Chapter 3

QTL Mapping in Experimental Populations

8

Genome Scanning for Quantitative Trait Loci

In the previous chapters, we learned the basic concept of quantitative genetics, the quantitative genetics model, the method for major gene detection (genotypes of the major gene are observed) and the algorithm for segregation analysis (genotypes of the major gene are not observed). We also learned some analytical techniques to analyze a molecular marker linked to a major gene. The real focus of statistical genomics, however, is to identify functional genes that are responsible for the genetic variation of quantitative traits or complex traits if they are not normally distributed. These chapters provide the necessary technology (knowledge preparation) for gene identification, which is the theme of this book.

Molecular markers are not genes but they are inherited following Mendel's laws and their genotypes are observable. The functional genes also follow Mendel's laws of inheritance but their genotypes are not observable. Since both markers and genes are carried by a limited number of chromosomes in the genome, some genes must be physically linked with some markers. If a marker sits in the neighborhood of a gene, the segregation pattern of the marker must be associated with the phenotypic variation of a trait that is controlled by the gene due to linkage. Therefore, we can study marker and trait association and hope to identify important markers that are closely linked to the gene. Since a quantitative trait is often controlled by the segregation of more than one gene, more markers are needed to identify all genes for a quantitative trait. These multiple genes are called quantitative trait loci (QTL). This chapter deals with marker-trait association study in line crosses. The association study using line crosses is different from the association study using randomly sampled populations. The former takes advantage of linkage disequilibrium while the latter assumes no linkage disequilibrium. As a result, markers associated with the trait of interest in line crosses are not equivalent to the genes while markers associated with the traits in randomly sampled populations are most likely the actual genes. The statistical methods for association study, however, are the same, regardless whether the populations are derived from line crosses or not.

8.1 The mouse data

A dataset from an F_2 mouse population consisting of 110 individuals was used as an example for the genetic analysis. The data were published by Lan et al. (2006) and are freely available from the internet. The parents of the F_2 population were B6 (29 males) and BTBR (31 females). The F_2 mice used in this study were measured for various clinical traits related to obesity and diabetes. The framework map consists of 194 microsatellite markers, with an average marker spacing of about 10 cM. The mouse genome has 19 chromosomes (excluding the sex chromosome). The data analyzed in this chapter contain 110 F_2 mice and 193 markers. The second marker (D7Mit76) on chromosome 7 was excluded from the analysis because it overlaps with the first marker (D7Mit56). The 193 markers cover about 1,800 cM of the entire mouse genome. The trait of interest was the 10th week body weight. The marker map, the genotypes of the 110 mice for the 193 markers and the 10th week body weights of the F_2 mice are also provided in the author's personal website (www.statgen.ucr.edu). The files stored in our website are not the original data but preprocessed by our laboratory members and thus they are ready for analysis using QTL mapping software packages such us the QTL procedure in SAS (Hu and Xu, 2009).

8.2 Genome scanning

In major gene identification, we used an F-test statistic to test the significance of a major gene. In genome scanning, we simply treat each marker as a major

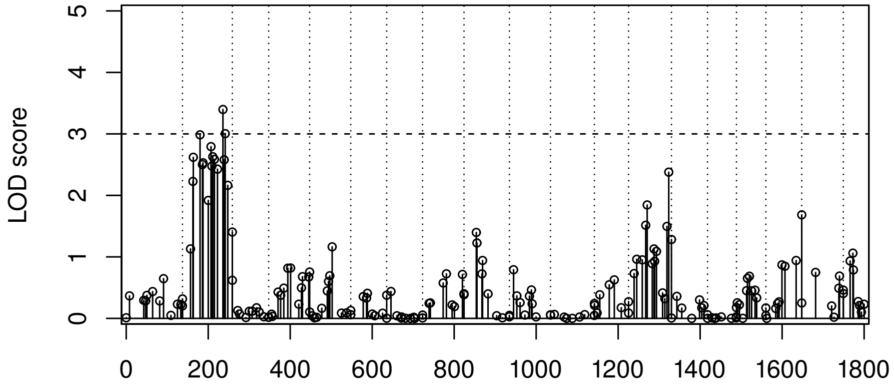

Fig. 8.1. LOD score profile of the entire genome (19 chromosomes) for the 10th week body weight of F_2 mice derived from the cross of two inbred lines. The 19 chromosomes are separated by the dotted reference lines.

gene and analyze every single marker. The test statistics of all markers across the genome are plotted against the genome location of the markers, forming a test statistic profile. Some regions of the genome may show peaks while majority of the genome may be flat. The regions with peaks may suggest QTL nearby the peaks. If the marker density is sufficiently high, some markers may actually overlap with the QTL. The genome scanning is also called individual marker analysis. We scan all markers across the genome, but with one marker at a time. The entire genome scanning requires many repeated single marker analysis. The genetic model and test statistic in genome scanning are of no difference from the major gene detection except that we now deal with multiple markers. Sometimes, investigators already have prior knowledge about the functions of some genes. The functions may be related to the development of the quantitative trait of interest. These genes are called candidate genes for the trait of interest. Genome scanning may also include these candidate genes. Figure 8.1 shows the LOD score profile of the mouse genome for the trait of 10th week body weight (wt10week). Note that the LOD (log of odds) score is often used in human genetics. The Wald-test statistic is often converted into the LOD score using (see a later section for the definition of LOD)

$$\mathrm{LOD} = \frac{W}{2 \times \ln(10)} \tag{8.1}$$

There are many peaks in the LOD score profile, but the peaks in chromosome 2 appear to be too high ($\mathrm{LOD} > 3$) to be explained by chance. Therefore, one or more QTL may exist in chromosome two for this trait.

The model used for the analysis is called the additive genetic model because the dominance effect has been ignored. Figure 8.2 shows the additive effects plotted against markers, the so called QTL effect profile. We can see that QTL effects in some regions of the genome are positive while in in other regions they are negative. The way we coded the genotypes determined the signs of the QTL effects. Assume that the original genotypic data were coded as 'A' for line B6, 'B' for line BTBR and 'H' for heterozygote. We numerically recoded the genotype as 1 for 'A', 0 for 'H' and -1 for 'B'. Based on this coding system, a negative QTL effect means that the B6 allele is "low" and the BTBR allele is "high". Therefore, the QTL allele carried by B6 in the second chromosome is the "low" allele, i.e., it is responsible for the low body weight. Of course, if 'A' and 'B' alleles represent BTBR and B6, respectively, the negative and positive signs should be explained in just the opposite way.

8.3 Missing genotypes

In the section of major gene detection, we assumed that the genotype of a major gene is observed for every individual. In the section of segregation analysis, the genotype of the major gene is missing for every individual. This section deals with marker analysis. Although most individuals in the mapping population should be genotyped for all markers, still some individuals may not be genotyped for some markers, either due to technical errors or human errors. If an individual is not genotyped for all markers, this individual should be eliminated from the analysis. However, most individuals may just have a few missing genotypes. These individuals must be included in the analysis; otherwise, we may not have enough sample size to perform genetic mapping. We now use the $\mathtt{F}_2$ population as an example to show how to deal with the missing marker problem.

Let y_j be the phenotypic value of individual j and it can be described by the following linear model,

$$y_j = b_0 + X_j b_1 + e_j, \tag{8.2}$$

where b_0 is the intercept, b_1 is the additive genetic effect, i.e., a, and e_j is the residual error. The genotype indicator variable X_j depends on the genotype of the marker under consideration. Let us define X_j as

$$X_j = \begin{cases} +1 & \text{for} A_1A_1 \\ \ \ 0 & \text{for} A_1A_2 \\ -1 & \text{for} A_2A_2 \end{cases} . \tag{8.3}$$

Let G_j be the genotype of the marker under consideration and $p_j(1) = \Pr(G_j = A_1A_1|\text{marker})$ be the probability of $G_j = A_1A_1$ given the genotypes of the two markers flanking the marker of interest. Similarly, let $p_j(0) = \Pr(G_j = A_1A_2|\text{marker})$ and $p_j(-1) = \Pr(G_j = A_2A_2|\text{marker})$. The conditional expectation of X_j given the flanking marker genotypes is

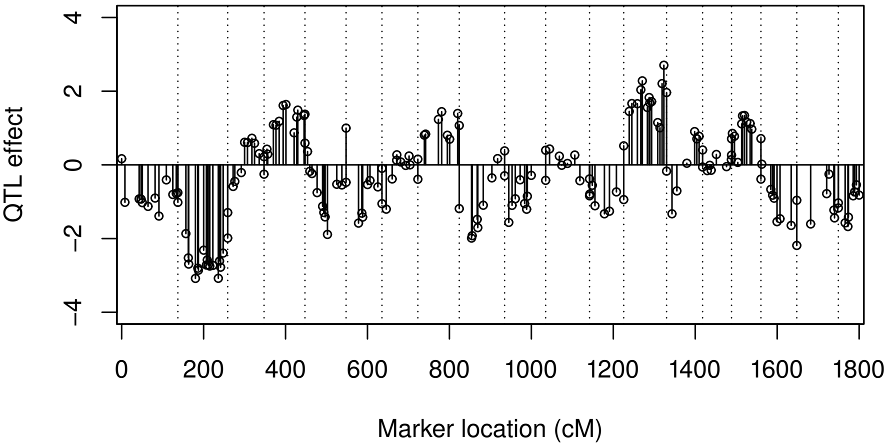

Fig. 8.2. QTL effect profile of the entire genome (19 chromosomes) for the 10th week body weight of $\mathtt{F}_2$ mice derived from the cross of two inbred lines. The 19 chromosomes are separated by the dotted reference lines.

$E(X_j) = p_j(1) - p_j(-1)$. The model for missing markers is the same as equation (8.2) except that X_j is replaced by $E(X_j)$, i.e.,

$$y_j = b_0 + E(X_j)b_1 + e_j. \tag{8.4}$$

To include the dominance effect, the revised model is

$$y_j = b_0 + E(X_{j1})b_1 + E(X_{j2})b_2 + e_j \tag{8.5}$$

where X_{j1} is the genotype indicator variable for the additive effect, as defined early (equation 8.3) and X_{j2} is the genotype indicator variable for the dominance effect,

$$X_{j2} = \begin{cases} 0 & \text{for} A_1A_1 \\ 1 & \text{for} A_1A_2 \\ 0 & \text{for} A_2A_2 \end{cases}. \tag{8.6}$$

The conditional expectation of X_{j2} is simply $E(X_j) = p_j(0)$. The second regression coefficient b_2 is the dominance effect, i.e., $b_2 = d$.

8.4 Test statistics

There are many different test statistics we can use for genome scanning. The one we learned in the major gene detection is the F-test statistic. We now discuss the test statistics when only a single model effect is subject to test. In genetic analysis, this is equivalent to testing only the additive effect. Let $\hat{b}_1$ be the estimated genetic effect and $\sigma^2_{\hat{b}_1}$ be the variance of the estimate. The F-test statistic for the null hypothesis $H_0 : b_1 = 0$ is

$$F = \frac{\hat{b}_1^2}{\sigma^2_{\hat{b}_1}}. \tag{8.7}$$

This F-test statistic appears to be different from the F-test statistic occurring in the analysis of variances (ANOVA). The latter is defined as the ratio of the between group mean squares $\mathrm{MS_B}$ to the within group mean squares $\mathrm{MS_W}$. However, the two test statistics are two different forms of the same test statistic (derivation is not shown). As an F-test statistic, it will follow an F distribution with a numerator degree of freedom 1 and a denominator degree of freedom $n - 2$.

A single genetic effect can also be tested using the t-test statistic. The t-test statistic is simply the square root of the F-test statistic,

$$t = \sqrt{F} = \frac{|\hat{b}_1|}{\sigma_{\hat{b}_1}} \tag{8.8}$$

Under the null hypothesis $H_0 : b_1 = 0$, this test statistic will follow a t distribution with $n-2$ degrees of freedom. As the sample size increases, $n-2$

is not much different from n, therefore, the degrees of freedom in the F-test and the t-test is approximately equal to the sample size.

When $n \to \infty$, the F-test will be identical to the χ^2-test statistic, which follows a χ^2 distribution with one degree of freedom. The corresponding t-test statistic will approach to the Z-test statistic.

The F-test statistic in the form of $F = \frac{\hat{b}_1^2}{\sigma^2_{\hat{b}_1}}$ is actually called the Wald-test statistic or simply W-test statistic (Wald, 1943). Although the Wald-test statistic is not called as often as the F- and t-test statistics in genome scanning, it will be used more often here in this text book due to the fact that the Wald-test statistic is comparable or similar to the likelihood ratio test statistic.

The likelihood ratio test (LRT) statistic is defined as

$$\lambda = -2[L_0(\hat{\theta}_0) - L_1(\hat{\theta}_1)] \tag{8.9}$$

where $L_0(\hat{\theta}_0)$ is the log likelihood function evaluated under the null model ($H_0 : b_1 = 0$) and $L_1(\hat{\theta}_1)$ is the log likelihood function evaluated under the full model ($H_1 : b_1 \neq 0$). The null model and the full model differ by one parameter, i.e., $\theta_0 = \{b_0, \sigma^2\}$ and $\theta_1 = \{b_0, b_1, \sigma^2\}$. We often call the null model the restricted model or reduced model because it has $b_1 = 0$ as the restriction or simply have one parameter less than the full model. Because $L_1(\hat{\theta}_1)$ is guaranteed to be larger than $L_0(\hat{\theta}_0)$, the log likelihood difference is negative. A negative test statistic looks strange, and thus we put a minus sign in front of the different to make the test statistic positive. The constant multiplier 2 is simply to make the likelihood ratio test statistic follow a standard distribution under the null model. This standard distribution happens to be a χ^2 distribution with one degree of freedom. The degree of freedom is one, not any other value, because the null model has one parameter less than the full model.

