=1}^{n} [\pi(G_i, \boldsymbol{z_i}) f_0(y_i{}^* - \mu - \boldsymbol{\beta}\mathbf{z_i} - \beta_g G_i)]^{\delta_i} \times [1 - \pi(G_i, \boldsymbol{z_i}) + \pi(G_i, \boldsymbol{z_i}) S_0(y_i{}^* - \mu - \boldsymbol{\beta}\mathbf{z_i} - \beta_g G_i)]^{1-\delta_i}. \quad (3.2)$$

We can obtain the estimates through maximizing $L_{obs}(\mu, \boldsymbol{\gamma}, \gamma_g, \boldsymbol{\beta}, \beta_g, S_0; O)$ directly if the distribution of the error term, S_0, is known to us. But this is not feasible in the nonparametric AFT mixture cure model due to lack of information about S_0. Below we propose an alternative procedure for obtaining maximum likelihood estimates of $\mu, \boldsymbol{\gamma}, \gamma_g, \boldsymbol{\beta}, \beta_g$ in Equation (3.2).

Given the values of $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)$ the complete log-likelihood is

$$\begin{aligned} l_c(\mu, \boldsymbol{\gamma}, \gamma_g, \boldsymbol{\beta}, \beta_g, S_0; O, \boldsymbol{\eta}) &= l_{c_1}(\boldsymbol{\gamma}, \gamma_g; O, \boldsymbol{\eta}) \\ &+ l_{c_2}(\mu, \boldsymbol{\beta}, \beta_g, S_0; O, \boldsymbol{\eta}), \end{aligned}$$

where

$$l_{c_1}(\boldsymbol{\gamma}, \gamma_g; O, \boldsymbol{\eta}) = \sum_{i=1}^{n} \eta_i \log[\pi(G_i, \boldsymbol{z_i})] + (1 - \eta_i) \log[1 - \pi(G_i, \boldsymbol{z_i})], \quad (3.3)$$

and

$$\begin{aligned} l_{c_2}(\mu, \boldsymbol{\beta}, \beta_g, S_0; O, \boldsymbol{\eta}) &= \sum_{i=1}^{n} \eta_i \delta_i \log[h_0(y_i{}^* - \mu - \boldsymbol{\beta z_i} - \beta_g G_i] \\ &+ \eta_i (1 - \delta_i) S_0(\log(y_i{}^*) - \mu - \boldsymbol{\beta z_i} - \beta_g G_i), \quad (3.4) \end{aligned}$$

$h_0 = f_0/S_0$ is the hazard function of ϵ.

Let ${\eta_i}^{(k)} = E(\eta_i|\boldsymbol{\alpha^{(k)}}, O)$. Note that ${\eta_i}^{(k)}$ can be interpreted as the conditional probability of the i^{th} individual being cured at the k^{th} iteration of the algorithm. It is clear that

$$\eta_i^{(k)} = \delta_i + \frac{(1-\delta_i)\pi(G_i, \boldsymbol{z_i})S_0(y_i{}^* - \mu - \boldsymbol{\beta z_i} - \beta_g G_i)}{[1-\pi(G_i, \boldsymbol{z_i}) + \pi(G_i, \boldsymbol{z_i})S_0(\log(y_i{}^*) - \mu - \boldsymbol{\beta z_i} - \beta_g G_i)]}. \tag{3.5}$$

Then, $E[l_c(\mu, \boldsymbol{\gamma}, \gamma_g, \boldsymbol{\beta}, \beta_g, S_0; O, \eta|\boldsymbol{\alpha^{(k)}})] = l_{c_1}(\boldsymbol{\gamma}, \gamma_g) + l_{c_2}(\mu, \boldsymbol{\beta}, \beta_g, S_0)$, where

$$l_{c_1}(\boldsymbol{\gamma}, \gamma_g) = \sum_{i=1}^{n} {\eta_i}^{(k)} \log[\pi(G_i, \boldsymbol{z_i})] + (1 - {\eta_i}^{(k)}) \log[1 - \pi(G_i, \boldsymbol{z_i})], \tag{3.6}$$

and

$$\begin{aligned} l_{c_2}(\mu, \boldsymbol{\beta}, \beta_g, S_0) &= \sum_{i=1}^{n} {\eta_i}^{(k)} \delta_i \log[h_0(y_i{}^* - \mu - \boldsymbol{\beta z_i} - \beta_g G_i)] \\ &+ {\eta_i}^{(k)}(1-\delta_i)[S_0(y_i{}^*) - \mu - \boldsymbol{\beta z_i} - \beta_g G_i]. \end{aligned} \tag{3.7}$$

As mentioned earlier, our interest is to estimate $\boldsymbol{\alpha}$. To accomplish this we want to maximize Equations (3.3) and (3.4) with respect to the unknown parameters $\mu, \boldsymbol{\gamma}, \gamma_g, \boldsymbol{\beta}, \beta_g$ and S_0. Maximizing Equation (3.3) with respect to $\boldsymbol{\gamma}$ and γ_g can be done using the Newton-Raphson algorithm. However, maximizing Equation (3.4) with respect to $\boldsymbol{\beta}, \beta_g$ and S_0 is not easy because we do not have any information about the survival function S_0. The nonparametric approach involves the development of rank-based monotone estimating equations for estimating the unknown parameters $\mu, \boldsymbol{\beta}$ and β_g. One way to estimate the parameters $\mu, \boldsymbol{\beta}, \beta_g$ is to use the fact that if $\delta_i \log \eta_i{}^{(k)} \equiv 0$ and $\delta_i \eta_i{}^{(k)} \equiv \delta_i$, then Equation (3.4) can be written as

$$\begin{aligned} l_{c_2}(\mu, \boldsymbol{\beta}, \beta_g, S_0) &= \sum_{i=1}^{n} \delta_i \log[\eta_i{}^{(k)} h_0(y_i{}^* - \mu - \boldsymbol{\beta z_i} - \beta_g G_i)] \\ &+ {\eta_i}^{(k)} \log[S_0(y_i{}^* - \mu - \boldsymbol{\beta z_i} - \beta_g G_i)], \end{aligned} \tag{3.8}$$

which can be simplified to

$$\prod_{i=1}^{n} [\eta_i(k) h_0(y_i{}^* - \mu - \boldsymbol{\beta z_i} - \beta_g G_i)]^{\delta_i} [S_0(y_i{}^* - \mu - \boldsymbol{\beta z_i} - \beta_g G_i)]. \tag{3.9}$$

The idea is to change the AFT cure model into an AFT model with no cured part involved by modifying the error term and to follow the nonparametric procedure of JIN (2003) which is independent of the cure part.

Since $S_0(y_i{}^* - \boldsymbol{\beta}\mathbf{z_i} - \beta_g G_i)^{\eta_i^{(k)}}$ is the survival function corresponding to the hazard function $\eta_i^{(k)} h_0(y_i{}^* - \boldsymbol{\beta}\mathbf{z_i} - \beta_g G_i)$, Equation (3.9) can be written as the log-likelihood function of the AFT model with

$$U^* = \mu + \boldsymbol{\beta}\mathbf{z} + \beta_g G + \epsilon^*, \tag{3.10}$$

where the hazard function of ϵ^* is $\eta^{(k)} h_0(\epsilon^*)$. Thus, Equation (3.8) for the nonparametric AFT model can be written in terms of the standard nonparametric AFT model (except for the constant term $\eta^{(k)}$). Then we can estimate $\boldsymbol{\beta}, \beta_g$ based on the methods for the nonparametric AFT model. With a similar approach to rank estimation, see YUHYUN, WEI (2003). The estimating equation using the Gehan-type weight function is

$$Z^*{}_G(\mu, \boldsymbol{\beta}, \beta_g) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i \eta_i^{(k)} (\boldsymbol{z_i}{}^* - \boldsymbol{z_j}{}^*) I(\epsilon_i(\mu, \boldsymbol{\beta}, \beta_g) \leq \epsilon_j(\mu, \boldsymbol{\beta}, \beta_g)), \tag{3.11}$$

where

$$\boldsymbol{z_i}{}^* = \begin{pmatrix} \mu \\ z_{i1} \\ \cdots \\ z_{iK} \\ G_i \end{pmatrix}$$

is a (K+2) vector for the i^{th} subject. Now, (3.10) can be solved by constructing a convex function where $Z^*{}_G(\mu, \boldsymbol{\beta}, \beta_g)$ is the gradient of (3.10). Then we can get the parameter estimates $(\mu, \boldsymbol{\beta}, \beta_g)$ through the minimization of convex function. The regression parameter estimates obtained using (3.10) have been shown to be consistent and asymptotically normal. See JIN (2003) and YUHYUN, WEI (2003).