We now realize that the Wald-test statistic, the F-test statistic and the likelihood ratio test statistic, all approach a χ^2 distribution with one degree of freedom as the sample size is sufficiently large. Therefore, these three test statistics can be used interchangeably with very little difference, although the likelihood ratio test statistic is considered a slightly better test-statistic than the others.

The likelihood ratio test statistic is defined using the natural logarithm, i.e., the logarithm with base $e \approx 2.718281828459$. In human genetics, people often use the LOD (Log of Odds) score as the test statistic. Let $L_0 = L_0(\hat{\theta}_0)$ and $L_1 = L_1(\hat{\theta}_1)$ be short expressions of the natural logarithms of the likelihood functions under the null model and the full model, respectively. The original likelihood functions (before taking the natural log) are $l_0 = e^{L_0}$ and $l_1 = e^{L_1}$, respectively. The LOD score is defined as

$$\begin{aligned}\mathrm{LOD} =&\log_{10}\left(\frac{l_1}{l_0}\right) = \log_{10}\left(\frac{e^{L_1}}{e^{L_0}}\right)\\ =&\log_{10} e^{L_1} - \log_{10} e^{L_0} = \log_{10} e^{(L_1-L_0)}\end{aligned} \tag{8.10}$$

It is the log of the likelihood ratio with base 10 rather than base e. The relationship between LOD and the likelihood ratio test statistic (λ) is

$$\begin{aligned}\mathrm{LOD} =&\log_{10} e^{(L_1-L_0)} = \tfrac{1}{2}\log_{10} e^{[-2(L_0-L_1)]}\\ =&[-2(L_0-L_1)]\left(\tfrac{1}{2}\log_{10} e\right) = \lambda\left(\tfrac{1}{2}\log_{10} e\right)\end{aligned} \tag{8.11}$$

The constant $\frac{1}{2}\log_{10} e \approx 0.2171$ and the inverse of the constant is approximately 4.6052. Therefore, we may use the following approximation to convert λ to LOD,

$$\mathrm{LOD} = 0.2171\,\lambda = \frac{\lambda}{4.6052} \tag{8.12}$$

The LOD score has an intuitive interpretation because of the base 10. A LOD score of x means that the full model is 10^x times more likely than the restricted model. For example, a LOD score 3 means that the full model (with the marker effect) is 1000 times more likely than the reduce model (without the marker effect).

We now turn our attention to the hypotheses where two or more genetic effects are tested simultaneously. For example, in an $\mathtt{F}_2$ population, we can test both the additive and dominance effects. The null hypothesis is $H_0 : a = d = 0$ or $H_0 : b_1 = b_2 = 0$. In this case, the t-test is not a valid choice, because it is designed for testing only a single effect. The F-test, although can be used, is rarely chosen as the test statistic for genome scanning. The F-test statistic for testing two effects is defined as

$$F = \frac{1}{2}[\,\hat{b}_1\ \hat{b}_2\,]\begin{bmatrix}\mathrm{var}(\hat{b}_1) & \mathrm{cov}(\hat{b}_1,\hat{b}_2)\\ \mathrm{cov}(\hat{b}_1,\hat{b}_2) & \mathrm{var}(\hat{b}_2)\end{bmatrix}^{-1}\begin{bmatrix}\hat{b}_1\\ \hat{b}_2\end{bmatrix} \tag{8.13}$$

The $\frac{1}{2}$ multiplier appears because we are testing two effects. If we test k effects simultaneously, the multiplier will be $\frac{1}{k}$, and the dimensionality of the effect vector and the variance matrix will be changed to $k \times 1$ and $k \times k$ accordingly. The F-test statistic follows an F-distribution with degrees of freedom k and $n - (k + 1)$ or simply k and n when n is sufficiently large.

In contrast to the test for a single genetic effect where the F-test statistic is equivalent to the W-test statistic, when testing two or more effects, the W-test statistic is

$$W = [\,\hat{b}_1\ \hat{b}_2\,]\begin{bmatrix}\mathrm{var}(\hat{b}_1) & \mathrm{cov}(\hat{b}_1,\hat{b}_2)\\ \mathrm{cov}(\hat{b}_1,\hat{b}_2) & \mathrm{var}(\hat{b}_2)\end{bmatrix}^{-1}\begin{bmatrix}\hat{b}_1\\ \hat{b}_2\end{bmatrix} \tag{8.14}$$

The relationship between the W-test and the F-test is $W = kF$. When the sample size is sufficiently large, the W-test statistic will approach a χ^2 distribution with k degrees of freedom ($k = 2$ in this case).

The corresponding likelihood ratio test statistic for two or more effects has the same form as that for testing a single effect except that $\theta_0 = \{b_0, \sigma^2\}$ under the null model has two parameters less than the $\theta_1 = \{b_0, b_1, b_2, \sigma^2\}$ under the full model. As a result, the λ test statistic follows a χ^2 distribution with $k = 2$ degrees of freedom. Both the W-test and the likelihood ratio test statistics follow the same χ^2 distribution, and thus they can be used interchangeably with very little difference.

The W-test statistic requires calculation of the variance-covariance matrix of the estimated parameters and its inverse. However, some of the algorithms for parameter estimation, e.g., the EM algorithm, do not have an automatic way to calculate this matrix. For these methods, the likelihood ratio test statistic may be preferred because of the ease of calculating the test statistic. When both the W-test and the λ-test statistics are available, which one is better? The answer is that the λ-test statistic is more desirable if the sample size is small. For large samples sizes, these two test-statistics are virtually the same.

In summary, the W-test and the λ-test statistics are preferable for genome scanning because they can be used for testing both a single effect and multiple effects (compared to the t-test and the Z-test which are only useful for testing a single effect). The LOD score test statistic is simply a rescaled likelihood ratio test statistic, and thus they are used interchangeably without any difference at all.

8.5 Bonferroni correction

Genome scanning involves multiple tests. Sometimes the number of tests may reach hundreds or even thousands. For a single test, the critical value for any test statistic simply takes the 95% or 99% quantile of the distribution that the test statistic follows under the null hypothesis. For example, the F-test statistic follows an F distribution, the likelihood ratio test statistic follows a chi-square distribution and the W-test statistic also follows a chi-square distribution. When multiple tests are involved, the critical value used for a single test must be adjusted to make the experiment-wise Type I error at a desired level, say 0.05.

The Bonferroni correction is a multiple test correction used when multiple statistical tests are being performed in a single experiment (Dunn, 1961). While a given alpha value α may be appropriate for each individual test, it is not for the set of all tests involved in a single experiment. In order to avoid spurious positives, the alpha value needs to be lowered to account for the number of tests being performed. The Bonferroni correction sets the Type I error for the entire set of k tests equal to β by taking the alpha value for each test equal to α. The β is now called the experiment-wise Type I error rate and α is called the test-wise Type I error rate or nominal Type I error rate. The Bonferroni correction states that, in an experiment involving k tests, if you want to control the experiment-wise Type I error rate at β, the nominal Type I error rate for a single test should be

$$\alpha = \frac{\beta}{k} \tag{8.15}$$

For example, if an experiment involves 100 tests and the investigator wants to control the experiment-wise Type I error at $\beta = 0.05$, for each of the individual tests, the nominal Type I error rate should be $\alpha = \frac{\beta}{k} = \frac{0.05}{100} = 0.0005$. In other words, for any individual test the p-value should be less than 0.0005 in order to declare significance for that test. The Bonferroni correction does not require independence of the multiple tests.

When the multiple tests are independent, there is an alternative correction for the Type I error, which is called the Šidák correction (Abdi, 2007). This correction is often confused with the Bonferroni correction. If a test-wise Type I error is α, the probability of non-significance is $1 - \alpha$ for this particular test. For k independent tests and none of them is significant, the probability is $(1 - \alpha)^k$. The experiment-wise Type I error is defined as the probability that at least one of the k tests is significant. This probability is

$$\beta = 1 - (1 - \alpha)^k \tag{8.16}$$

To find the nominal α value given the experiment wise value β, we use the reverse function

$$\alpha = 1 - (1 - \beta)^{1/k} \tag{8.17}$$

This correction is the Šidák correction. The two corrections are approximately the same when β is small because $(1 - \beta)^{1/k} \approx 1 - \frac{\beta}{k}$ and thus

$$\alpha \approx \frac{\beta}{k} \tag{8.18}$$

Therefore, the Bonferroni correction is an approximation of the Šidák correction for multiple independent tests for small β.

8.6 Permutation test

When the number of tests is large, the Bonferroni and Šidák corrections tend to be over conservative. In addition, if a test statistic does not follow any standard distribution under the null model, calculation of the p-value may be difficult for each individual test. In this case, we can adopt the permutation test to draw an empirical critical value. This method was developed by Churchill and Doerge (1994) for QTL mapping. The idea is simple, but implementation can be time consuming. When the sample size n is small, we can evaluate all $n!$ different permuted samples of the original phenotypic values while keeping the marker genotype data intact. In other words, we

Table 8.1. Phenotypic values of trait y and the genotypes of five markers from ten plants (the original data set)

plant	y	M_1	M_2	M_3	M_4	M_5
1	55.0	H	H	H	H	A
2	54.2	H	H	H	H	U
3	61.6	H	H	U	A	A
4	66.6	H	H	H	H	U
5	67.4	H	H	H	B	U
6	64.3	H	H	H	H	H
7	54.0	H	A	B	B	B
8	57.2	H	B	H	H	H
9	63.7	H	H	H	H	H
10	55.0	H	H	A	H	U

only reshuffle the phenotypes, not the marker genotypes. For each permuted sample, we apply any method of genome scanning to calculate the test statistical values for all markers. In each of the permuted sample, the association of the phenotype and genotypes of markers has been (purposely) destroyed so that the distribution of the test statistics will mimic the actual distribution under the null model, from which a desirable critical value can be drawn from the empirical null distribution. The number of permuted samples can be extremely large if the sample size is large. In this case, we can randomly reshuffle the data to purposely destroy the association between the phenotype and the marker genotype. By random reshuffling the phenotypes, individual j may take the phenotypic value of individual i for $i \neq j$ while the marker genotype of individual j remains unchanged. After reshuffling the phenotypes, we analyze the data and scan the entire genome. By chance, we may find some peaks in the test statistic profile. We know that these peaks are false because we have already destroyed the association between markers and phenotypes. We record the value of the test statistic at the highest peak of the profile and denote it by λ_1. We then reshuffle the data and scan the genome again. We may find some false peaks again. We then record the highest peak and write down the value, λ_2, and put it in the data set. We repeat the reshuffling process many times to form a large sample of λ's, denoted by $\{\lambda_1, \ldots, \lambda_M\}$, where M is a large number, say 1000. These λ values will form a distribution, called the null distribution. The 95% or 99% quantile of the null distribution is the empirical critical value for our test statistic. We then compare our test statistic for each marker (from the original data analysis) against this empirical critical value. If the test statistic of a marker is larger than this critical value, we can declare this marker as being significant. Note that permutation test is time consuming, but it is realistic with the advanced computing system currently available in most laboratories.

We now provide an example to show how to use the permutation test to draw the critical value of a test statistic. Table 8.1 gives a small sample of ten plants (the original dataset).

Ten randomly reshuffled samples are demonstrated in Table 8.2. We can see that the first observation of sample 1 (S_1) takes the phenotype of plant number 6 while the genotypes of the five markers remain unchanged. Another example is that the second observation of sample 2 (S_2) takes the phenotype of plant number 8 while the genotypes of the five markers are still the genotypes for plant number 2. The phenotypic values corresponding to the ten reshuffled samples are given in Table 8.3. Each sample is subject to genome scanning, i.e., five F-test statistics are calculated, one for each marker. The maximum F-test statistic value for each reshuffled sample is given in the last row of Table 8.3. For example, the maximum F-value in the first sample (S_1) is 2.12, the maximum F-value for S_2 is 1.28 and so on. We then sorted the ten F-values from the ten samples in descending order as shown in the following sequence,

Table 8.2. Plant ID's of ten randomly reshuffled samples, denoted by $S_1, S_2, \ldots, S_{10}$, respectively

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
6	5	9	10	5	6	2	10	1	10
2	8	6	8	1	2	7	6	4	5
4	1	5	7	8	4	1	9	3	9
3	10	1	4	10	3	10	8	5	3
8	3	2	1	7	8	4	5	8	1
9	4	3	5	2	9	3	3	7	8
10	7	10	2	4	10	8	2	9	2
1	9	7	6	3	1	9	1	10	6
7	6	8	3	6	7	6	7	6	7
5	2	4	9	9	5	5	4	2	4

$$\{4.51, 3.95, 3.51, 3.21, 2.33, 2.12, 1.85, 1.28, 1.11, 0.95\}$$

The ten F-values are assumed to be sampled from the null distribution. The empirical 90% quantile is 3.95, which can be used as the critical value for the F-test statistic to compare under the Type I error of 0.10. The number of reshuffled samples in the example is not sufficiently large to give 95% quantile for the Type I error of $\alpha = 0.05$. In practice, the number of randomly reshuffled samples depends on $\alpha = 0.05$ due to Monte Carlo error. Nettleton and Doerge (2000) recommended that the permutation sample size should be at least $\frac{5}{\alpha}$, where α is the experiment-wise Type I error rate. In permutation analysis, there is no such a thing as nominal Type I error. In practice, we often choose 1000 as the permutation sample size.