Estimate of the survival function S_0, at the $(k+1)^{th}$ step is then obtained by

$$\hat{S}_0^{(k+1)} = \exp\left(-\sum_{j:\tau_j<\epsilon} \frac{d_{\tau_j}}{\sum_{i\in R(\tau_j)} \eta_i^{(k)}}\right), \tag{3.12}$$

where τ_j are different uncensored failure residuals, d_{τ_j} denotes the set of failures and $R(\tau_j)$ denotes the risk set at τ_j. See ZHANG and PENG (2007) for more details. Thus, by updating S_0 at the kth step, we can update η_i's at the kth step by using Equation (3.5). This feature will enable us to maximize l_{c1} and l_{c2} because everything will be known to us through η_i and S_0.

To clarify this, first we start making an initial guess on the parameter vector $\boldsymbol{\alpha}^* = (\mu, \boldsymbol{\gamma}, \gamma_g, \boldsymbol{\beta}, \beta_g)$. In the E-step we compute the expectation of the complete likelihood with respect to the unknown variables $\eta_1, \cdots, \eta_n$'s which are functions of S_0. Then we maximize the $E(l_c)$ with the optimization software and get the new estimates of $\boldsymbol{\alpha}^*$. These new estimates replace the initial guess, which is called updating. We continue to maximize $E(l_c)$ and update our parameters until the estimates converge.

3.2 Applying Methodology to the Listeria Data Set

Now, we are ready to apply our methodology to the data set we described in Section 2.8. First we provide the observed log-likelihoods of 29 locations on Chromosome 13 for the nonparametric method in Table 3.1. From Table 3.1 it is clear that location 22 still has the highest log-likelihood. That is, we feel that location 22 is possible QTL using the nonparametric method. The parameter estimates for location 22 are given in Table 3.2.

Again, similar to parametric and imputation methods we interpret the parameter β such that mice with the heterozygous genotype AB at the QTL tend to have more late event times than those with homozygous genotypes AA at the QTL; as well,

Table 3.1: Locations with Observed Loglikelihood Values for Nonparametric Method.

Location	Loglikelihood
1	−79.547
2	−79.220
3	−79.319
4	−79.116
5	−79.05897
6	−79.139
7	−78.940
8	−78.875
9	−78.424
10	−77.904
11	−77.904
12	−78.161
13	−77.875
14	−77.330
15	−76.947
16	−76.792
17	−76.0222
18	−76.035
19	−76.274
20	−76.098
21	−74.704
22	−74.685
23	−75.783
24	−75.525
25	−75.525
26	−76.529
27	−77.449
28	−78.398
29	−78.712

Table 3.2: Estimates of Parameters and Standard Errors for Nonparametric Method.

Parameters	Estimate	Bias	SSE
β	-0.188	0.0372	0.004
μ	4.656	-0.0335	0.091
γ	1.348	0.011	0.02
θ	0.222	-0.025	0.013

we interpret γ that the odds that homozygous mice are not cured compared to heterozygous mice are higher.

It should be noted that comparing estimates given in Tables 2.5, 2.6 and 3.2, it seems all are close to each other. Here we have implemented our application using *R*. We should mention that our program is a modified version of the original *R* program written by ZHANG and PENG (2007), where they used the same model with observable covariates.

CHAPTER 4

PERMUTATION AND BOOTSTRAP

In this chapter, we move from point estimation of unknown parameters described in Chapters 2 and 3 to LOD scores. The purpose here is to identify the regions of a genome that show evidence for a trait. We show this through a genome scan of backcross data from listeria monocytogenes using standard interval mapping. See LANDER and BOTSTEIN (1989). We use interval mapping, which is performing a genome scan with a single-QTL model in which one assumes the presence of a single QTL and considers each point on a dense grid across the genome, one at a time, as the location of the putative QTL to detect loci with important marginal effects.

4.1 LOD Score and Empirical Threshold

For a given location ξ, suppose we want to test,

$$H_0 : \beta_g(\xi) = 0, \gamma_g(\xi) = 0. \tag{4.1}$$

Then it is clear that under H_0, the LOD score at a given putative location ξ can be defined as

$$LOD(\xi) \;=\; \log_{10} \frac{L_{\hat{\beta}_g(\xi),\hat{\gamma}_g(\xi),\hat{\mu}(\xi),\hat{\sigma}(\xi)}}{(L_{\hat{\mu}(\xi),\hat{\sigma}(\xi)} | \beta_g(\xi) = 0, \gamma_g(\xi) = 0)}. \tag{4.2}$$

Now, in order to find the significance thresholds, we need to find the distribution of LOD scores under H_0. In our situation, however, it is hard to find the distribution of LOD scores under H_0.

For standard interval mapping, significance thresholds are typically obtained by permutation tests; see CHURCHILL and DOERGE (1994). We show an approximate result for the null distribution of the genome-wide maximum LOD score by permutation, conditioning on observed genotypes and phenotypes: This is done simply permuting (shuffling) phenotype data relative to the genotype data. That is, we keep the genotype data intact and the observed data will be shuffled in a way that the initial genotype phenotype correspondence will not be preserved; in particular it will be randomized. We then apply our QTL method that we developed in Chapters 2 and 3 to this shuffled version of the data and obtain the set of LOD scores. The process is repeated some number of times, which results in a number of maximum LOD scores, $P_1, P_2, \cdots, P_v$, where v is the number of permutation replicates (generally $v = 1000$ or 10000). We then can use the 95^{th} percentile of the $P's$ as an estimated genome-wide LOD threshold value.

In our setup, we have individuals for which we observed their failure times as well as individuals that are censored at a fixed time. Since it does not make sense to shuffle the individuals' phenotypes for a fixed time, we shuffled all the individuals in this study. Following this procedure the estimated 95% LOD thresholds for the three methods, parametric, imputation and nonparametric for the data set are 4.56, 3.92 and 2.51 respectively.

Tables 4.1, 4.2 and 4.3 give the LOD scores for all parametric, imputation and nonparametric methods for the accelerated failure time model parameter β and for the logistic model parameter γ for all the 29 locations.

Using thresholds 4.56, 3.92 and 2.51 for parametric, imputation and nonparametric methods respectively, the results from Tables 4.1, 4.2 and 4.3 indicated evidence for QTL on chromosome 13 between locations 20 and 26, the most likely on location

Table 4.1: LOD Scores for Parametric Method for the Parameters β and γ.

Location	LOD
1	0.108
2	0.262
3	0.246
4	0.254
5	0.282
6	0.327
7	0.385
8	0.397
9	0.603
10	0.767
11	0.767
12	0.842
13	1.02
14	1.223
15	1.325
16	1.436
17	1.59
18	1.862
19	3.531
20	4.971
21	5.715
22	5.829
23	4.89
24	4.711
25	4.713
26	3.961
27	3.273
28	2.81
29	2.534

Table 4.2: LOD Scores for Imputation Method for the Parameters β and γ.

Location	LOD
1	0.198
2	0.894
3	0.833
4	0.364
5	1.014
6	0.233
7	0.72
8	0.547
9	0.884
10	1.098
11	1.098
12	0.734
13	0.424
14	0.824
15	1.322
16	1.588
17	1.588
18	2.026
19	2.85
20	4.975
21	5.661
22	6.183
23	4.86
24	4.69
25	4.69
26	3.43
27	3.26
28	2.85
29	2.56

Table 4.3: LOD Scores for Nonparametric Method for the Parameters β and γ.

Location	LOD
1	1.199
2	1.341
3	1.297
4	1.382
5	1.405
6	1.372
7	1.452
8	1.479
9	1.672
10	1.895
11	1.895
12	1.781
13	1.896
14	2.103
15	2.264
16	2.341
17	2.712
18	2.689
19	2.489
20	2.529
21	3.079
22	3.087
23	2.671
24	2.803
25	2.803
26	2.415
27	2.039
28	1.632
29	1.496

22 based on our previous results.

4.2 Interval Estimates of QTL Location

Once one has obtained evidence for a QTL, one may seek an interval estimate of the location of the QTL. The estimated QTL location by interval mapping will deviate somewhat from the true location, and so we seek the range of possible QTL locations that are supported by the data. We assume, in this section, that there is one and only one QTL on the chromosome of interest.

One approach is the bootstrap. The bootstrap was introduced as a computer-based approach in estimating the standard error of $\hat{\theta}$. In the bootstrap, we sample, with replacement, from the individuals to create a new data set of the same size as the original with replacement. With these new data, interval mapping that we developed in Chapters 2 and 3 are performed and the estimated location of the QTL recorded. The process is repeated multiple times, to create a set of QTL location estimates $\hat{\theta}_1, \cdots, \hat{\theta}_b$ which are coming from each bootstrap sample, $b = 1, \cdots, B$. The bootstrap estimate of the variance of the estimator $\hat{\theta}$, is the variance of the set $\hat{\theta}_{(b)}$, $b = 1, \cdots, B$. That is,

$$\mathrm{Var}_{boot}[\hat{\theta}] = \mathrm{Var}[\hat{\theta}_{(b)}] = \frac{\sum_{b=1}^{n}(\hat{\theta}_{(b)} - \hat{\theta}_{(.)})}{(B-1)}$$

where

$$\hat{\theta}_{(.)} = \frac{\sum_{b=1}^{n}(\hat{\theta}_{(b)})}{B}$$

In our case, if $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}, \hat{\beta}_g, \hat{\beta}, \hat{\gamma}_g, \hat{\gamma})$, for $B \to \infty$, the bootstrap estimate $\mathrm{Var}_{boot}[\hat{\theta}]$ converges to the variances of $\hat{\mu}, \hat{\sigma}, \hat{\beta}_g, \hat{\beta}, \hat{\gamma}_g$ and $\hat{\gamma}$ individually.