Permutation test is not a method for genome scanning; rather, it is only a way to draw an empirical critical value of a test statistic for us to decide statistical significance of a marker. It applies to all test statistics, e.g., the F-test, the W-test, the likelihood ratio test and so on. The phrase "permutation test" can be confusing because it is not a method for significance test. "Permutation analysis" may be a better phrase to describe this empirical approach of critical value calculation.

8.7 Piepho's approximate critical value

Permutation analysis is perhaps the best method for drawing the empirical critical value to control the genome wise Type I error rate. However, it can be time consuming.Piepho (2001) developed an approximate method, which does not require randomly reshuffling of the data. The method simply uses exiting test statistical values of all points across the genome. The test statistic must be the likelihood ratio test statistic. If the test statistic is the LOD score, a simple conversion to the likelihood ratio test statistic is required. The W-test statistic may also be used because it also follows a chi-square distribution under the null model. Let $\beta = 0.05$ be the genome wise Type I error, $C = \chi^2_{k,1-\alpha}$ be the $(1-\alpha) \times 100\%$ quantile of the chi-square distribution and k (the degrees of freedom of the test statistic) is the number of genetic effects subject to statistical test, where $k = 1$ for a BC design and $k = 2$ for an F_2 design. The following relations provides a way to solve for $C = \chi^2_{k,1-\alpha}$, the critical value for the likelihood ratio test statistic to compare so that the genome wise Type I error is controlled at β.

$$\beta = m \Pr(\chi^2_k > C) + \frac{2^{-\frac{1}{2}k} C^{-\frac{1}{2}(1-k)} e^{-\frac{1}{2}C}}{\Gamma(\frac{k}{2})} \sum_{i=1}^{m} v_i \tag{8.19}$$

where m is the number of chromosomes and $\Gamma(\frac{k}{2})$ is the Gamma function. The v_i for the ith chromosome is defined as

Table 8.3. The corresponding phenotypic values and the maximum F-test statistical values (last row) in the ten randomly reshuffled samples

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
	64.3	67.4	63.7	55.0	67.4	64.3	54.2	55.0	55.0	55.0
	54.2	57.2	64.3	57.2	55.0	54.2	54.0	64.3	66.6	67.4
	66.6	55.0	67.4	54.0	57.2	66.6	55.0	63.7	61.6	63.7
	61.6	55.0	55.0	66.6	55.0	61.6	55.0	57.2	67.4	61.6
	57.2	61.6	54.2	55.0	54.0	57.2	66.6	67.4	57.2	55.0
	63.7	66.6	61.6	67.4	54.2	63.7	61.6	61.6	54.0	57.2
	55.0	54.0	55.0	54.2	66.6	55.0	57.2	54.2	63.7	54.2
	55.0	63.7	54.0	64.3	61.6	55.0	63.7	55.0	55.0	64.3
	54.0	64.3	57.2	61.6	64.3	54.0	64.3	54.0	64.3	54.0
	67.4	54.2	66.6	63.7	63.7	67.4	67.4	66.6	54.2	66.6
F-test	2.12	1.28	3.51	1.85	2.33	4.51	3.21	3.95	1.11	0.95

$$v_i = \left|\sqrt{\lambda_1} - \sqrt{\lambda_2}\right| + \left|\sqrt{\lambda_2} - \sqrt{\lambda_3}\right| + ... + \left|\sqrt{\lambda_{m_i-1}} - \sqrt{\lambda_{m_i}}\right| \tag{8.20}$$

where λ_l for $l = 1, ..., m_i$ is the likelihood ratio test statistic for marker l in chromosome i, and m_i is the total number of markers in chromosome i. Once β is given, the above equation is simply a function of C. A numerical solution can be found using the bi-section algorithm. Once C is found, the Type I error for an individual marker can be obtained using the inverse function of the chi-square distribution function. The Gamma function $\Gamma(\frac{k}{2})$ depends on k, the number of genetic effects. For the common designs of experiments, k only takes 1, 2 or 3. For example, $k = 1$ for BC, DH (double haploid) and RIL (recombinant inbred line) designs. If the additive effect is the only one to be tested in the F_2 design, k also equals 1. If both the additive and dominance effects are tested in the F_2 design, $k = 2$. In a four-way cross design, k equals 3. Therefore, we only need the value of $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, $\Gamma(\frac{2}{2}) = \Gamma(1) = 1$ and $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$.

8.8 Theoretical consideration

Genome scanning described in this chapter refers to marker analysis. Since a marker is not a QTL, the estimated marker effect only reflects a fraction of the QTL effect. Take a BC design as an example. Assume that a QTL with effect a is d cM away from a marker. If we use this marker to estimate the QTL effect, the marker effect will not be equal to a; rather, it will be $(1 - 2r)a$, where $r = \frac{1}{2}(1 - e^{-d/(2\times100)})$ is the recombination fraction between the marker and the QTL. The marker effect is only a fraction of the QTL effect. This fraction is $(1 - 2r)$, which is the correlation coefficient between the marker and the QTL genotype indicator variables. When $r = 0$, a situation where the marker overlaps with the QTL, the marker effect is identical to the QTL effect. On the other hand, if $r = 0.5$, a situation where the marker is not linked to the QTL, the marker effect equals zero, regardless how large the QTL effect is. When $0 < r < 0.5$, what we estimate for the marker is a confounded effect between the QTL effect and the linkage parameter. A small marker effect may be due to a large QTL effect but weakly linked to the marker or a small QTL effect with a strong linkage. There is no way to tell the actual QTL effect unless more markers are taken into account simultaneously, which is the topic to be addressed in the next chapter when interval mapping is introduced.

We now prove that the correlation between the marker and the QTL is $1 - 2r$. Let X be the indicator variable for the QTL genotype, i.e., $X = 1$ for A_1A_1 and $X = 0$ for A_1A_2. Let M be the corresponding indicator variable for the marker genotype, i.e., $M = 1$ and $M = 0$, respectively, for the two genotypes of the marker. The joint distribution of X and M is given in Table 8.4. This joint distribution table is symmetrical, meaning that both M and X have the same marginal distribution. First, let us look at the marginal distribution of variable

		M		
		1	0	
X	1	$(1-r)/2$	$r/2$	1/2
	0	$r/2$	$(1-r)/2$	1/2
		1/2	1/2	

Table 8.4. Joint distribution of X (QTL genotype) and M (marker genotype)

M. From the joint distribution table, we get $\Pr(M=1) = \frac{1-r}{2} + \frac{r}{2} = \frac{1}{2}$ and $\Pr(M=0) = \frac{r}{2} + \frac{1-r}{2} = \frac{1}{2}$. Therefore, the variance of M is

$$\begin{aligned}\mathrm{var}(M) =& E(M^2) - E^2(M) \\ =& \left(\frac{1}{2}\times 1^2 + \frac{1}{2}\times 0^2\right) - \left(\frac{1}{2}\times 1 + \frac{1}{2}\times 0\right)^2 \\ =& \frac{1}{4}\end{aligned} \tag{8.21}$$

Similarly, $\mathrm{var}(X) = \frac{1}{4}$, due to the symmetrical nature. We now evaluate the covariance between M and X.

$$\begin{aligned}\mathrm{cov}(M,X) =& E(MX) - E(M)E(X) \\ =& \frac{1-r}{2}\times 1\times 1 - \frac{1}{2}\times\frac{1}{2} = \frac{1}{4}\left[2(1-r)-1\right] \\ =& \frac{1}{4}(1-2r)\end{aligned} \tag{8.22}$$

The correlation coefficient between the two variables is

$$\rho_{MX} = \frac{\mathrm{cov}(M,X)}{\sqrt{\mathrm{var}(M)\mathrm{var}(X)}} = \frac{\frac{1}{4}(1-2r)}{\frac{1}{4}} = 1-2r \tag{8.23}$$

Because of the symmetry of X and M, this correlation is also equal to the regression coefficient, i.e.,

$$\beta_{XM} = \frac{\mathrm{cov}(M,X)}{\mathrm{var}(M)} = \frac{\frac{1}{4}(1-2r)}{\frac{1}{4}} = 1-2r \tag{8.24}$$

Recall that when a marker is used to estimate the effect of a linked QTL, the QTL effect will be biased by a factor $(1-2r)$. This fraction is the correlation between X and M. In fact, it is the regression coefficient of X on M. Because $\rho_{MX} = \beta_{XM}$, we say $(1-2r)$ is the correlation coefficient. We now show why the factor of reduction is β_{XM}. Recall that the QTL model is

$$y_j = b_0 + E(X_j|M_j)b_1 + e_j \tag{8.25}$$

where $E(X_j|M)$ is the conditional mean of X_j given M_j. We use marker genotype M_j to infer the QTL genotype X_j. The conditional mean can be expressed as the predicted value of X from M using the following regression equation,

$$E(X_j|M_j) = E(X_j) + M_j\beta_{XM} = \frac{1}{2} + M_j(1-2r) \tag{8.26}$$

Substituting this equation (8.26) into the above model (equation 8.25), we get

$$\begin{aligned} y_j &= \left(b_0 + \tfrac{1}{2}\right) + M_j(\beta_{XM}b_1) + e_j \\ &= b_0^* + M_j b_1^* + e_j \end{aligned} \tag{8.27}$$

where $b_0^* = b_0 + \frac{1}{2}$ and $b_1^* = \beta_{XM}b_1 = (1-2r)a$. Note that $b_1 = a$ is the genetic effect and $\beta_{XM} = 1-2r$ as given in equation (8.24).

9

Interval Mapping

Interval mapping is an extension of the individual marker analysis so that two markers are analyzed at a time. In the marker analysis (Chapter 8), we cannot estimate the exact position of a QTL. With interval mapping, we use two markers to determine an interval, within which a putative QTL position is proposed. The genotype of the putative QTL is not observable but can be inferred with a certain probability using the three-point or multipoint method introduced in Chapter 4. Once the genotype of the QTL is inferred, we can estimate and test the QTL effect at that particular position. We divide the interval into many putative positions of QTL with one or two cM apart and investigate every putative position within the interval. Once we have searched the current interval, we move on to the next interval and so on until all intervals have been searched. The putative QTL position (not necessarily at a marker) that has the maximum test statistical value is the estimated QTL position. Figure 9.1 demonstrates the process of genome scanning for markers only (panel a), for markers and virtual markers (panel b) and for every point of the chromosome (panel c).

Interval mapping was originally developed by Lander and Botstein (1989) and further modified by numerous authors. Interval mapping has revolutionized genetic mapping because we can really pinpoint the exact location of a QTL. In each of the four sections that follow, we will introduce one specific statistical method of interval mapping based on the F_2 design. Methods of interval mapping for a BC design is straightforward and thus will not be discussed in this chapter. Maximum likelihood (ML) method of interval mapping (Lander and Botstein, 1989) is the optimal method for interval mapping. Least squares (LS) method (Haley and Knott, 1992) is a simplified approximation of Lander and Botstein method. The iteratively reweighted least squares (IRLS) method (Xu, 1998a,b) is a further improved method over the least squares method. Recently Feenstra et al. (2006) developed an estimating equation (EE) method for QTL mapping, which is an extension of the IRLS with improved performance. Han and Xu (2008) developed a Fisher scoring algorithm (FISHER) for QTL mapping. Both the EE and FISHER algorithms maximize the same likelihood function and thus they generate identical result. In this chapter, we introduce the methods based on their simplicity rather than their chronological orders of development. Therefore, the methods will be introduced in the following order: LS, IRLS, FISHER and ML. Bayesian method will be discussed in a later chapter where multiple QTL mapping is addressed.

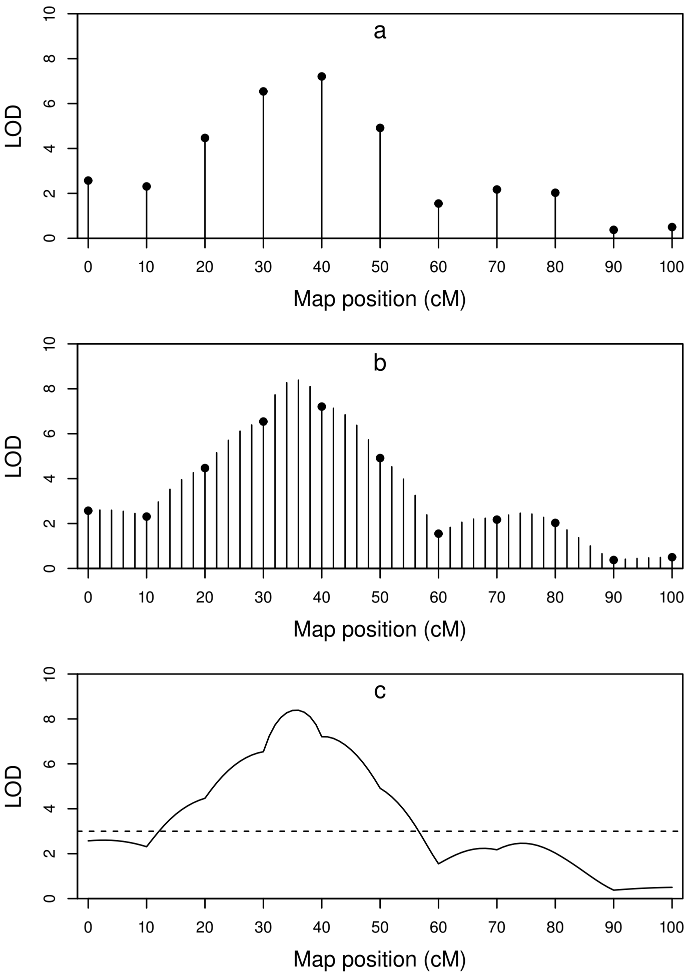

Fig. 9.1. The LOD test statistics for (a) marker effects (top panel), (b) virtual marker effects (panel in the middle), and (c) every point of a simulated chromosome (bottom panel).

9.1 Least squares method

The LS method was introduced by Haley and Knott (1992) aiming to improving the computational speed. The statistical model for the phenotypic value of the jth individual is

$$y_j = X_j\beta + Z_j\gamma + \varepsilon_j \tag{9.1}$$

where β is a $p \times 1$ vector for some model effects that are irrelevant to QTL effects, X_j is a $1 \times p$ known design vector, $\gamma = \{a, d\}$ is a 2×1 vector for QTL effects of a putative locus (a for additive effect and d for dominance effect), Z_j is a 1×2 vector for the genotype indicator variable defined as

$$Z_j = \begin{cases} H_1 & \text{for } A_1A_1 \\ H_2 & \text{for } A_1A_2 \\ H_3 & \text{for } A_2A_2 \end{cases} \tag{9.2}$$

where H_k for $k = 1, 2, 3$ is the kth row of matrix

$$H = \begin{bmatrix} +1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{9.3}$$

The residual error ε_j is assumed to be a $N(0, \sigma^2)$ variable. Although normal distribution for ε_j is not a required assumption for the LS method, it is required for the ML method. It is important to include non-QTL effects β in the model to control the residual error variance as small as possible. For example, location and year effects are common in replicated experiments. These effects are not related to QTL but will contribute to the residual error if not included in the model. If there is no such a non-QTL effect to consider in a nice designed experiment, β will be a single parameter (intercept) and X_j will be unity across all $j = 1, \ldots, n$.

With interval mapping, the QTL genotype is never known unless the putative QTL position overlaps with a fully informative marker. Therefore,Haley and Knott (1992) suggested to replace the unknown Z_j by the expectation of Z_j conditional on flanking marker genotype. Let $p_j(1)$, $p_j(0)$ and $p_j(-1)$ be the conditional probabilities for the three genotypes given flanking marker information (see Chapter 4 for the method of calculating conditional probability). The LS model of Haley and Knott (1992) is

$$y_j = X_j\beta + U_j\gamma + e_j \tag{9.4}$$

where

$$U_j = E(Z_j) = p_j(+1)H_1 + p_j(0)H_2 + p_j(-1)H_3 \tag{9.5}$$

is the conditional expectation of Z_j. The residual error e_j (different from ε_j) remains normal with mean zero and variance σ^2, although this assumption has been violated (see next section). The least squares estimate of β and γ is

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} X_j^T X_j & \sum_{j=1}^{n} X_j^T U_j \\ \sum_{j=1}^{n} U_j^T X_j & \sum_{j=1}^{n} U_j^T U_j \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n} X_j^T y_j \\ \sum_{j=1}^{n} U_j^T y_j \end{bmatrix} \tag{9.6}$$

and the estimated residual error variance is

$$\hat{\sigma}^2 = \frac{1}{n-p-2} \sum_{j=1}^{n} (y_j - X_j\hat{\beta} - U_j\hat{\gamma})^2 \tag{9.7}$$

The variance-covariance matrix of the estimated parameters is

$$\text{var} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} X_j^T X_j & \sum_{j=1}^{n} X_j^T U_j \\ \sum_{j=1}^{n} U_j^T X_j & \sum_{j=1}^{n} U_j^T U_j \end{bmatrix}^{-1} \hat{\sigma}^2 \tag{9.8}$$

which is a $(p+2) \times (p+2)$ matrix. Let

$$\text{var}(\hat{\gamma}) = V = \begin{bmatrix} \text{var}(\hat{a}) & \text{cov}(\hat{a}, \hat{d}) \\ \text{cov}(\hat{a}, \hat{d}) & \text{var}(\hat{d}) \end{bmatrix} \tag{9.9}$$

be the 2×2 lower diagonal bock of matrix (9.8). The standard errors of the estimated additive and dominance effects are the square roots of the diagonal elements of matrix (9.9).

We can use either the F-test or the W-test statistic to test the hypothesis of $H_0 : \gamma = 0$. The W-test statistic is

$$W = \hat{\gamma}^T V^{-1} \hat{\gamma} = \begin{bmatrix} \hat{a} & \hat{d} \end{bmatrix} \begin{bmatrix} \text{var}(\hat{a}) & \text{cov}(\hat{a}, \hat{d}) \\ \text{cov}(\hat{a}, \hat{d}) & \text{var}(\hat{d}) \end{bmatrix}^{-1} \begin{bmatrix} \hat{a} \\ \hat{d} \end{bmatrix} \tag{9.10}$$

The likelihood ratio test statistic can also be applied if we assume that $e_j \sim N(0, \sigma^2)$ for all $j = 1, \ldots, n$. The log likelihood function for the full model is

$$\begin{aligned} L_1 &= -\frac{n}{2} \ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \sum_{j=1}^{n} (y - X_j\hat{\beta} - U_j\hat{\gamma})^2 \\ &\approx -\frac{n}{2} \left[\ln(\hat{\sigma}^2) + 1 \right] \end{aligned} \tag{9.11}$$

The reduced model under $H_0 : \gamma = 0$ is

$$\begin{aligned} L_0 &= -\frac{n}{2} \ln(\hat{\hat{\sigma}}^2) - \frac{1}{2\hat{\hat{\sigma}}^2} \sum_{j=1}^{n} (y - X_j\hat{\hat{\beta}})^2 \\ &\approx -\frac{n}{2} \left[\ln(\hat{\hat{\sigma}}^2) + 1 \right] \end{aligned} \tag{9.12}$$

where

$$\hat{\hat{\beta}} = \left[\sum_{j=1}^{n} X_j^T X_j\right]^{-1} \left[\sum_{j=1}^{n} X_j^T y_j\right] \tag{9.13}$$

and

$$\hat{\hat{\sigma}}^2 = \frac{1}{n-p}\sum_{j=1}^{n}(y_j - X_j\hat{\hat{\beta}})^2 \tag{9.14}$$

The likelihood ratio test statistic is

$$\lambda = -2(L_0 - L_1) \tag{9.15}$$

9.2 Weighted least squares

Xu (1995) realized that the LS method is flawed because the residual variance is heterogeneous after replacing X_j by its conditional expectation U_j. The conditional variance of X_j given marker information varies from one individual to another and it will contribute to the residual variance. Xu (1998a,b) modified the exact model

$$y_j = X_j\beta + Z_j\gamma + \varepsilon_j \tag{9.16}$$

by

$$y_j = X_j\beta + U_j\gamma + (Z_j - U_j)\gamma + \varepsilon_j \tag{9.17}$$

which differs from the Haley and Knott's (1992) model by $(Z_j - U_j)\gamma$. Since Z_j is not observable, this additional term is merged into the residual error if ignored. Let

$$e_j = (Z_j - U_j)\gamma + \varepsilon_j \tag{9.18}$$

be the new residual error. The Haley and Knott's (1992) model can be rewritten as

$$y_j = X_j\beta + U_j\gamma + e_j \tag{9.19}$$

Although we assume $\varepsilon_j \sim N(0, \sigma^2)$, this does not validate the normal assumption of e_j. The expectation for e_j is

$$E(e_j) = [E(Z_j) - U_j]\,\gamma + E(\varepsilon_j) = 0 \tag{9.20}$$

The variance of e_j is

$$\text{var}(e_j) = \sigma_j^2 = \gamma^T\text{var}(Z_j)\gamma + \sigma^2 = \left(\frac{1}{\sigma^2}\gamma^T\Sigma_j\gamma + 1\right)\sigma^2 \tag{9.21}$$

where $\Sigma_j = \text{var}(Z_j)$, which is defined as a conditional variance-covariance matrix given flanking marker information. The explicit forms of Σ_j is

$$\Sigma_j = E(Z_j^T Z_j) - E(Z_j^T)E(Z_j), \tag{9.22}$$

where

$$E(Z_j^T Z_j) = p_j(1)H_1^T H_1 + p_j(0)H_2^T H_2 + p_j(-1)H_3^T H_3 \tag{9.23}$$

and

$$E(Z_j) = U_j = p_j(1)H_1 + p_j(0)H_2 + p_j(-1)H_3. \tag{9.24}$$

Let

$$\sigma_j^2 = \left(\frac{1}{\sigma^2}\gamma^T \Sigma_j \gamma + 1\right)\sigma^2 = \frac{1}{W_j}\sigma^2 \tag{9.25}$$

where

$$W_j = \left(\frac{1}{\sigma^2}\gamma^T \Sigma_j \gamma + 1\right)^{-1} \tag{9.26}$$

is the weight variable for the jth individual. The weighted least squares estimate of the parameters are

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} X_j^T W_j X_j & \sum_{j=1}^{n} X_j^T W_j U_j \\ \sum_{j=1}^{n} U_j^T W_j X_j & \sum_{j=1}^{n} U_j^T W_j U_j \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n} X_j^T W_j y_j \\ \sum_{j=1}^{n} U_j^T W_j y_j \end{bmatrix} \tag{9.27}$$

and

$$\hat{\sigma}^2 = \frac{1}{n-p-2}\sum_{j=1}^{n} W_j (y_j - X_j\hat{\beta} - U_j\hat{\gamma})^2 \tag{9.28}$$

Since W_j is a function of σ^2, iterations are required. The iteration process is demonstrated as below.

1. Initialize γ and σ^2
2. Update β and γ using equation 9.27
3. Update σ^2 using equation 9.28
4. Repeat Step 2 to Step 3 until a certain criterion of convergence is satisfied.

The iteration process is very fast, usually taking less than 5 iterations to converge. Since the weight is not a constant (it is a function of the parameters), repeatedly updating the weight is required. Therefore, the weighted least squares method is also called iteratively reweighted least squares (IRLS). The few cycle of iterations make the results of IRLS very close to that of the maximum likelihood method (to be introduce later). A nice property of the IRLS is that the variance-covariance matrix of the estimated parameters is automatically given as a by-product of the iteration process. This matrix is

$$\text{var}\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} X_j^T W_j X_j & \sum_{j=1}^{n} X_j^T W_j U_j \\ \sum_{j=1}^{n} U_j^T W_j X_j & \sum_{j=1}^{n} U_j^T W_j U_j \end{bmatrix}^{-1} \hat{\sigma}^2 \tag{9.29}$$

As a result, the F- or W-test statistic can be used for significance test. Like the least squares method, a likelihood ratio test statistic can also be established for significance test. The L_0 under the null model is the same as that described in the section of least squares method. The L_1 under the alternative model is

$$\begin{aligned} L_1 &= -\frac{n}{2}\ln(\hat{\sigma}^2) + \frac{1}{2}\sum_{j=1}^{n}\ln(W_j) - \frac{1}{2\hat{\sigma}^2}\sum_{j=1}^{n} W_j(y - X_j\hat{\beta} - U_j\hat{\gamma})^2 \\ &\approx -\frac{n}{2}\left[\ln(\hat{\sigma}^2) + 1\right] + \frac{1}{2}\sum_{j=1}^{n}\ln(W_j) \end{aligned} \tag{9.30}$$

9.3 Fisher scoring

The weighted least squares solution described in the previous section does not maximize the log likelihood function (9.30). We can prove that it actually maximizes equation (9.30) if W_j is treated as a constant. The fact that W_j is a function of parameters makes the above weighted least squares estimates suboptimal. The optimal solution should be obtained by maximizing equation (9.30) fully without assuming W_j being a constant.