We are also interested in the bootstrap estimate of bias. For this, let θ be the estimator based on the original sample. Because the bias is defined as $\mathrm{Bias}[\hat{\theta}] =$

$E[\hat{\theta}] - \theta$ the bootstrap estimate of bias is $\text{Bias}[\hat{\theta}] = \hat{\theta}_{(.)} - \hat{\theta}$.

An approach for constructing a $100(1-\alpha)\%$ confidence interval is to use the upper and lower $\alpha/2$ values of the bootstrap distribution. The approaches using the full bootstrap distribution are referred as percentile confidence limits. If $\hat{\theta}_{L,\alpha/2}$ denotes the value such that only a fraction $\alpha/2$ of all bootstrap estimates are inferior to it, and likewise $\hat{\theta}_{H,\alpha/2}$ is the value exceeded by only $\alpha/2$ of all bootstrap estimates, then an approximate confidence interval is given by

$$[\hat{\theta}_{L,\alpha/2}, \hat{\theta}_{H,\alpha/2}].$$

Furthermore, confidence interval coverage is the probability that the confidence interval includes the true parameter, under repeated sampling from the same underlying population. For example, percentile bootstrap confidence is where the endpoints of the 95% confidence interval are given by the 25th and 975th sorted bootstrap values for a 1000 bootstrap samples. In our case the coverage probabilities that a QTL will be within locations 20 and 26 for three methods, parametric, imputation and nonparametric are 94%, 92% and 89%, which are given in Figures 4.1, 4.2 and 4.3 respectively. These are an empirical bootstrap confidence intervals.

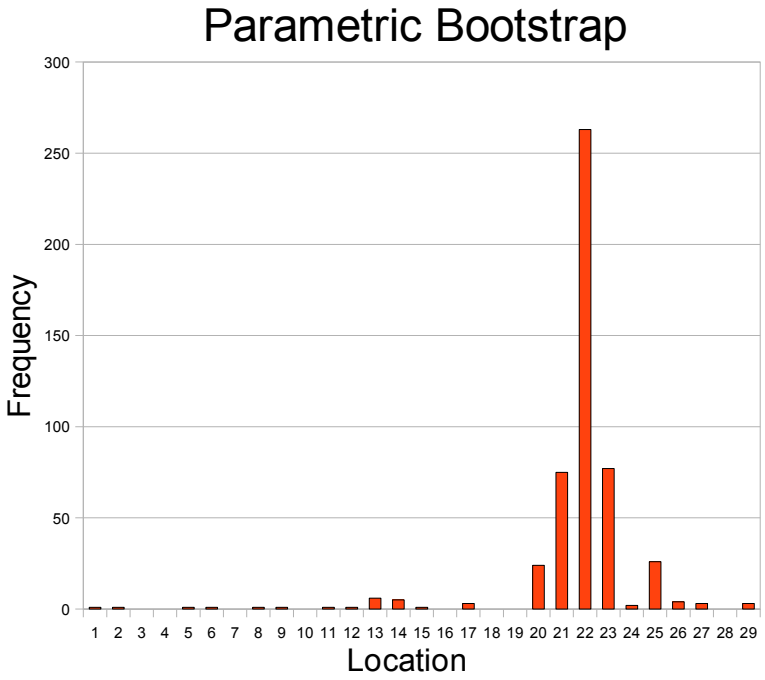

Figure 4.1: Parametric Bootstrap

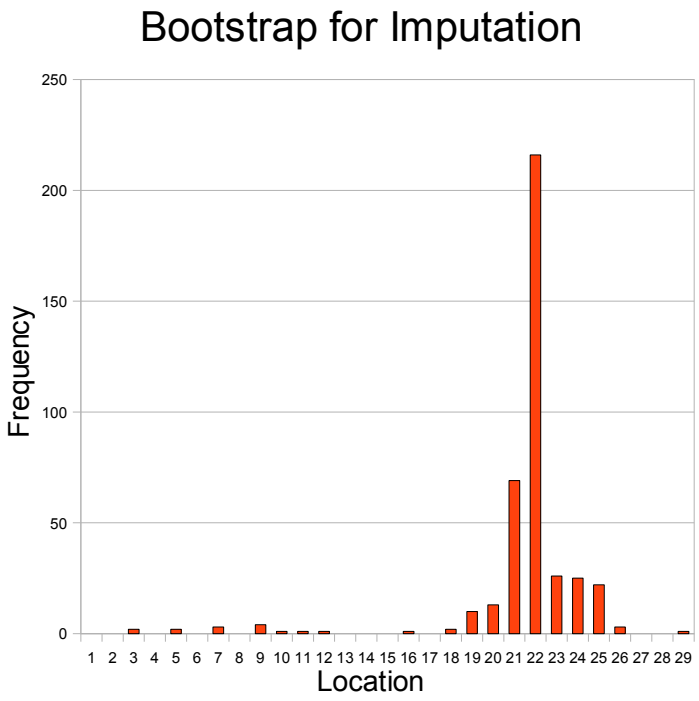

Figure 4.2: Bootstrap for Imputation

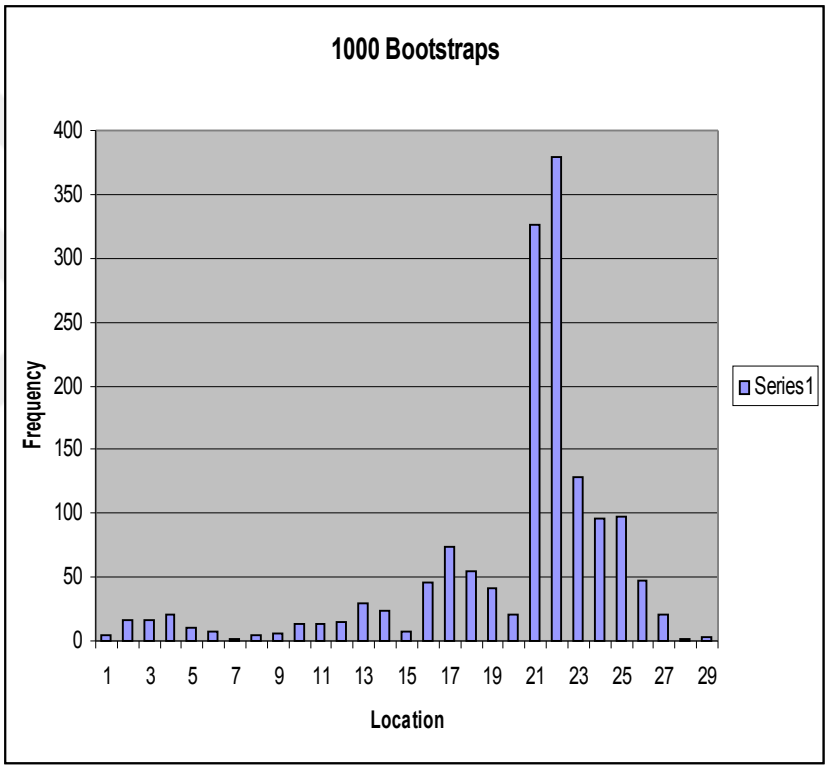

Figure 4.3: Nonparametric Bootstrap

CHAPTER 5

ASYMPTOTIC RESULTS

In Chapters 2 and 4, we proposed two methods, namely EM and imputation methods, to estimate the unknown parameters of the accelerated failure time cure model in (2.7) and to obtain the LOD scores. Our goal in this chapter is to study asymptotic behavior of LOD scores under both methods. Since these two methods have the same asymptotic behavior, (see, LITTLE and RUBIN [2002] for more details), we will only concentrate on the LOD scores based on imputation.