Recall that the linear model for y_j is

$$y_j = X_j\beta + U_j\gamma + e_j \tag{9.31}$$

where the residual error $e_j = (Z_j - U_j)\gamma + \varepsilon_j$ has a zero mean and variance

$$\sigma_j^2 = \left(\frac{1}{\sigma^2}\gamma^T \Sigma_j \gamma + 1\right)\sigma^2 = \frac{1}{W_j}\sigma^2 \tag{9.32}$$

If we assume that $e_j \sim N(0, \sigma_j^2)$, we can construct the following log likelihood function,

$$L(\theta) = -\frac{n}{2}\ln(\sigma^2) + \frac{1}{2}\sum_{j=1}^{n}\ln(W_j) - \frac{1}{2\sigma^2}\sum_{j=1}^{n} W_j(y - X_j\beta - U_j\gamma)^2 \tag{9.33}$$

where $\theta = \{\beta, \gamma, \sigma^2\}$ is the vector of parameters. The maximum likelihood solution for the above likelihood function is hard to obtain because W_j is not a constant but a function of the parameters. The Newton-Raphson algorithm may be adopted but it requires the second partial derivative of the log likelihood function with respect to the parameter, which is very complicated. In addition, the Newton-Raphson algorithm often misbehaves when the dimensionality of θ is high. We now introduce the Fisher scoring algorithm for finding the MLE of θ. The method requires the first partial derivative of $L(\theta)$ with respect to the parameters, called the score vector and denoted by $S(\theta)$, and the information matrix, denoted by $I(\theta)$. The score vector has the following form,

$$S(\theta) = \begin{bmatrix} \frac{1}{\sigma^2}\sum\limits_{j=1}^{n} X_j^T W_j (y_j - \mu_j) \\ \frac{1}{\sigma^2}\sum\limits_{j=1}^{n} U_j^T W_j (y_j - \mu_j) - \frac{1}{\sigma^2}\sum\limits_{j=1}^{n} W_j \Sigma_j \gamma + \frac{1}{\sigma^4}\sum\limits_{j=1}^{n} (y_j - \mu_j)^2 W_j^2 \Sigma_j \gamma \\ \frac{1}{2\sigma^4}\sum\limits_{j=1}^{n} W_j^2 (y_j - \mu_j)^2 - \frac{1}{2\sigma^2}\sum\limits_{j=1}^{n} W_j \end{bmatrix} \tag{9.34}$$

where

$$\mu_j = X_j\beta + U_j\gamma \tag{9.35}$$

The information matrix is given below

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2}\sum\limits_{j=1}^{n} X_j^T W_j X_j & \frac{1}{\sigma^2}\sum\limits_{j=1}^{n} X_j^T W_j U_j & 0 \\ \frac{1}{\sigma^2}\sum\limits_{j=1}^{n} U_j W_j X_j, & \frac{1}{\sigma^2}\sum\limits_{j=1}^{n} U_j^T W_j U_j + \frac{2}{\sigma^4}\sum\limits_{j=1}^{n} W_j^2 \Sigma_j \gamma\gamma^T \Sigma_j, & \frac{1}{\sigma^4}\sum\limits_{j=1}^{n} W_j^2 \Sigma_j \gamma \\ 0 & \frac{1}{\sigma^4}\sum\limits_{j=1}^{n} W_j^2 \gamma^T \Sigma_j & \frac{1}{2\sigma^4}\sum\limits_{j=1}^{n} W_j^2 \end{bmatrix} \tag{9.36}$$

The Fisher scoring algorithm is implemented using the following iteration equation,

$$\theta^{(t+1)} = \theta^{(t)} + I^{-1}(\theta^{(t)})S(\theta^{(t)}) \tag{9.37}$$

where $\theta^{(t)}$ is the parameter value at iteration t and $\theta^{(t+1)}$ is the updated value. Once the iteration process converges, the variance-covariance matrix of the estimated parameters is automatically given, which is

$$\mathrm{var}(\hat{\theta}) = I^{-1}(\hat{\theta}) \tag{9.38}$$

The detailed expression of this matrix is

$$\mathrm{var}\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \sum\limits_{j=1}^{n} X_j^T W_j X_j & \sum\limits_{j=1}^{n} X_j^T W_j U_j & 0 \\ \sum\limits_{j=1}^{n} U_j W_j X_j, & \sum\limits_{j=1}^{n} U_j^T W_j U_j + \frac{2}{\hat{\sigma}^2}\sum\limits_{j=1}^{n} W_j^2 \Sigma_j \hat{\gamma}\hat{\gamma}^T \Sigma_j, & \frac{1}{\hat{\sigma}^2}\sum\limits_{j=1}^{n} W_j^2 \Sigma_j \hat{\gamma} \\ 0 & \frac{1}{\hat{\sigma}^2}\sum\limits_{j=1}^{n} W_j^2 \hat{\gamma}^T \Sigma_j & \frac{1}{2\hat{\sigma}^2}\sum\limits_{j=1}^{n} W_j^2 \end{bmatrix}^{-1} \hat{\sigma}^2 \tag{9.39}$$

which can be compared with the variance-covariance matrix of the iteratively reweighted least squares estimate given in the previous section (equation 9.29).

We now give the derivation of the score vector and the information matrix. We can write the log likelihood function as

$$L(\theta) = \sum_{j=1}^{n} L_j(\theta) \tag{9.40}$$

where

$$L_j(\theta) = -\frac{1}{2}\ln(\sigma^2) + \frac{1}{2}\ln W_j - \frac{1}{2\sigma^2}W_j(y_j - \mu_j)^2 \tag{9.41}$$

and

$$\mu_j = X_j\beta + U_j\gamma \tag{9.42}$$

The score vector is a vector of the first partial derivatives, as shown below

$$S(\theta) = \sum_{j=1}^{n} S_j(\theta) \tag{9.43}$$

where

$$S_j(\theta) = \begin{bmatrix} \frac{\partial}{\partial\beta}L_j(\theta) \\ \frac{\partial}{\partial\gamma}L_j(\theta) \\ \frac{\partial}{\partial\sigma^2}L_j(\theta) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}X_j^T W_j(y_j - \mu_j) \\ \frac{1}{\sigma^2}U_j^T W_j(y_j - \mu_j) - \frac{1}{\sigma^2}W_j\Sigma_j\gamma + \frac{1}{\sigma^4}(y_j - \mu_j)^2 W_j^2\Sigma_j\gamma \\ \frac{1}{2\sigma^4}W_j^2(y_j - \mu_j)^2 - \frac{1}{2\sigma^2}W_j \end{bmatrix} \tag{9.44}$$

Therefore, we only need to take the sum of the first partial derivatives across individuals to get the score vector. Note that when deriving $S_j(\theta)$ we need the following derivatives,

$$\frac{\partial W_j}{\partial\theta} = \begin{bmatrix} \frac{\partial W_j}{\partial\beta} \\ \frac{\partial W_j}{\partial\gamma} \\ \frac{\partial W_j}{\partial\sigma^2} \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{2}{\sigma^2}W_j^2\Sigma_j\gamma \\ \frac{1}{\sigma^2}W_j(1 - W_j) \end{bmatrix} \tag{9.45}$$

and

$$\frac{\partial\mu_j}{\partial\theta} = \begin{bmatrix} \frac{\partial\mu_j}{\partial\beta} \\ \frac{\partial\mu_j}{\partial\gamma} \\ \frac{\partial\mu_j}{\partial\sigma^2} \end{bmatrix} = \begin{bmatrix} X_j^T \\ U_j^T \\ 0 \end{bmatrix} \tag{9.46}$$

The information matrix is

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta) = \sum_{j=1}^{n} -E[H_j(\theta)] \tag{9.47}$$

where

$$H_j(\theta) = \frac{\partial^2 L_j(\theta)}{\partial\theta\partial\theta^T} = \begin{bmatrix} \frac{\partial^2 L_j(\theta)}{\partial\beta\partial\beta^T} & \frac{\partial^2 L_j(\theta)}{\partial\beta\partial\gamma^T} & \frac{\partial^2 L_j(\theta)}{\partial\beta\partial\sigma^2} \\ \frac{\partial^2 L_j(\theta)}{\partial\gamma\partial\beta^T} & \frac{\partial^2 L_j(\theta)}{\partial\gamma\partial\gamma^T} & \frac{\partial^2 L_j(\theta)}{\partial\gamma\partial\sigma^2} \\ \frac{\partial^2 L_j(\theta)}{\partial\sigma^2\partial\beta^T} & \frac{\partial^2 L_j(\theta)}{\partial\sigma^2\partial\gamma^T} & \frac{\partial^2 L_j(\theta)}{\partial\sigma^2\partial\sigma^2} \end{bmatrix} \tag{9.48}$$

is the second partial derivative of $L_j(\theta)$ with respect to the parameters and called the Hessian matrix. Derivation of this matrix is very tedious, but the negative expectation of the Hessian matrix is identical to the expectation of the product of the score vector (Wedderburn, 1974),

$$-E[H_j(\theta)] = E[S_j(\theta)S_j^T(\theta)] \tag{9.49}$$

Using this identity, we can avoid the Hessian matrix. Therefore, the information matrix is

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta) = \sum_{j=1}^{n} E[S_j(\theta)S_j^T(\theta)] \tag{9.50}$$

where

$$E[S_j(\theta)S_j^T(\theta)] = \begin{bmatrix} E\left(\frac{\partial L_j(\theta)}{\partial \beta}\frac{\partial L_j(\theta)}{\partial \beta^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial \beta}\frac{\partial L_j(\theta)}{\partial \gamma^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial \beta}\frac{\partial L_j(\theta)}{\partial \sigma^2}\right) \\ E\left(\frac{\partial L_j(\theta)}{\partial \gamma}\frac{\partial L_j(\theta)}{\partial \beta^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial \gamma}\frac{\partial L_j(\theta)}{\partial \gamma^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial \gamma}\frac{\partial L_j(\theta)}{\partial \sigma^2}\right) \\ E\left(\frac{\partial L_j(\theta)}{\partial \sigma^2}\frac{\partial L_j(\theta)}{\partial \beta^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial \sigma^2}\frac{\partial L_j(\theta)}{\partial \gamma^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial \sigma^2}\frac{\partial L_j(\theta)}{\partial \sigma^2}\right) \end{bmatrix} \tag{9.51}$$

Note that the expectation is taken with respect to the phenotypic value y_j . In other words, after taking the expectation, variable y_j will disappear from the expressions. There are six different blocks in the above matrix. We will only provide the derivation for one block as an example. The derivations of the remaining five blocks are left to students for practice. The result can be found in Han and Xu (2008). We now show the derivation of the first block of the matrix. The product (before taking the expectation) is

$$\begin{aligned} \frac{\partial L_j(\theta)}{\partial \beta}\frac{\partial L_j(\theta)}{\partial \beta^T} &= \left[\frac{1}{\sigma^2}X_j^T W_j (y_j - \mu_j)\right]\left[\frac{1}{\sigma^2}X_j^T W_j (y_j - \mu_j)\right]^T \\ &= \frac{1}{\sigma^4}X_j^T W_j^2 X_j^T (y_j - \mu_j)^2 \end{aligned} \tag{9.52}$$

The expectation of it is

$$\begin{aligned} E\left(\frac{\partial L_j(\theta)}{\partial \beta}\frac{\partial L_j(\theta)}{\partial \beta^T}\right) &= \frac{1}{\sigma^4}X_j^T W_j^2 X_j^T E\left[(y_j - \mu_j)^2\right] \\ &= \frac{1}{\sigma^2}X_j^T W_j X_j^T \end{aligned} \tag{9.53}$$

The second line of the above equation requires the following identity,

$$E\left[(y_j - \mu_j)^2\right] = \frac{1}{W_j}\sigma^2 \tag{9.54}$$

Taking the sum of equation (9.53) across individuals, we get

$$I_{11}(\theta) = \frac{1}{\sigma^2}\sum_{j=1}^{n} X_j^T W_j X_j \tag{9.55}$$

which is the first block of the information matrix. When deriving the expectations for the remaining five blocks, we need the following expectations,

$$E[(y_j - \mu_j)^k] = \begin{cases} 0 & \text{for odd } k \\ W_j^{-1}\sigma^2 & \text{for } k = 2 \\ 3W_j^{-2}\sigma^4 & \text{for } k = 4 \end{cases} \tag{9.56}$$

The above expectations requires the assumption of $y_j \sim N(\mu_j, \sigma_j^2)$ where $\sigma_j^2 = W_j^{-1}\sigma^2$.

9.4 Maximum likelihood method

The maximum likelihood method (Lander and Botstein, 1989) is the optimal one compared to all other methods described in this chapter. The linear model for the phenotypic value of y_j is

$$y_j = X_j\beta + Z_j\gamma + \varepsilon_j \tag{9.57}$$

where $\varepsilon_j \sim N(0, \sigma^2)$ is assumed. The genotype indicator variable Z_j is a missing value because we cannot observe the genotype of a putative QTL. Rather than replacing Z_j by U_j as done in the least squares and the weighted least squares methods, the maximum likelihood method takes into consideration the mixture distribution of y_j. We have learned the mixture distribution in Chapter 7 when we deal with segregation analysis of quantitative traits. We now extend the mixture model to interval mapping. When the genotype of the putative QTL is observed, the probability density of y_j is

$$\begin{aligned} f_k(y_j) &= \Pr(y_j|Z_j = H_k) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(y_j - X_j\beta + H_k\gamma)^2\right] \end{aligned} \tag{9.58}$$

When flanking marker information is used, the conditional probability that $Z_j = H_k$ is

$$p_j(k) = \Pr(Z_j = H_k), \forall k = 1, 2, 3 \tag{9.59}$$

for the three genotypes, A_1A_1, A_1A_2 and A_2A_2. These probabilities are different from the Mendelian segregation ratio $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ as described in the segregation analysis. They are the conditional probabilities given marker information and thus vary from one individual to another because different individuals may have different marker genotypes. Using the conditional probabilities as weights, we get the mixture distribution

$$f(y_j) = \sum_{k=1}^{3} p_j(2-k)f_k(y_j) \tag{9.60}$$

where

$$p_j(2-k) = \begin{cases} p_j(-1) & \text{for } k = 1 \\ p_j(0) & \text{for } k = 2 \\ p_j(+1) & \text{for } k = 3 \end{cases} \tag{9.61}$$

is a special notation for the conditional probability and should not be interpreted as p_j times $(2-k)$. The log likelihood function is

$$L(\theta) = \sum_{j=1}^{n} L_j(\theta) \tag{9.62}$$

where $L_j(\theta) = \ln f(y_j)$.

9.4.1 EM algorithm

The MLE of θ can be obtained using any numerical algorithms but the EM algorithm is generally more preferable than others because we can take advantage of the mixture distribution. Derivation of the EM algorithm has been given in Chapter 7 when segregation analysis was introduced. Here we simply give the result of the EM algorithm. Assuming that the genotypes of all individuals are observed, the maximum likelihood estimates of parameters would be

$$\begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \sum\limits_{j=1}^{n} X_j^T X_j & \sum\limits_{j=1}^{n} X_j^T Z_j \\ \sum\limits_{j=1}^{n} Z_j^T X_j & \sum\limits_{j=1}^{n} Z_j^T Z_j \end{bmatrix}^{-1} \begin{bmatrix} \sum\limits_{j=1}^{n} X_j^T y_j \\ \sum\limits_{j=1}^{n} Z_j^T y_j \end{bmatrix} \tag{9.63}$$

and

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^{n} (y_j - X_j\beta - Z_j\gamma)^2 \tag{9.64}$$

The EM algorithm takes advantage of the above explicit solutions of the parameters by substituting all entities containing the missing value Z_j by their posterior expectations, i.e.,

$$\begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \sum\limits_{j=1}^{n} X_j^T X_j & \sum\limits_{j=1}^{n} X_j^T E(Z_j) \\ \sum\limits_{j=1}^{n} E(Z_j^T) X_j & \sum\limits_{j=1}^{n} E(Z_j^T Z_j) \end{bmatrix}^{-1} \begin{bmatrix} \sum\limits_{j=1}^{n} X_j^T y_j \\ \sum\limits_{j=1}^{n} E(Z_j^T) y_j \end{bmatrix} \tag{9.65}$$

and

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^{n} E\left[(y_j - X_j\beta - Z_j\gamma)^2\right] \tag{9.66}$$

where the expectations are taken using the posterior probabilities of QTL genotypes, which is defined as

$$p_j^*(2-k) = \frac{p_j(2-k) f_k(y_j)}{\sum_{k'=1}^{3} p_j(2-k') f_{k'}(y_j)}, \forall k = 1, 2, 3 \tag{9.67}$$

The posterior expectations are

$$
\begin{aligned}
E(Z_j) &= \sum_{k=1}^{3} p_j^*(2-k)H_k \\
E(Z_j^T Z_j) &= \sum_{k=1}^{3} p_j^*(2-k)H_k^T H_k \\
E\left[(y_j - X_j\beta - Z_j\gamma)^2\right] &= \sum_{k=1}^{3} p_j^*(2-k)(y_j - X_j\beta - H_k\gamma)^2
\end{aligned}
\tag{9.68}
$$

Since $f_k(y_j)$ is a function of parameters and thus $p_j^*(2-k)$ is also a function of the parameters. However, the parameters are unknown and they are the very quantities we want to find out. Therefore, iterations are required. Here is the iteration process,

1. Initialize $\theta = \theta^{(t)}$ for $t = 0$
2. Calculate the posterior expectations using equations (9.67) and (9.68)
3. Update parameters using equations (9.65) and (9.66)
4. Increment t by 1 and repeat Step 2 to Step 3 until a certain criterion of convergence is satisfied.

Once the iteration converges, the MLE of the parameters is $\hat{\theta} = \theta^{(t)}$, where t is the number of iterations required for convergence.

9.4.2 Variance-covariance matrix of $\hat{\boldsymbol{\theta}}$

Unlike the weighted least squares and the Fisher scoring algorithms where the variance-covariance matrix of the estimated parameters is automatically given as a by-product of the iteration process, the EM algorithm requires an additional step to calculate this matrix. The method was developed by Louis (1982) and it requires the score vectors and the Hessian matrix for the complete-data log likelihood function rather than the actual observed log likelihood function. The complete-data log likelihood function is the log likelihood function as if Z_j were observed, which is

$$L(\theta, Z) = \sum_{j=1}^{n} L_j(\theta, Z) \tag{9.69}$$

where

$$L_j(\theta, Z) = -\frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y_j - X_j\beta - Z_j\gamma)^2 \tag{9.70}$$

The score vector is

$$S(\theta, Z) = \sum_{j=1}^{n} S_j(\theta, Z) \tag{9.71}$$

where

$$S_j(\theta,Z)=\begin{bmatrix}\frac{\partial}{\partial\beta}L_j(\theta,Z)\\ \frac{\partial}{\partial\gamma}L_j(\theta,Z)\\ \frac{\partial}{\partial\sigma^2}L_j(\theta,Z)\end{bmatrix}=\begin{bmatrix}\frac{1}{\sigma^2}X_j^T(y_j-X_j\beta-Z_j\gamma)\\ \frac{1}{\sigma^2}Z_j^T(y_j-X_j\beta-Z_j\gamma)\\ -\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}(y_j-X_j\beta-Z_j\gamma)^2\end{bmatrix}\tag{9.72}$$

The second partial derivative (Hessian matrix) is

$$H(\theta,Z)=\sum_{j=1}^{n}H_j(\theta,Z)\tag{9.73}$$

where

$$H_j(\theta,Z)=\begin{bmatrix}\frac{\partial^2L_j(\theta,Z)}{\partial\beta\partial\beta^T}&\frac{\partial^2L_j(\theta,Z)}{\partial\beta\partial\gamma^T}&\frac{\partial^2L_j(\theta,Z)}{\partial\beta\partial\sigma^2}\\ \frac{\partial^2L_j(\theta,Z)}{\partial\gamma\partial\beta^T}&\frac{\partial^2L_j(\theta,Z)}{\partial\gamma\partial\gamma^T}&\frac{\partial^2L_j(\theta,Z)}{\partial\gamma\partial\sigma^2}\\ \frac{L_j(\theta,Z)}{\partial\sigma^2\partial\beta^T}&\frac{L_j(\theta,Z)}{\partial\sigma^2\partial\gamma^T}&\frac{L_j(\theta,Z)}{\partial\sigma^2\partial\sigma^2}\end{bmatrix}\tag{9.74}$$

The six different blocks of the above matrix are

$$\begin{aligned}\frac{\partial^2L_j(\theta)}{\partial\beta\partial\beta^T}&=-\frac{1}{\sigma^2}X_j^TX_j\\ \frac{\partial^2L_j(\theta)}{\partial\beta\partial\gamma^T}&=-\frac{1}{\sigma^2}X_j^TZ_j\\ \frac{\partial^2L_j(\theta)}{\partial\beta\partial\sigma^2}&=-\frac{1}{\sigma^4}X_j^T(y_j-X_j\beta-Z_j\gamma)\\ \frac{\partial^2L_j(\theta)}{\partial\gamma\partial\gamma^T}&=-\frac{1}{\sigma^2}Z_j^TZ_j\\ \frac{\partial^2L_j(\theta)}{\partial\gamma\partial\sigma^2}&=-\frac{1}{\sigma^4}Z_j^T(y_j-X_j\beta-Z_j\gamma)\\ \frac{\partial^2L_j(\theta)}{\partial\sigma^2\partial\sigma^2}&=\frac{1}{2\sigma^4}-\frac{1}{\sigma^6}(y_j-X_j\beta-Z_j\gamma)^2\end{aligned}\tag{9.75}$$

We now have the score vector and the Hessian matrix available for the complete-data log likelihood function. The Louis information matrix is

$$I(\theta)=-E\left[H(\theta,Z)\right]-E\left[S(\theta,Z)S^T(\theta,Z)\right]\tag{9.76}$$

where the expectations are taken with respect the missing value (Z_j) using the posterior probabilities of QTL genotypes. At the MLE of parameters, $E\left[S(\hat\theta,Z)\right]=0$. Therefore,

$$\begin{aligned}E\left[S(\theta,Z)S^T(\theta,Z)\right]&=\operatorname{var}\left[S(\theta,Z)\right]+E\left[S(\theta,Z)\right]E\left[S^T(\theta,Z)\right]\\ &=\operatorname{var}\left[S(\theta,Z)\right]\end{aligned}\tag{9.77}$$

As a result, an alternative expression of the Louis information matrix is

$$\begin{aligned} I(\theta) &= -E\left[H(\theta,Z)\right] - \text{var}\left[S(\theta,Z)\right] \\ &= -\sum_{j=1}^{n} E\left[H_j(\theta,Z)\right] - \sum_{j=1}^{n} \text{var}\left[S_j(\theta,Z)\right] \qquad (9.78) \\ & \qquad (9.79) \end{aligned}$$

The expectations are

$$E\left[H_j(\theta,Z)\right] = \begin{bmatrix} E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\beta\partial\beta^T}\right) & E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\beta\partial\gamma^T}\right) & E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\beta\partial\sigma^2}\right) \\ E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\gamma\partial\beta^T}\right) & E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\gamma\partial\gamma^T}\right) & E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\gamma\partial\sigma^2}\right) \\ E\left(\frac{L_j(\theta,Z)}{\partial\sigma^2\partial\beta^T}\right) & E\left(\frac{L_j(\theta,Z)}{\partial\sigma^2\partial\gamma^T}\right) & E\left(\frac{L_j(\theta,Z)}{\partial\sigma^2\partial\sigma^2}\right) \end{bmatrix} \qquad (9.80)$$

The six different blocks of the above matrix are

$$\begin{aligned} E\left(\frac{\partial^2 L_j(\theta)}{\partial\beta\partial\beta^T}\right) &= -\frac{1}{\sigma^2}X_j^T X_j \\ E\left(\frac{\partial^2 L_j(\theta)}{\partial\beta\partial\gamma^T}\right) &= -\frac{1}{\sigma^2}X_j^T E(Z_j) \\ E\left(\frac{\partial^2 L_j(\theta)}{\partial\beta\partial\sigma^2}\right) &= -\frac{1}{\sigma^4}X_j^T\left[y_j - X_j\beta - E(Z_j)\gamma\right] \\ E\left(\frac{\partial^2 L_j(\theta)}{\partial\gamma\partial\gamma^T}\right) &= -\frac{1}{\sigma^2}E(Z_j^T Z_j) \\ E\left(\frac{\partial^2 L_j(\theta)}{\partial\gamma\partial\sigma^2}\right) &= -\frac{1}{\sigma^4}E\left[Z_j^T(y_j - X_j\beta - Z_j\gamma)\right] \\ E\left(\frac{\partial^2 L_j(\theta)}{\partial\sigma^2\partial\sigma^2}\right) &= \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}E\left[(y_j - X_j\beta - Z_j\gamma)^2\right] \qquad (9.81) \end{aligned}$$

Again, all the expectations are taken with respect to the missing value Z_j, not the observed phenotype y_j. This is very different from the information matrix of the Fisher scoring algorithm. The variance-covariance matrix of the score vector is

$$\text{var}\left[S(\theta,Z)\right] = \sum_{j=1}^{n}\text{var}\left[S_j(\theta,Z)\right] \qquad (9.82)$$

where var$[S_j(\theta,Z)]$ is a symmetric matrix as shown below,

$$\begin{bmatrix} \text{var}\left(\frac{\partial L_j(\theta,Z)}{\partial\beta}\right) & \text{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\beta}, \frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right) & \text{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\beta}, \frac{\partial L_j(\theta,Z)}{\partial\sigma^2}\right) \\ \text{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\gamma}, \frac{\partial L_j(\theta,Z)}{\partial\beta^T}\right) & \text{var}\left(\frac{\partial L_j(\theta,Z)}{\partial\gamma}\right) & \text{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\gamma}, \frac{\partial L_j(\theta,Z)}{\partial\sigma^2}\right) \\ \text{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\sigma^2}, \frac{\partial L_j(\theta,Z)}{\partial\beta^T}\right) & \text{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\sigma^2}, \frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right) & \text{var}\left(\frac{\partial L_j(\theta,Z)}{\partial\sigma^2}\right) \end{bmatrix} \qquad (9.83)$$

The variances are calculated with respect to the missing value Z_j using the posterior probabilities of QTL genotypes. We only provide the detailed expression of one block of the above matrix. The remaining blocks are left to students for practice. The block that is used as an example is the (1,2) block.

$$\begin{aligned}
&\operatorname{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\beta},\frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right)= \\
&\qquad E\left[\frac{\partial L_j(\theta,Z)}{\partial\beta}\frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right]-E\left[\frac{\partial L_j(\theta,Z)}{\partial\beta}\right]E\left[\frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right]
\end{aligned} \tag{9.84}$$

where

$$\begin{aligned}
E\left[\frac{\partial L_j(\theta,Z)}{\partial\beta}\right]&=\frac{1}{\sigma^2}X_j^T\left[y_j-X_j\beta-E(Z_j)\gamma\right] \\
E\left[\frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right]&=\frac{1}{\sigma^2}E\left[(y_j-X_j\beta-Z_j\gamma)Z_j\right] \\
E\left[\frac{\partial L_j(\theta,Z)}{\partial\beta}\frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right]&=\frac{1}{\sigma^4}E\left[X_j^T(y_j-X_j\beta-Z_j\gamma)^2Z_j\right]
\end{aligned} \tag{9.85}$$

We already learned how to calculate $E(Z_j)$ using the posterior probability of QTL genotype. The other expectations are

$$\begin{aligned}
E\left[(y_j-X_j\beta-Z_j\gamma)Z_j\right]&=\sum_{k=1}^{3}p_j^*(2-k)(y_j-X_j\beta-H_k\gamma)H_k \\
E\left[X_j^T(y_j-X_j\beta-Z_j\gamma)^2Z_j^T\right]&=\sum_{k=1}^{3}p_j^*(2-k)X_j^T(y_j-X_j\beta-H_k\gamma)^2H_k^T
\end{aligned} \tag{9.86}$$

When calculating the information matrix, the parameter θ is substituted by $\hat{\theta}$, the MLE of θ . Therefore, the observed information matrix is

$$I(\hat{\theta})=-E\left[H(\hat{\theta},Z)\right]-\operatorname{var}\left[S(\hat{\theta},Z)\right] \tag{9.87}$$

and the variance-covariance matrix of the estimated parameters is $\operatorname{var}(\hat{\theta})=I^{-1}(\hat{\theta})$.