First consider the asymptotic behavior of LOD score for a given location, say ξ. Let $\sum_{j=1}^{n}\delta_j = q$ and let $y_1, \cdots, y_q$ corresponds to $\delta_1 = 1, \cdots, \delta_q = 1$ and $y_{q+1}, \cdots, y_n$ corresponds to $\delta_{q+1} = 0, \cdots, \delta_n = 0$ respectively. Then from Equation (2.19) we have

$$\begin{aligned} l_n(\theta,\xi) &= \sum_{i=1}^{q} log[f(y_i)\frac{e^{\mu+\gamma G_i}}{1+e^{\mu+\gamma G_i}}] + \sum_{i=q+1}^{n} log[S(y_i)\frac{e^{\mu+\gamma G_i}}{1+e^{\mu+\gamma G_i}} + \frac{1}{1+e^{\mu+\gamma G_i}}] \\ &= \sum_{i=1}^{q} logf(y_i) + \sum_{i=1}^{q}\mu+\gamma G_i - \sum_{i=1}^{q} log(1+e^{\mu+\sigma G_i}) \\ &- \sum_{i=q+1}^{n} log(1+e^{\mu+\gamma G_i}) + \sum_{i=q+1}^{n} log[S(y_i)e^{\mu+\sigma G_i}+1], \end{aligned} \tag{5.1}$$

where

$$f(y_i,\gamma^*,\mu^*,\sigma) = (1/\sigma)exp[\frac{y_i-\mu^*-\gamma^* G_i}{\sigma}]exp[-exp[\frac{y_i-\mu^*-\gamma^* G_i}{\sigma}]]$$

and $\theta = (\sigma, \mu^*, \gamma^*, \mu, \gamma)$. Equation (5.1) reduces to

$$
\begin{aligned}
&= -q\log\sigma + (1/\sigma)\sum_{i=1}^{q}(y_i - \mu^* - \gamma^* G_i) + \sum_{i=1}^{q} logS(y_i) + \sum_{i=1}^{q}(\mu + \gamma G_i) \\
&- \sum_{i=1}^{q} log(1 + e^{\mu+\gamma G_i}) + \sum_{i=q+1}^{n} log[S(y_i)e^{\mu+\gamma G_i} + 1] - \sum_{i=q+1}^{n} log(1 + e^{\mu+\gamma G_i}) \\
&= -q\log\sigma + (1/\sigma)\sum_{i=1}^{q}(y_i - \mu^* - \gamma^* G_i) + \sum_{i=1}^{q}(\mu + \gamma G_i) - \sum_{i=1}^{q} log(1 + e^{\mu+\gamma G_i}) \\
&+ \sum_{i=1}^{q} log(S(y_i)) + \sum_{i=q+1}^{n} log[\frac{1}{S(y_i)} + e^{\mu+\gamma G_i}] + \sum_{i=q+1}^{n} log(S(y_i)) \\
&- \sum_{i=q+1}^{n} log(1 + e^{\mu+\gamma G_i}) \\
&= -q\log\sigma + (1/\sigma)\sum_{i=1}^{q}(y_i - \mu^* - \gamma^* G_i) + \sum_{i=1}^{n}(\mu + \gamma G_i) - \sum_{i=1}^{n} log(1 + exp(\mu + \gamma G_i)) \\
&- \sum_{i=1}^{n} exp(y_i - \mu^* - \gamma^* G_i)/\sigma + \sum_{i=q+1}^{n} log[(\frac{exp(-(\mu + \gamma G_i))}{S(y_i)}) + 1], \qquad (5.2)
\end{aligned}
$$

where $S(y_i) = exp[-exp[\frac{y_i - \mu^* - \gamma^* G_i}{\sigma}]]$.

Using Taylor expansion of second order terms in Equation (5.2) of $l_n(\theta, \xi)$ around $\mu = \gamma = \mu^* = \gamma^* = 0$ and $\sigma = 1$ a second order approximation to (5.1) is given by,

$$
\begin{aligned}
l_n(\theta, \xi) &\approx -q\log\sigma + (1/\sigma)\sum_{i=1}^{q}(y_i - \mu^* - \gamma^* G_i) + \sum_{i=1}^{n}(\mu + \gamma G_i) \\
&- \sum_{i=1}^{n}[log2 + (1/2)\mu + (1/2)\gamma G_i + (1/8)\mu^2 + (1/8)\gamma^2 G_i] \\
&- \sum_{i=1}^{n}[e^{y_i} - e^{y_i}\mu^* - e^{y_i}\gamma^* G_i - y_i e^{y_i}(\sigma - 1) + (1/2)e^{y_i}\mu^{*2} + (1/2)e^{y_i}\gamma^{*2} G_i] \\
&- \frac{e^{e^{y_i}} G_i \gamma}{1 + e^{e^{y_i}}} - \frac{e^{y_i + e^{y_i}} G_i \gamma^*}{1 + e^{e^{y_i}}} - \frac{e^{e^{y_i}}\mu}{1 + e^{e^{y_i}}} - \frac{e^{y_i + e^{y_i}}\mu^*}{1 + e^{e^{y_i}}} - \frac{e^{y_i + e^{y_i}}(\sigma - 1)y_i}{1 + e^{e^{y_i}}} \\
&+ \frac{e^{y_i + e^{y_i}}\left(1 + e^{e^{y_i}} + e^{y_i}\right)}{2\left(1 + e^{e^{y_i}}\right)^2}\mu^{*2} + \frac{e^{y_i + e^{y_i}}\left(1 + e^{e^{y_i}} + e^{y_i}\right) G_i}{2\left(1 + e^{e^{y_i}}\right)^2}\gamma^{*2} + \frac{e^{e^{y_i}} G_i}{2\left(1 + e^{e^{y_i}}\right)^2}\gamma^2 \\
&+ \frac{e^{e^{y_i}}}{2\left(1 + e^{e^{y_i}}\right)^2}\mu^2 + \frac{e^{y_i + e^{y_i}} y_i\left(e^{y_i} y_i + y_i + e^{e^{y_i}}(y_i + 2) + 2\right)}{2\left(1 + e^{e^{y_i}}\right)^2}(\sigma - 1)^2. \qquad (5.3)
\end{aligned}
$$

The following theorem gives asymptotic behavior of $2[l_n(\hat{\theta}, \xi) - l_n(\hat{\hat{\theta}}, \xi)]$ under H_0. Here $\hat{\theta}$ is the MLE of θ under H_a and $\hat{\hat{\theta}}$ is the MLE under H_0 where $H_0 : \gamma = 0, \gamma^* = 0$. It should be noted that under H_0, $\hat{\hat{\theta}} = (\hat{\hat{\sigma}}, \hat{\hat{\mu}}^*, 0, \hat{\hat{\mu}}, 0)$.

Theorem 5.1 *As $n \to \infty$, $2[l_n(\hat{\theta}, \xi) - l_n(\hat{\hat{\theta}}, \xi)]$ goes to chi-square with two degrees of freedom.*

Proof: From the Equation (5.3), using the fact that both estimators $\hat{\theta}$ and $\hat{\hat{\theta}}$ which are obtained under H_a and H_0 respectively are consistent estimates of θ, one can easily show under the null hypothesis

$$
\begin{aligned}
2[l_n(\hat{\theta},\xi) - l_n(\hat{\hat{\theta}},\xi)] &\approx -2qlog\frac{\hat{\sigma}}{\hat{\hat{\sigma}}} \\
&+ 2\sum_{i=1}^{q}[\frac{1}{\hat{\sigma}}(y_i - \mu^* - \gamma^* G_i) - \frac{1}{\hat{\hat{\sigma}}}(y_i - \hat{\hat{\mu}}^*)] + 2[n(\hat{\mu} - \hat{\hat{\mu}}) + \sum_{i=1}^{q}\hat{\gamma}G_i] \\
&- n(\hat{\mu} - \hat{\hat{\mu}}) - (n/4)(\hat{\mu^2} - \hat{\hat{\mu}}^2) - \sum_{i=1}^{n} G_i(1/4)\hat{\gamma}^2 \\
&- 2[(\hat{\hat{\mu}}^* - \hat{\mu}^*)\sum_{i=1}^{n} e^{y_i} \\
&+ ((\hat{\hat{\sigma}} - 1) - (\hat{\sigma} - 1))\sum_{i=1}^{n} y_i e^{y_i} + (1/2)(\hat{\mu}^{*2} - \hat{\hat{\mu}}^{*2})\sum_{i=1}^{n} e^{y_i} \\
&+ ((\hat{\sigma} - 1)^2 - (\hat{\hat{\sigma}} - 1)^2)\sum_{i=1}^{n} e^{y_i}(y_i + (1/2){y_i}^2) \\
&- \hat{\gamma}^* \sum_{i=1}^{n} e^{y_i} G_i + (1/2)\hat{\gamma}^{*2}\sum_{i=1}^{n} e^{y_i} G_i \\
&- 2\sum_{i=n-q}^{n}[\frac{e^{e^{y_i}} G_i}{1 + e^{e^{y_i}}}(\hat{\gamma} - e^{y_i}\hat{\gamma^*}) + \frac{e^{y_i + e^{y_i}}}{1 + e^{e^{y_i}}}[\hat{\hat{\mu}} - \hat{\mu}] + \frac{e^{y_i + e^{y_i}}}{1 + e^{e^{y_i}}}[\hat{\hat{\mu}}^* - \hat{\mu^*}] \\
&+ \frac{e^{y_i + e^{y_i}} y_i}{1 + e^{e^{y_i}}}[(\hat{\hat{\sigma}} - 1) - (\hat{\sigma} - 1)] \\
&+ \frac{e^{y_i + e^{y_i}}\left(1 + e^{e^{y_i}} + e^{y_i}\right)}{2\left(1 + e^{e^{y_i}}\right)^2}[\hat{\mu^*}^2 - \hat{\hat{\mu}}^{*2}] + \frac{e^{e^{y_i}}}{2\left(1 + e^{e^{y_i}}\right)^2}[\hat{\mu}^2 - \hat{\hat{\mu}}^2] \\
&+ \frac{e^{y_i + e^{y_i}}\left(1 + e^{e^{y_i}} + e^{y_i}\right) G_i}{2\left(1 + e^{e^{y_i}}\right)^2}\hat{\gamma^*}^2 \\
&+ \frac{e^{e^{y_i}} G_i}{2\left(1 + e^{e^{y_i}}\right)^2}\hat{\gamma}^2 \\
&+ \frac{e^{y_i + e^{y_i}} y_i\left(e^{y_i} y_i + y_i + e^{e^{y_i}}(y_i + 2) + 2\right)}{2\left(1 + e^{e^{y_i}}\right)^2}[(\hat{\sigma} - 1)^2 - (\hat{\hat{\sigma}} - 1)^2)]] \\
&\approx \frac{\hat{\gamma}^2}{\frac{\partial l_n(\hat{\theta},\xi)}{\partial\gamma}} + \frac{\hat{\gamma^*}^2}{\frac{\partial l_n(\hat{\theta},\xi)}{\partial\gamma^*}} \qquad (5.4) \\
&\approx \frac{\hat{\gamma}^2}{(\hat{\sigma}(\hat{\gamma}))^2} + \frac{\hat{\gamma}^{*2}}{(\hat{\sigma}(\hat{\gamma}^*))^2} \\
&\approx {\chi_1}^2 + {\chi_1}^2 \\
&= {\chi_2}^2. \qquad (5.5)
\end{aligned}
$$

This completes the proof.

Since we want to control Type I error, we are interested in the limiting distribution of $max({\chi_2}^2{}_{(\xi_1)}, \cdots, {\chi_2}^2{}_{(\xi_q)})$. For that we will consider two cases.

Case 1: The markers are discrete.

Case 2: The markers are dense.

Under case 1, if we define $X_q = max(2[l_n(\hat{\theta}, \xi_1) - l_n(\hat{\hat{\theta}}, \xi_1)], \cdots, 2[l_n(\hat{\theta}, \xi_q) - l_n(\hat{\hat{\theta}}, \xi_q)]$, then the following theorem gives the asymptotic distribution of X_q.

Theorem 5.2 *Under H_0, as $q \to \infty$ asymptotic distribution of ${a_q}^{-1}(X_q - b_q)$ is the extreme value of type I. That is,*

$$P({a_q}^{-1}(X_q - b_q) \leq x) \approx exp(-e^{-x}), \tag{5.6}$$

where $a_q = 1$ and $b_q = log\ q$.

Proof: This result comes from Equation (5.4) and Theorem 5.1 of LOYNES (1965).

The next theorem gives asymptotic distribution of $max_\xi(2[l_n(\hat{\theta}, \xi) - l_n(\hat{\hat{\theta}}, \xi)])$ in the case of dense markers (case 2).

Theorem 5.3 *For case 2, under the null hypothesis an approximation to the tail distribution of the $max_\xi(2[l_n(\hat{\theta}, \xi) - l_n(\hat{\hat{\theta}}, \xi)] \geq a)$ is given by*

$$P(max_\xi 2[l_n(\hat{\theta}, \xi) - l_n(\hat{\hat{\theta}}, \xi)] \geq a) \approx 1 - exp\{-[C + \nu b^2 L(\frac{\beta_1 + \beta_2}{2})exp(-b^2/2)]\} \tag{5.7}$$

where C is the number of chromosomes, L is genetic length measured in cM, λ is the rate of crossovers and Δ is the intermarker distance,$\beta_1 = \beta_2 = 2\lambda$, $\nu = \nu(b\{\Delta(\beta_1 + \beta_2)\}^{1/2})$ and $a = b^2$.

Proof: Let $X(\xi) = \frac{\hat{\gamma}(\xi)}{\sqrt{\frac{\hat{\gamma}^2}{(\hat{\sigma}(\hat{\gamma}))^2}}}$ and $Y(\xi) = \frac{\hat{\gamma^*}(\xi)}{\sqrt{\frac{\hat{\gamma}^{*2}}{(\hat{\sigma}(\hat{\gamma}^*))^2}}}$. Following arguments of DUPUIS and SIEGMUND (1999), and the facts that both $X(\xi)$ and $Y(\xi)$ are approximately independent Ornstein-Uhlenbeck processes with mean 0 and covariance functions $e^{-2\lambda|\Delta|}$ and $e^{-2\lambda|\Delta|}$ respectively, we get the result.

As an application of Theorems 5.2 and 5.3 we provide threshold values for the discrete and dense set of markers for our data set. Under case 1, we look for 0.05 false positive threshold for chromosome 13. In other words, we want a such that

$$P[X_q > a] = 0.05.$$

Under H_0, using Theorem 5.2, we have

$$P[X_q - \log\, q > a - \log\, q] = 1 - e^{(a - \log\, q)} = 0.05. \tag{5.8}$$

In our situation we have $q = 29$ locations on chromosome 13; solving Equation (5.8) gives us a threshold value of 3.31 for the discrete case.

For dense markers, for our situation $L = 80$cM, C=1, $\lambda = 0.01$, $\Delta = 0$ hence $\nu = 1$. By simply applying Theorem 5.3, the threshold value for chromosome 13 is 2.59.

It is clear that the threshold under case 1 is closer to the empirical threshold than case 2. In fact, one expects this because the empirical threshold value was obtained based on those 29 locations on chromosome 13.

CHAPTER 6

FUTURE RESEARCH

In this chapter we describe several problems that we plan to work on in the future.

Problem 1: We discuss different paths for further research in identifying QTL that allows models such as the EHR (Extented Hazard Regression) and the nonproportinal hazards regression models.

The extended hazard regression (EHR) model was introduced for censored survival data by ETEZADI-AMOLI and CIAMPI (1985, 1987). The model includes the Cox PH and the AFT models as special cases and thus includes two Weibull distributions differing in scale parameters with a fixed shape parameter.

More specifically, let $h(x|\boldsymbol{z})$ denote the conditional hazard function for an individual with covariate vector $\boldsymbol{z}$ and let $h_0(x)$ denote the baseline hazard function. Then, in general form, the EHR model is given by $h(x|\boldsymbol{z}) = g(\boldsymbol{z})h_0[v(\boldsymbol{z})x]$, where $g(\boldsymbol{z})$ and $v(\boldsymbol{z})$ are positive functions such that $g(\mathbf{0}) = v(\mathbf{0}) = 1$. The functions $g(\boldsymbol{z})$ and $v(\boldsymbol{z})$ may involve unknown parameters. The most commonly used form for the functions g and v is the exponential function, $g(\boldsymbol{z}) = \exp(\boldsymbol{\beta'z})$ and $v(\boldsymbol{z}) = \exp(\boldsymbol{\alpha'z})$ where α and β are unknown regression parameters. Specifically, the EHR model will take the form

$$h(x|\boldsymbol{z}) = exp(\boldsymbol{\beta'z})h_0(exp(\boldsymbol{\alpha'z})x).$$

A reparameterization of the above equation gives

$$h(x|\boldsymbol{z}) = e^{(\boldsymbol{\alpha}+\boldsymbol{\beta})'\boldsymbol{z}}\lambda_0(e^{\boldsymbol{\alpha}'\boldsymbol{z}}). \tag{6.1}$$

The Cox PH model is obtained as a special case by letting $\boldsymbol{\alpha} = 0$ and the AFT model is obtained by letting $\boldsymbol{\beta} = 0$. This general model describes the situation in which the covariate vector $\boldsymbol{z}$ affects survival by changing both the time scale by the factor of $e^{\boldsymbol{\alpha}'\boldsymbol{z}}$ and the scale in which the hazard is measured by the factor $e^{\boldsymbol{\beta}'\boldsymbol{z}}$.

The conditional survival function $S(x|\boldsymbol{z})$ given the covariate vector $\boldsymbol{z}$ for this model is given by

$$\begin{aligned} S(x|\boldsymbol{z}) &= \{S_0(e^{\boldsymbol{\alpha}'\boldsymbol{z}}x)\}^{exp(\boldsymbol{\beta}'\boldsymbol{z})} \\ &= exp\{-e^{\boldsymbol{\beta}'\boldsymbol{z}}(-log\{S_0(e^{\boldsymbol{\alpha}'\boldsymbol{z}}x)\})\} \end{aligned} \tag{6.2}$$

For QTL detection, our covariates are genotypes and environmental factors. One of our future plans is to follow the procedure of ETEZADI-AMOLI and CIAMPI (1985, 1987) to estimate unknown parameters and the function $\lambda_0(x)$ in Equation (6.1) and obtain results similar to those described in Chapters 2 and 4.