9.4.3 Hypothesis test

The hypothesis that $H_0:\gamma=0$ can be tested using several different ways. If $\operatorname{var}(\hat{\theta})$ is already calculated, we can use the F- or W-test statistic, which requires $\operatorname{var}(\hat{\gamma})$, the variance-covariance matrix of the estimated QTL effects. It is a submatrix of $\operatorname{var}(\hat{\theta})$. The W-test statistic is

$$W=\hat{\gamma}^T\operatorname{var}^{-1}(\hat{\gamma})\hat{\gamma} \tag{9.88}$$

Alternatively, the likelihood ratio test statistic can be applied to test H_0. We have presented two log likelihood functions, one is the complete-data log likelihood function, denoted by $L(\theta,Z)$, and the other is the observed log likelihood function, denoted by $L(\theta)$. The log likelihood function used to construct the likelihood ratio test statistic is $L(\theta)$, not $L(\theta, Z)$. This complete-data log likelihood function, $L(\theta, Z)$, is only used to derive the EM algorithm and the observed information matrix. The likelihood ratio test statistic is

$$\lambda = -2(L_0 - L_1)$$

where $L_1 = L(\hat{\theta})$ is the observed log likelihood function evaluated at $\hat{\theta} = \{\hat{\beta}, \hat{\gamma}, \hat{\sigma}^2\}$ and L_0 is the log likelihood function evaluated at $\hat{\hat{\theta}} = \{\hat{\hat{\beta}}, 0, \hat{\hat{\sigma}}^2\}$ under the restricted model. The estimated parameter $\hat{\hat{\theta}}$ under the restricted model and L_0 are the same as those given in the section of the least squares method.

9.5 Remarks on the four methods of interval mapping

The LS method (Haley and Knott, 1992) is an approximation of the ML method, aiming to improve the computational speed. The method has been extended substantially to many other situations, e.g., multiple trait QTL mapping (Knott and Haley, 2000) and QTL mapping for binary traits (Visscher et al., 1996). When used for binary and other non-normal traits, the method is no longer called LS. Because of the fast speed, the method remains a popular

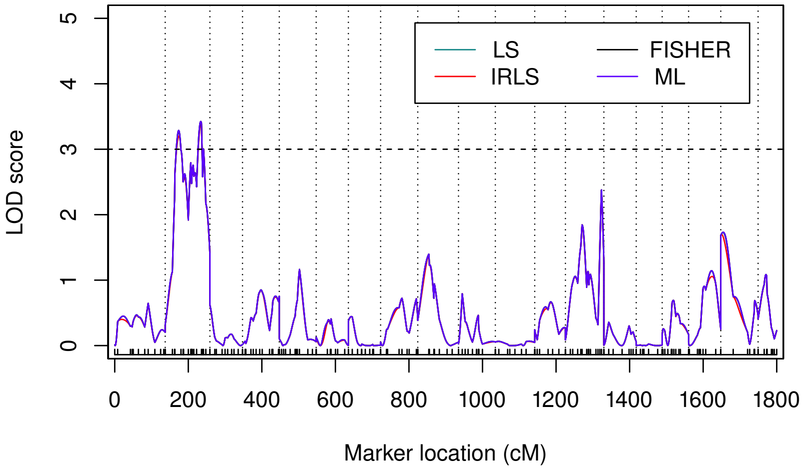

Fig. 9.2. The LOD test statistic profiles for four methods of interval mapping (LS - least square, IRLS - iteratively reweighted least square, FISHER - Fisher scoring, and ML - maximum likelihood). The mouse data were obtained from Lan et al. (2006). The trait investigated is the 10th week body weight. The 19 chromosomes (excluding the sex chromosome) are separated by the vertical dotted lines. The unevenly distributed black ticks on the horizontal axis indicate the marker locations.

method, even though the computer power has increased by many orders of magnitude since the LS was developed. In some literature (e.g., Feenstra et al. (2006)), the LS method is also called the H-K method in honor of the authors, Haley and Knott (1992). Xu (1995) noticed that the LS method, although a good approximation to ML in terms of estimates of QTL effects and test statistic, may lead to a biased (inflated) estimate for the residual error variance. Based on this work, Xu (1998a,b) eventually developed the iteratively reweighted least squares (IRLS) method. In these works (Xu, 1998a,b), the iteratively reweighted least squares was abbreviated IRWLS. Xu (1998b) compared LS, IRLS and ML in a variety of situations and conclude that IRLS is always better than LS and as efficient as ML. When the residual error does not have a normal distribution, which is required by the ML method, LS and IRLS can be better than ML. In other words, LS and IRLS are more robust than ML to the departure from normality. Kao (2000) and Feenstra et al. (2006) conducted more comprehensive investigation on LS, IRLS and ML and found that when epistatic effects exist, LS can generate unsatisfactory results, but IRLS and ML usually map QTL better than LS. In addition, Feenstra et al. (2006) modified the weighted least square method by using the estimating equations (EE) algorithm. This algorithm further improved the efficiency of the weighted least squares by maximizing an approximate likelihood function. Most recently, Han and Xu (2008) developed a Fisher scoring (FISHER) algorithm to maximize the approximate likelihood function. Both the EE and Fisher algorithm maximize the same likelihood function, and thus they produce identical results.

The LS method ignores the uncertainty of the QTL genotype. The IRLS, FISHER (or EE) and ML methods use different ways to extract information from the uncertainty of QTL genotype. If the putative location of QTL overlaps with a fully informative marker, all four methods produce identical result. Therefore, if the marker density is sufficiently high, there is virtually no difference for the four methods. For low marker density, when the putative position is far away from either flanking marker, the four methods will show some difference. This difference will be magnified by large QTL. Han and Xu (2008) compared the four methods in a simulation experiment and showed that when the putative QTL position is fixed in the middle of a 10 cM interval, the four methods generated almost identical results. However, when the interval expands to 20 cM, the differences among the four methods become noticeable.

Interval mapping with a 1 cM increment for the mouse 10th week body weight data were conducted using all the four methods by Han and Xu (2008). The LOD test statistic profiles are shown in Figure 9.2 for the four methods of interval mapping (LS, IRLS, FISHER and ML). There is virtually no differences for the four methods. The difference in LOD profiles is noticeable when the marker density is low. Comparisons for the estimated QTL effects were also conducted for the mouse data. Figure 9.3 shows the estimated QTL effect profiles along the genome for the four methods. Again the difference is barely noticeable.

A final remark on interval mapping is the way to infer the QTL genotype using flanking markers. If only flanking markers are used to infer the genotype of a putative position bracketed by the two markers, the method is called interval mapping. Strictly speaking, interval mapping only applies to fully informative markers because we always use flanking markers to infer the QTL genotype. However, almost all datasets obtained from real life experiments contain missing, uninformative or partially informative markers. To extract maximum information from markers, people always use the multipoint method (Jiang and Zeng, 1997) to infer a QTL genotype. The multipoint method uses more markers or even all markers of the entire chromosome (not just flanking markers) to infer the genotype of a putative position. With the multipoint analysis, we no longer have the notion of interval, and thus interval mapping is no longer an appropriate phrase to describe QTL mapping. Unfortunately, a more appropriate phrase has not been proposed and people are used to the phrase of interval mapping. Therefore, the so called interval mapping in the current literature means QTL mapping under a single QTL model, regardless whether the genotype of a putative QTL position is inferred from flanking markers or all markers.

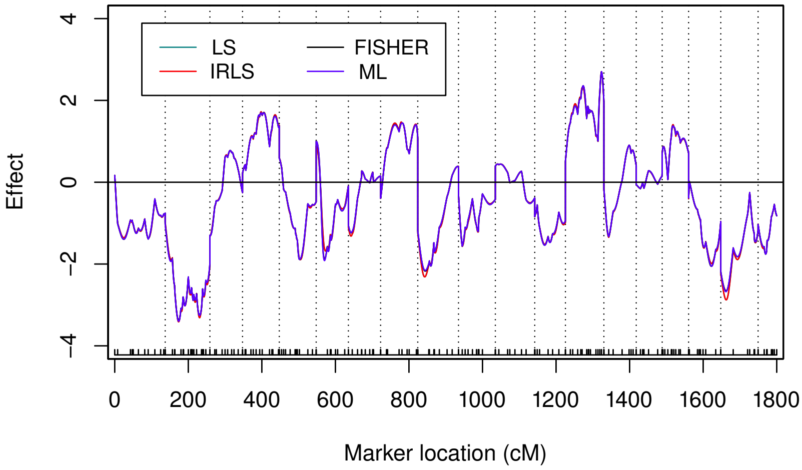

Fig. 9.3. The QTL effect profiles for four methods of interval mapping (LS - least square, IRLS - iteratively reweighted least square, FISHER - Fisher scoring, and ML - maximum likelihood). The mouse data were obtained from Lan et al. (2006). The trait investigated is the 10th week body weight. The 19 chromosomes (excluding the sex chromosome) are separated by the vertical dotted lines. The unevenly distributed black ticks on the horizontal axis indicate the marker locations.

8

Genome Scanning for Quantitative Trait Loci

In the previous chapters, we learned the basic concept of quantitative genetics, the quantitative genetics model, the method for major gene detection (genotypes of the major gene are observed) and the algorithm for segregation analysis (genotypes of the major gene are not observed). We also learned some analytical techniques to analyze a molecular marker linked to a major gene. The real focus of statistical genomics, however, is to identify functional genes that are responsible for the genetic variation of quantitative traits or complex traits if they are not normally distributed. These chapters provide the necessary technology (knowledge preparation) for gene identification, which is the theme of this book.

Molecular markers are not genes but they are inherited following Mendel's laws and their genotypes are observable. The functional genes also follow Mendel's laws of inheritance but their genotypes are not observable. Since both markers and genes are carried by a limited number of chromosomes in the genome, some genes must be physically linked with some markers. If a marker sits in the neighborhood of a gene, the segregation pattern of the marker must be associated with the phenotypic variation of a trait that is controlled by the gene due to linkage. Therefore, we can study marker and trait association and hope to identify important markers that are closely linked to the gene. Since a quantitative trait is often controlled by the segregation of more than one gene, more markers are needed to identify all genes for a quantitative trait. These multiple genes are called quantitative trait loci (QTL). This chapter deals with marker-trait association study in line crosses. The association study using line crosses is different from the association study using randomly sampled populations. The former takes advantage of linkage disequilibrium while the latter assumes no linkage disequilibrium. As a result, markers associated with the trait of interest in line crosses are not equivalent to the genes while markers associated with the traits in randomly sampled populations are most likely the actual genes. The statistical methods for association study, however, are the same, regardless whether the populations are derived from line crosses or not.

8.1 The mouse data

A dataset from an F_2 mouse population consisting of 110 individuals was used as an example for the genetic analysis. The data were published by Lan et al. (2006) and are freely available from the internet. The parents of the F_2 population were B6 (29 males) and BTBR (31 females). The F_2 mice used in this study were measured for various clinical traits related to obesity and diabetes. The framework map consists of 194 microsatellite markers, with an average marker spacing of about 10 cM. The mouse genome has 19 chromosomes (excluding the sex chromosome). The data analyzed in this chapter contain 110 F_2 mice and 193 markers. The second marker (D7Mit76) on chromosome 7 was excluded from the analysis because it overlaps with the first marker (D7Mit56). The 193 markers cover about 1,800 cM of the entire mouse genome. The trait of interest was the 10th week body weight. The marker map, the genotypes of the 110 mice for the 193 markers and the 10th week body weights of the F_2 mice are also provided in the author's personal website (www.statgen.ucr.edu). The files stored in our website are not the original data but preprocessed by our laboratory members and thus they are ready for analysis using QTL mapping software packages such us the QTL procedure in SAS (Hu and Xu, 2009).