Our next task is to consider nonproportional hazard regression models in which the hazard functions corresponding to different values of the covariates can cross. Specifically, we use the model where the cumulative hazard corresponding to a covariate vector $\boldsymbol{z} = (1, z_1, \cdots, z_p)'$ is given by

$$H(x|\boldsymbol{z}) = g(\boldsymbol{z})\{H_0(x)\}^{v(\boldsymbol{z})}, \tag{6.3}$$

where $H_0(x)$ is an arbitrary continuous baseline cumulative hazard function and $g(\boldsymbol{z})$ and $v(\boldsymbol{z})$ are nonnegative functions of the covariate vector $\boldsymbol{z}$ such that $g(\mathbf{0}) = v(\mathbf{0}) = \mathbf{1}$. The functions $H_0(x)$, $g(\boldsymbol{z})$ and $v(\boldsymbol{z})$ may involve unknown parameters. We refer to model (6.3) as a nonproportional hazard regression model. The CoxPH family

of models is a special case of the above model obtained by letting $v(\boldsymbol{z}) = 1$. Again, in our case since our covariates are genotypes and environmental factors, we plan to follow the procedure of DEVARAJAN (2000) to estimate unknown quantities in (6.3) and obtain results similar to those described in Chapters 2 and 4.

Problem 2: In classical studies for finding the QTL, the likelihood is maximized over unknown parameters and is then tested for whether the QTL is present or not. The Bayesian approach can also be used in identifying the QTL by simply sampling the unknowns from posterior distribution and estimating where the QTL is. In classical theory we use LOD scores representing the log of odds, which is obtained by maximizing the likelihood. However, in a Bayesian setup, we use LPD (Bayesian Log Posterior Density), which is obtained by averaging over the missing QTL genotypes. Our goal is to compute LPD for the model described in Equation (2.7) and underlying parameters in the model. To accomplish this we need prior information on parameters of Equation (2.7). Some examples of prior information are Dirichlet, the Gamma and the Beta priors for vectors of regression parameters in Equation (2.7).

Problem 3: We want to analyze different ways of genotyping and phenotyping methods in terms of information perspective for our model given in Equation (2.7). We are interested in how genotyping strategy, for example genotyping only selected individuals, or the placement of markers in the genome and its density, in the hopes of reducing cost and time, have an effect in terms of identifying the QTL. SEN(2005) assumes normal distribution for phenotype given the QTL genotype, and suggests that it is reasonable to genotype every 10cM for the cost efficiency using the Fisher information. We plan to follow his arguments and see if we get similar results for our model.

REFERENCES

ALMGREN. P., BENDAHL. P.O., BENGTSSON. H., HOSSJER.O., PERFEKT. R., *Lecture notes: Statistics in Genetics* (2003).

BATESON. W, PUNNETT.R.C,*On the Interrelations of Genetic Factors*, Proceedings of Royal Society, B. **84** (1911).

BAUM, L. E., *A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains*, The Annals of Mathematical Statistics, **41** (1970), 164–171.

BOYARTCHUK, V.L., BROMAN,K.W., MOSHER, R.E. , D'ORAZIO S.E., STARNBACH.M.N., *Multigenic Control of listeria Monocytogenes Susceptibility in Mice*, Nat Genet **27** (2001), 259–260.

BROMAN, K.W, *Mapping Quantitative Trait loci in the Case of a Spike in the Phenotype Distribution*, Genetics, **163** (2003), 1169–1175.

BROMAN,K.W., WU.H, SEN.S. and CHURCHILL G.A., *R/qtl QTL Mapping in Experimenatl Crosses*, Bioinformatics, **19** (2003), 889–890.

BROMAN, K.W., *Use of Hidden Markov Models for QTL Mapping. Working paper 125, Department of Biostatistics, Johns Hopkins University* (2006).

CHEN, H., M. HANSEN, J. JONES, R. VLAHOS, G. ANDERSON., *Detrimental Metabolic Effects of Combining long Term Cigarette Smoke Exposure and High-Fat Diet in Mice*, Am J Physiol Endocrinol Metab (2007), 1564–1571.

CHEN-HUNG K., ZENG.Z.B and TEASDALE.R.D, *Multiple Interval Mapping for Quantitative Trait Loci*, Genetics **152** (1999), 1203–1216.

CHURCHILL G.A, R.W. DOERGE *Emprical Threshold Values for Quantitative Trait Mapping*, Genetics **138** (1994), 963–971.

COX, D. R. *Regression Models and Life Tables (with Discussion).* Journal of the Royal Statistical Society, Series B **34**:(1972), 187220

DEVARAJAN. K., *Inference for a Nonproportional Hazards Regression Model and Applications*, Northern Illinois University, Ph.D Dissertation, (2000).

DUPUIS, J. and D. SIEGMUND., *Statistical Methods for Mapping Quantitative Trait loci from a Dense Set of Markers*, Genetics**151** (1999), 373–386.

ELSTON R.C., STEWART J., *A General Model for the Analysis of Pedigree Data*, Hum Hered **21** (1971),523–542.

ETAZADI-AMOLI, J., CIAMPI. A., *A General model for Testing the Proportional Hazards and the Accelerated Failure Time Hazard Hypothesis in the Analysis of Censored Survival Data with Covarites*, Communications in Statistics-A, **14** (1985), 651–667.

ETAZADI-AMOLI, J., CIAMPI. A., *Extended Hazard Regression for Censored Survival Data with Covariates: A Spline Approximation to the Baseline Hazard Function*, Biometrics **43** (1987), 181–192.

FESTING, M. F., *Inbred Strains of Rats*, Behav Genet **11** (1981), 431–435.

HALDANE, J.B.S., *The Combination of Linkage Values and the Calculation of Distances between the Loci of Linked Factors*, Journal of Genetics **8** (1919), 299–309.

HALEY, C. S. and KNOTT, S. A., *A Statistical Framework for Quantitative Trait Mapping*, Genetics **159** (1992),315–324.

JIN.Z., LIN.D.Y., WEI.L.J., and YING.Z. *Rank-based Inference for the Accelerated Failure Time Model*, Biometrika **90** (2003), 341–353.

KALBFLEISCH J.D., PRENTICE R.L, *The Statistical Analysis of Failure Time Data*, Wiley Series in Probability and Statistics, (2002), 341–353.

KAO.C.H., ZENG.Z.B. and TEASDALE.R.D., *Multiple Interval Mapping for Quantitative Trait Loci*, Genetics **152** (1999), 1203–1216.

KRUGLYAK.L., E.S. LANDER., *A Nonparametric Approach for Mapping Quantitative Trait Loci*, Genetics **139** (1995), 1421–1428.

LANDER, E.S., D. BOTSTEIN., *Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps*, Genetics **121** (1989), 185–199.

LITTLE.J.A.R., RUBIN.D.B., *Statistical Analysis with Missing Data*,second ed., Wiley Series in Probability and Statistics(2002).

LIU, M., LU, W., SHAO, Y., *Interval Mapping of Quantitative Trait Loci for Time-to-Event Data with the Proportional Hazards Mixture Cure Model*, Biometrics **62** (2006),1053–1061.

LOYNES.R.M., *Extreme Values in Uniformly Mixing Stationary Stochastic Processes*, The Annals of Mathematical Statistics **36** (1965), 993–999.

MANICHAIKUL.A., *Statistical Methods for Mapping Quantitative Trait Loci in Experimental Crosses*, Ph.D Dissertation, Department of Biostatistics, Johns Hopkins University (2007).

RABINER.L.R. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE **77** (1989), 257–286.

RUBIN, C., E. CAPILLA, Y. LUU, B. BUSA, H. CRAWFORD., *Adipogenesis is Inhibited by Brief, Daily Exposure to High-frequency, Extremely Low-magnitude Mechanical Signals*, Proc Natl Acad Sci U S A (2007).

SAWCER, S., H.B.JONES., D.JUDGE., F.VISSER, A.COMPSTON., *Emprical Genomewide Significance Levels Established by Whole Genome Simulations*, Genet Epidemiol **14** (1997), 223–229.

SEN, S., J. M. SATAGOPAN, K. W. BROMAN and G. A. CHURCHILL., *R/qtl Design: Inbred Line Cross Experimental Design*, Mamm Genome **18** (2007), 87–93.

SEN,S. and CHURCHILL,G.A., *A Simple Regression Method for Mapping Quantitative Trait Loci in Line Crosses Using Flanking markers*, Heredity **159** (2001), 371–387.

SILVER, L. M., 1995 Mouse Genetics: Concepts and Applications. Oxford University Press.

STYLIANOU.I.M., R. KORSTANJE., R.LI., S.SHEEHAN., B.PAIGEN., *Quantitaive Trait Locus Analysis for Obesity Reveals Multiple Networks of Interacting Loci*, Mamm Genome **17** (2006), 22–36.

WEILING.F., *Historical Study: Johann Gregor Mendel 1822-1884*, Am J Med Genet **40** (1991), 1–25.

WRIGHT.S., *Evolution in Mendelian Populations*, Genetics **16** (1931), 97–159.

WUSCHKE, S., S. DAHM., C. SCHMIDT., H.,G. JOOST and H. AL-HASANI., *A Metaanalysis of Quantitative Trait Loci Associated with Body Weight and Adiposity in Mice*, Int J Obes (Lond)**31** (2007), 829–841.

XU.S., W.R. ATCHLEY., *Mapping Quantitaive Trait Loci for Complex Binary Diseases Using Line Crosses*, Genetics **143** (1996), 1417–1424.

YUHYUN, P., WEI, L.J., *Estimating Subject-specific Survival Functions under the Accelerated Failure Time Model*, Biometrika **90** (2003), 717–723.

ZHAO, H. and T. P. SPEED., *On Genetic Map Functions*, Genetics **142** (1996), 1369–1377.

ZHANG.J., PENG.Y., *A New Estimation Method for the Semiparametric Accelerated Failure Time Mixture Cure Model*, Statistics in Medicine **26** (2007), 3157–3171.

APPENDIX

$mkdata <- function(imp)\{data(listeria)$
$listeria <- subset(listeria, chr = 1:19, ind = -c(30, 72, 76, 77))$
$listeria <- sim.geno(listeria, step = 2, n.draws = 1)$
$logical <- listeria\$geno[[13]]\$draws == 2$
$logical <- !logical$
$listeria\$geno[[13]]\$draws[logical] <- -1$
$listeria\$geno[[13]]\$draws[!logical] <- -0$
$data <- listeria\$geno[[13]]\$draws$
$y <- listeria\$pheno[, 1]$
$data <- matrix(data, ncol = ncol(data), nrow = nrow(data))$
$y <- matrix(y, nrow = nrow(data), ncol = 1)$
$yy <- y$
$logical <- yy == 264$
$logical <- !logical$
$yy[logical] <- -1$
$c <- c(rep(1, 3364))$
$c[!logical] <- -0$
$c <- matrix(c, ncol = 1, nrow = 3364)$
$c <- c[1:116,]$
$full.data <- cbind(y, c, data)$
$boot.data <- full.data$
$data <- boot.data[, 3:ncol(boot.data)]$
$y <- boot.data[, 1]$
$c <- boot.data[, 2]$
$data <- matrix(data, ncol = ncol(data), nrow = nrow(data))$
$y <- matrix(y, nrow = nrow(data), ncol = 1)$
$c <- matrix(c, nrow = nrow(data), ncol = 1)$
$y <- rbind(y, c)$
$temp <- matrix(data[, 1], ncol = 1, nrow = nrow(data))$
$for(i\ in\ 2:ncol(data))$
$temp <- rbind(temp, matrix(data[, i], ncol = 1, nrow = nrow(data)))$
$x <- array(temp, dim = c(nrow(data), 1, ncol(data)))$
$temp1 <- rbind(y, matrix(x[, , 1], ncol = 1, nrow = nrow(data)))$
$for(i\ in\ 2:(ncol(data)))$
$temp1 <- rbind(temp1, rbind(y, matrix(x[, , i], ncol = 1, nrow = nrow(data))))$
$data1 <- array(temp1, dim = c(nrow(data), 3, ncol(data)))$
$return(data1)$
$\}$
$H1\ starts\ here$

$ImputationsH1 <- function(imp, data1, i)\{$
$cure\_data10 <- NULL$
$cure\_data10 <- data.frame(data1[, , i])$
$x <- c("time","cens","geno")$
$names(cure\_data10) <- x$
$cure\_data <- cure\_data10$
$**********$
$lc1 = sum(w\_i^m * log(p1(geno)) + (1 - w\_i^m) * log(p0(geno)))$
$p1(geno) = exp(b[1] + b[2] * geno)/(1 + exp(b[1] + b[2] * geno))$
$p0(geno) = 1 - p1(geno)$
$**********$
$lc1 <- function(b)\{$
$temp <- rep(0, n)$
$i <- 1$
$while(i <= n)\{$
$temp[i] <- exp(sum(b * geno[i,]))$
$i <- i + 1$
$\}$
$p1 <- temp/(1 + temp)$
$- sum(w * log(p1) + (1 - w) * log(1 - p1))$
$\}$
$first\ derivative\ of\ the\ lc1$
$**********$
$dlc1 = sum(w * geno * (1 - p1) - (1 - w) * geno * p1)$
$**********$
$dlc1 <- function(b)\{$
$temp <- rep(0, n)$
$i <- 1$
$while(i <= n)\{$
$temp[i] <- exp(sum(b * geno[i,]))$
$i <- i + 1$
$\}$
$p1 <- temp/(1 + temp)$
$temp1 <- w\% * \%geno - p1\% * \%geno$
$- apply(temp1, 2, sum)$
$\}$
$get\ the\ estimation\ of\ the\ survival\ function\ from\ taylor\ (1995)$
$**********$
$lambda\_j = d\_j/(d\_j + sum(w\_l) on\ the\ risk\ set)$
$survival = prod\_\tau\_j < t(1 - lambda\_j)$
$**********$
$survival\_taylor <- function(cure\_data, error, w)\{$
$surv.object <- summary(survfit(Surv(error, cure\_data[, "cens"])))$

```
death_point <- cbind(time = surv.object$time, event = surv.object$n.event)
get the estimation of the hazard function
lambda <- rep(0, nrow(death_point))
i <- 1
while(i <= nrow(death_point)){
temp <- sum(as.numeric(error >= death_point[, "time"][i]) * w)
lambda[i] <- death_point[, "event"][i]/(temp)
i <- i + 1
}
get the estimate of the survival function
survival_value <- rep(0, length(error))
temp1 <- cumsum(lambda)
i <- 1
while(i <= length(error)){
temp2 <- sum(as.numeric(death_point[, "time"] < error[i]))
if(temp2 == length(death_point[, "time"]))survival_value[i] <- 0
elsesurvival_value[i] <- ifelse(temp2 == 0, 1, exp(-temp1[temp2]))
i <- i + 1
}
list(survival_value = survival_value)
}
lc2_modified <- function(beta){
error <- rep(0, n)
i <- 1
while(i <= n){
error[i] <- cure_data[, "time"][i] - sum(beta * geno[i, ])
i <- i + 1
}
tp <- rep(0, length(error))
i <- 1
while(i <= length(error)){
tp[i] <- sum(as.numeric((error[i] - error) < 0) * abs(error[i] - error) * w *
cure_data[, "cens"][i])
i <- i + 1
}
sum(tp)/n
}
take the log of the time
cure_data[, "time"] <- log(cure_data[, "time"])
n <- nrow(cure_data)
initial value to start the program
b <- c(2, -1)
beta <- c(0, 8)
```

```
w <- -cure_data[,"cens"]
geno <- -cbind(rep(1,n), cure_data[,"geno"])
tau <- -1
while(tau > 0.05){
temp <- -rep(0,n)
error <- -rep(0,n)
i <- -1
while(i <= n){
temp[i] <- -exp(sum(b * geno[i,]))
error[i] <- -cure_data[,"time"][i] - sum(beta * geno[i,])
i <- -i + 1
}
p1 <- -temp/(1 + temp)
p0 <- -1 - p1
update_s <- -survival_taylor(cure_data, error, w)$survival_value
update_w <- -((update_s * p1)/(p0 + p1 * update_s)) * (1 - cure_data[,"cens"]) +
cure_data[,"cens"]
tau3 <- -sum((update_w - w)^2)
w <- -update_w
update_b <- -optim(b, lc1, dlc1, method = "BFGS")$par
update_beta <- -optim(beta, lc2_modified, method = "Nelder - Mead")$par
control the convergence
tau1 <- -sum((update_b - b)^2)
tau2 <- -sum((update_beta - beta)^2)
tau <- -max(c(tau1, tau2, tau3))
b <- -update_b
beta <- -update_beta
}
lc1 <- -lc1(b)
lc2 <- -lc2_modified(beta)
return(list(lc1 = lc1, lc2 = lc2))
}
end of function ImputationsH1
H0 starts here
ImputationsH0 <- -function(imp, data1, i){
cure_data10 <- -NULL
cure_data10 <- -data.frame(data1[,,i])
x <- -c("time","cens","geno")
names(cure_data10) <- -x
cure_data <- -cure_data10
lc1 <- -function(b){
temp <- -rep(0,n)
i <- -1
```

```
while(i <= n){
temp[i] <- -exp(b)
i <- -i + 1
}
p1 <- -temp/(1 + temp)
- sum(w * log(p1) + (1 - w) * log(1 - p1))
}
the first derivative of the lc1
dlc1 = sum(w * geno * (1 - p1) - (1 - w) * geno * p1)
dlc1 <- -function(b){
temp <- -rep(0, n)
i <- -1
while(i <= n){
temp[i] <- -exp(b)
i <- -i + 1
}
p1 <- -temp/(1 + temp)
temp1 <- -((p1)^w) * ((1 - (p1))^(1-w)) * (w - p1)
temp1 <- -sum(temp1)
}
get the estimation of the survival function from taylor(1995)
lambda_j = d_j/(d_j + sum(w_l)on the risk set)
survival = prod_τ_j < t(1 - lambda_j)
survival_taylor <- -function(cure_data, error, w){
surv.object <- -summary(survfit(Surv(error, cure_data[, "cens"])))
death_point <- -cbind(time = surv.object$time, event = surv.object$n.event)
get the estimation of the hazard function
lambda <- -rep(0, nrow(death_point))
i <- -1
while(i <= nrow(death_point)){
temp <- -sum(as.numeric(error >= death_point[, "time"][i]) * w)
lambda[i] <- -death_point[, "event"][i]/(temp)
i <- -i + 1
}
get the estimate of the survival function
survival_value <- -rep(0, length(error))
temp1 <- -cumsum(lambda)
i <- -1
while(i <= length(error)){
temp2 <- -sum(as.numeric(death_point[, "time"] < error[i]))
if(temp2 == length(death_point[, "time"]))survival_value[i] <- -0
elsesurvival_value[i] <- -ifelse(temp2 == 0, 1, exp(-temp1[temp2]))
i <- -i + 1
```

```
}
list(survival_value = survival_value)
}
lc2_modified <- function(beta){
error <- rep(0, n)
i <- 1
while(i <= n){
error[i] <- cure_data[, "time"][i]
i <- i + 1
}
tp <- rep(0, length(error))
i <- 1
while(i <= length(error)){
tp[i] <- sum(as.numeric((error[i] - error) < 0) * abs(error[i] - error) * w *
cure_data[, "cens"][i])
i <- i + 1
}
sum(tp)/n
}
take the log of the time
cure_data[, "time"] <- log(cure_data[, "time"])
n <- nrow(cure_data)
initial value to start the program
b <- c(2)
beta <- c(0, 8)
w <- cure_data[, "cens"]
geno <- cbind(rep(1, n), cure_data[, "geno"])
tau <- 1
while(tau > 0.05){
temp <- rep(0, n)
error <- rep(0, n)
i <- 1
while(i <= n){
temp[i] <- exp(b)
error[i] <- cure_data[, "time"][i]
i <- i + 1
}
p1 <- temp/(1 + temp)
p0 <- 1 - p1
update_s <- survival_taylor(cure_data, error, w)$survival_value
update_w <- ((update_s * p1)/(p0 + p1 * update_s)) * (1 - cure_data[, "cens"]) +
cure_data[, "cens"]
tau3 <- sum((update_w - w)^2)
```

$w < -update\_w$
$update\_b < -optim(b, lc1, dlc1, method = "BFGS")\$par$
$update\_beta < -optim(beta, lc2\_modified, method = "Nelder - Mead")\$par$
$control\ the\ convergence$
$tau1 < -sum((update\_b - b)^2)$
$tau2 < -sum((update\_beta - beta)^2)$
$tau < -max(c(tau1, tau3))$
$b < -update\_b$
$beta < -update\_beta$
}
$lc1 < -lc1(b)$
$lc2 < -lc2\_modified(beta)$
$return(list(lc1 = lc1, lc2 = lc2))$
}
$end\ of\ function\ ImputationsH0$
$btlike13.H1 < -matrix(0, nrow = 29, ncol = 1)$
$btlike23.H1 < -matrix(0, nrow = 29, ncol = 1)$
$btlike33.H1 < -matrix(0, nrow = 29, ncol = 1)$
$btlike43.H1 < -matrix(0, nrow = 29, ncol = 1)$
$btlike53.H1 < -matrix(0, nrow = 29, ncol = 1)$
$btlike13.H0 < -matrix(0, nrow = 29, ncol = 1)$
$btlike23.H0 < -matrix(0, nrow = 29, ncol = 1)$
$btlike33.H0 < -matrix(0, nrow = 29, ncol = 1)$
$btlike43.H0 < -matrix(0, nrow = 29, ncol = 1)$
$btlike53.H0 < -matrix(0, nrow = 29, ncol = 1)$
$for(btloop\ in\ 1:1)\{Start\ super\ loop$
$cat("Bootstrap\ sample\ no.", btloop, "/n")$
$For\ imputation1$
$like13.H1 < -matrix(0, ncol = 1, nrow = 29)$
$like13.H0 < -matrix(0, ncol = 1, nrow = 29)$
$data1 < -mkdata(1)$
$for(i\ in\ 1:29)\{$
$cat("Location\ for\ Imputation1\ no = ", i, "/n")$
$result1.H1 < -ImputationsH1(1, data1, i)$
$result1.H0 < -ImputationsH0(1, data1, i)$
$sum.H1 < -result1.H1\$lc1 + result1.H1\$lc2$
$sum.H0 < -result1.H0\$lc1 + result1.H0\$lc2$
$like13.H1[i,] < -sum.H1$
$like13.H0[i,] < -sum.H0$
}
$return(list(like13.H1 = like13.H1, like13.H0 = like13.H0))$
$For\ imputation2$
$like23.H1 < -matrix(0, ncol = 1, nrow = 29)$

$like23.H0 < -matrix(0, ncol = 1, nrow = 29)$
$data1 < -mkdata(2)$
$for(i\ in\ 1:29)\{$
$cat("Location\ for\ Imputation2\ no = ", i, "/n")$
$result2.H1 < -ImputationsH1(2, data1, i)$
$result2.H0 < -ImputationsH0(2, data1, i)$
$sum.H1 < -result2.H1\$lc1 + result2.H1\$lc2$
$sum.H0 < -result2.H0\$lc1 + result2.H0\$lc2$
$like23.H1[i,] < -sum.H1$
$like23.H0[i,] < -sum.H0$
$\}$
$return(list(like23.H1 = like23.H1, like23.H0 = like23.H0))$
$For\ imputation3$
$like33.H1 < -matrix(0, ncol = 1, nrow = 29)$
$like33.H0 < -matrix(0, ncol = 1, nrow = 29)$
$data1 < -mkdata(3)$
$for(i\ in\ 1:29)\{$
$cat("Location\ for\ Imputation3\ no = ", i, "/n")$
$result3.H1 < -ImputationsH1(3, data1, i)$
$result3.H0 < -ImputationsH0(3, data1, i)$
$sum.H1 < -result3.H1\$lc1 + result3.H1\$lc2$
$sum.H0 < -result3.H0\$lc1 + result3.H0\$lc2$
$like33.H1[i,] < -sum.H1$
$like33.H0[i,] < -sum.H0$
$\}$
$For\ imputation4$
$like43.H1 < -matrix(0, ncol = 1, nrow = 29)$
$like43.H0 < -matrix(0, ncol = 1, nrow = 29)$
$data1 < -mkdata(4)$
$for(i\ in\ 1:29)\{$
$cat("Location\ for\ Imputation4\ no = ", i, "/n")$
$result4.H1 < -ImputationsH1(4, data1, i)$
$result4.H0 < -ImputationsH0(4, data1, i)$
$sum.H1 < -result4.H1\$lc1 + result4.H1\$lc2$
$sum.H0 < -result4.H0\$lc1 + result4.H0\$lc2$
$like43.H1[i,] < -sum.H1$
$like43.H0[i,] < -sum.H0$
$\}$
$For\ imputation5$
$like53.H1 < -matrix(0, ncol = 1, nrow = 29)$
$like53.H0 < -matrix(0, ncol = 1, nrow = 29)$
$data1 < -mkdata(5)$
$for(i\ in\ 1:29)\{$

$cat(\text{"}Location\ for\ Imputation5\ no = \text{"}, i, \text{"}/n\text{"})$
$result5.H1 < -ImputationsH1(5, data1, i)$
$result5.H0 < -ImputationsH0(5, data1, i)$
$sum.H1 < -result5.H1\$lc1 + result5.H1\$lc2$
$sum.H0 < -result5.H0\$lc1 + result5.H0\$lc2$
$like53.H1[i,] < -sum.H1$
$like53.H0[i,] < -sum.H0$
}
$btlike13.H1[, btloop] < -like13.H1$
$btlike23.H1[, btloop] < -like23.H1$
$btlike33.H1[, btloop] < -like33.H1$
$btlike43.H1[, btloop] < -like43.H1$
$btlike53.H1[, btloop] < -like53.H1$
$btlike13.H0[, btloop] < -like13.H0$
$btlike23.H0[, btloop] < -like23.H0$
$btlike33.H0[, btloop] < -like33.H0$
$btlike43.H0[, btloop] < -like43.H0$
$btlike53.H0[, btloop] < -like53.H0$
} super entire loop