8.2 Genome scanning

In major gene identification, we used an F-test statistic to test the significance of a major gene. In genome scanning, we simply treat each marker as a major

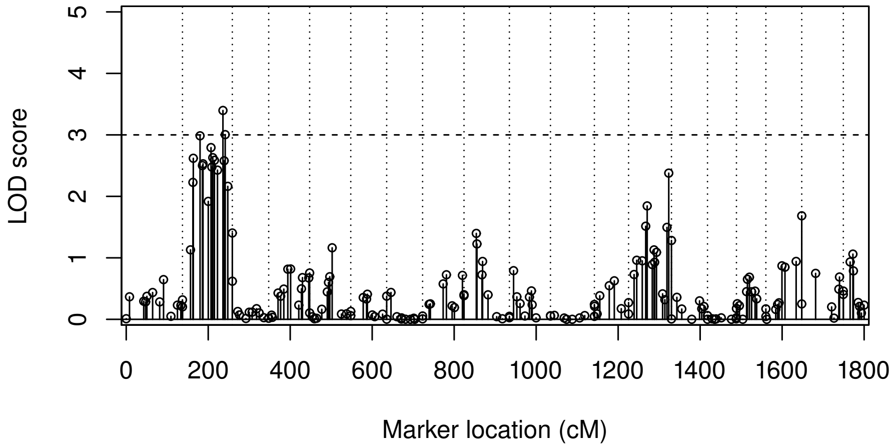

Fig. 8.1. LOD score profile of the entire genome (19 chromosomes) for the 10th week body weight of F_2 mice derived from the cross of two inbred lines. The 19 chromosomes are separated by the dotted reference lines.

gene and analyze every single marker. The test statistics of all markers across the genome are plotted against the genome location of the markers, forming a test statistic profile. Some regions of the genome may show peaks while majority of the genome may be flat. The regions with peaks may suggest QTL nearby the peaks. If the marker density is sufficiently high, some markers may actually overlap with the QTL. The genome scanning is also called individual marker analysis. We scan all markers across the genome, but with one marker at a time. The entire genome scanning requires many repeated single marker analysis. The genetic model and test statistic in genome scanning are of no difference from the major gene detection except that we now deal with multiple markers. Sometimes, investigators already have prior knowledge about the functions of some genes. The functions may be related to the development of the quantitative trait of interest. These genes are called candidate genes for the trait of interest. Genome scanning may also include these candidate genes. Figure 8.1 shows the LOD score profile of the mouse genome for the trait of 10th week body weight (wt10week). Note that the LOD (log of odds) score is often used in human genetics. The Wald-test statistic is often converted into the LOD score using (see a later section for the definition of LOD)

$$\text{LOD} = \frac{W}{2 \times \ln(10)} \tag{8.1}$$

There are many peaks in the LOD score profile, but the peaks in chromosome 2 appear to be too high ($\text{LOD} > 3$) to be explained by chance. Therefore, one or more QTL may exist in chromosome two for this trait.

The model used for the analysis is called the additive genetic model because the dominance effect has been ignored. Figure 8.2 shows the additive effects plotted against markers, the so called QTL effect profile. We can see that QTL effects in some regions of the genome are positive while in in other regions they are negative. The way we coded the genotypes determined the signs of the QTL effects. Assume that the original genotypic data were coded as ‘A’ for line B6, ‘B’ for line BTBR and ‘H’ for heterozygote. We numerically recoded the genotype as 1 for ‘A’, 0 for ‘H’ and -1 for ‘B’. Based on this coding system, a negative QTL effect means that the B6 allele is “low” and the BTBR allele is “high”. Therefore, the QTL allele carried by B6 in the second chromosome is the “low” allele, i.e., it is responsible for the low body weight. Of course, if ‘A’ and ‘B’ alleles represent BTBR and B6, respectively, the negative and positive signs should be explained in just the opposite way.

8.3 Missing genotypes

In the section of major gene detection, we assumed that the genotype of a major gene is observed for every individual. In the section of segregation analysis, the genotype of the major gene is missing for every individual. This section deals with marker analysis. Although most individuals in the mapping population should be genotyped for all markers, still some individuals may not be genotyped for some markers, either due to technical errors or human errors. If an individual is not genotyped for all markers, this individual should be eliminated from the analysis. However, most individuals may just have a few missing genotypes. These individuals must be included in the analysis; otherwise, we may not have enough sample size to perform genetic mapping. We now use the $\mathtt{F}_2$ population as an example to show how to deal with the missing marker problem.

Let y_j be the phenotypic value of individual j and it can be described by the following linear model,

$$y_j = b_0 + X_j b_1 + e_j, \tag{8.2}$$

where b_0 is the intercept, b_1 is the additive genetic effect, i.e., a, and e_j is the residual error. The genotype indicator variable X_j depends on the genotype of the marker under consideration. Let us define X_j as

$$X_j = \begin{cases} +1 & \text{for} A_1A_1 \\ \ \ 0 & \text{for} A_1A_2 \\ -1 & \text{for} A_2A_2 \end{cases} . \tag{8.3}$$

Let G_j be the genotype of the marker under consideration and $p_j(1) = \Pr(G_j = A_1A_1|\text{marker})$ be the probability of $G_j = A_1A_1$ given the genotypes of the two markers flanking the marker of interest. Similarly, let $p_j(0) = \Pr(G_j = A_1A_2|\text{marker})$ and $p_j(-1) = \Pr(G_j = A_2A_2|\text{marker})$. The conditional expectation of X_j given the flanking marker genotypes is

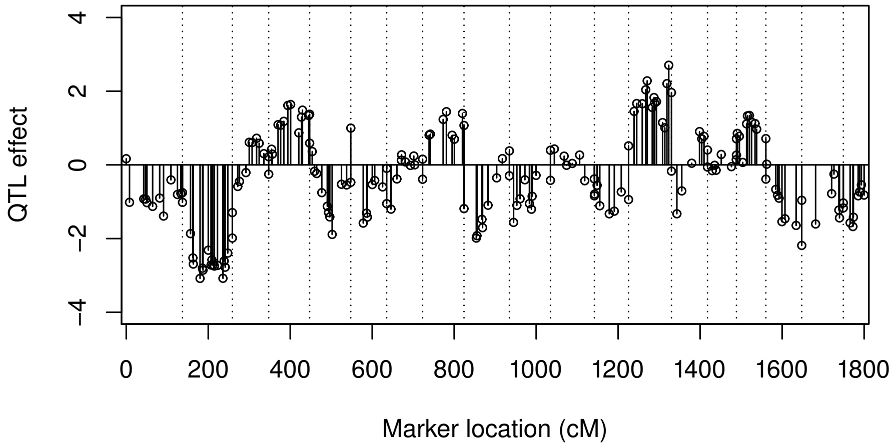

Fig. 8.2. QTL effect profile of the entire genome (19 chromosomes) for the 10th week body weight of $\mathtt{F}_2$ mice derived from the cross of two inbred lines. The 19 chromosomes are separated by the dotted reference lines.

$E(X_j) = p_j(1) - p_j(-1)$. The model for missing markers is the same as equation (8.2) except that X_j is replaced by $E(X_j)$, i.e.,

$$y_j = b_0 + E(X_j)b_1 + e_j. \tag{8.4}$$

To include the dominance effect, the revised model is

$$y_j = b_0 + E(X_{j1})b_1 + E(X_{j2})b_2 + e_j \tag{8.5}$$

where X_{j1} is the genotype indicator variable for the additive effect, as defined early (equation 8.3) and X_{j2} is the genotype indicator variable for the dominance effect,

$$X_{j2} = \begin{cases} 0 & \text{for} A_1A_1 \\ 1 & \text{for} A_1A_2 \\ 0 & \text{for} A_2A_2 \end{cases}. \tag{8.6}$$

The conditional expectation of X_{j2} is simply $E(X_j) = p_j(0)$. The second regression coefficient b_2 is the dominance effect, i.e., $b_2 = d$.

8.4 Test statistics

There are many different test statistics we can use for genome scanning. The one we learned in the major gene detection is the F-test statistic. We now discuss the test statistics when only a single model effect is subject to test. In genetic analysis, this is equivalent to testing only the additive effect. Let $\hat{b}_1$ be the estimated genetic effect and $\sigma^2_{\hat{b}_1}$ be the variance of the estimate. The F-test statistic for the null hypothesis $H_0 : b_1 = 0$ is

$$F = \frac{\hat{b}_1^2}{\sigma^2_{\hat{b}_1}}. \tag{8.7}$$

This F-test statistic appears to be different from the F-test statistic occurring in the analysis of variances (ANOVA). The latter is defined as the ratio of the between group mean squares MS_B to the within group mean squares MS_W. However, the two test statistics are two different forms of the same test statistic (derivation is not shown). As an F-test statistic, it will follow an F distribution with a numerator degree of freedom 1 and a denominator degree of freedom $n - 2$.

A single genetic effect can also be tested using the t-test statistic. The t-test statistic is simply the square root of the F-test statistic,

$$t = \sqrt{F} = \frac{|\hat{b}_1|}{\sigma_{\hat{b}_1}} \tag{8.8}$$

Under the null hypothesis $H_0 : b_1 = 0$, this test statistic will follow a t distribution with $n - 2$ degrees of freedom. As the sample size increases, $n - 2$

is not much different from n, therefore, the degrees of freedom in the F-test and the t-test is approximately equal to the sample size.

When $n \to \infty$, the F-test will be identical to the χ^2-test statistic, which follows a χ^2 distribution with one degree of freedom. The corresponding t-test statistic will approach to the Z-test statistic.

The F-test statistic in the form of $F = \frac{\hat{b}_1^2}{\sigma^2_{\hat{b}_1}}$ is actually called the Wald-test statistic or simply W-test statistic (Wald, 1943). Although the Wald-test statistic is not called as often as the F- and t-test statistics in genome scanning, it will be used more often here in this text book due to the fact that the Wald-test statistic is comparable or similar to the likelihood ratio test statistic.

The likelihood ratio test (LRT) statistic is defined as

$$\lambda = -2[L_0(\hat{\theta}_0) - L_1(\hat{\theta}_1)] \tag{8.9}$$

where $L_0(\hat{\theta}_0)$ is the log likelihood function evaluated under the null model ($H_0 : b_1 = 0$) and $L_1(\hat{\theta}_1)$ is the log likelihood function evaluated under the full model ($H_1 : b_1 \neq 0$). The null model and the full model differ by one parameter, i.e., $\theta_0 = \{b_0, \sigma^2\}$ and $\theta_1 = \{b_0, b_1, \sigma^2\}$. We often call the null model the restricted model or reduced model because it has $b_1 = 0$ as the restriction or simply have one parameter less than the full model. Because $L_1(\hat{\theta}_1)$ is guaranteed to be larger than $L_0(\hat{\theta}_0)$, the log likelihood difference is negative. A negative test statistic looks strange, and thus we put a minus sign in front of the different to make the test statistic positive. The constant multiplier 2 is simply to make the likelihood ratio test statistic follow a standard distribution under the null model. This standard distribution happens to be a χ^2 distribution with one degree of freedom. The degree of freedom is one, not any other value, because the null model has one parameter less than the full model.

We now realize that the Wald-test statistic, the F-test statistic and the likelihood ratio test statistic, all approach a χ^2 distribution with one degree of freedom as the sample size is sufficiently large. Therefore, these three test statistics can be used interchangeably with very little difference, although the likelihood ratio test statistic is considered a slightly better test-statistic than the others.

The likelihood ratio test statistic is defined using the natural logarithm, i.e., the logarithm with base $e \approx 2.718281828459$. In human genetics, people often use the LOD (Log of Odds) score as the test statistic. Let $L_0 = L_0(\hat{\theta}_0)$ and $L_1 = L_1(\hat{\theta}_1)$ be short expressions of the natural logarithms of the likelihood functions under the null model and the full model, respectively. The original likelihood functions (before taking the natural log) are $l_0 = e^{L_0}$ and $l_1 = e^{L_1}$, respectively. The LOD score is defined as

$$\begin{aligned}\text{LOD} =& \log_{10}\left(\frac{l_1}{l_0}\right) = \log_{10}\left(\frac{e^{L_1}}{e^{L_0}}\right) \\ =& \log_{10} e^{L_1} - \log_{10} e^{L_0} = \log_{10} e^{(L_1 - L_0)}\end{aligned} \tag{8.10}$$

It is the log of the likelihood ratio with base 10 rather than base e. The relationship between LOD and the likelihood ratio test statistic (λ) is

$$\begin{aligned}\text{LOD} =& \log_{10} e^{(L_1 - L_0)} = \tfrac{1}{2}\log_{10} e^{[-2(L_0 - L_1)]} \\ =& [-2(L_0 - L_1)]\left(\tfrac{1}{2}\log_{10} e\right) = \lambda\left(\tfrac{1}{2}\log_{10} e\right)\end{aligned} \tag{8.11}$$

The constant $\frac{1}{2}\log_{10} e \approx 0.2171$ and the inverse of the constant is approximately 4.6052. Therefore, we may use the following approximation to convert λ to LOD,

$$\text{LOD} = 0.2171\,\lambda = \frac{\lambda}{4.6052} \tag{8.12}$$

The LOD score has an intuitive interpretation because of the base 10. A LOD score of x means that the full model is 10^x times more likely than the restricted model. For example, a LOD score 3 means that the full model (with the marker effect) is 1000 times more likely than the reduce model (without the marker effect).

We now turn our attention to the hypotheses where two or more genetic effects are tested simultaneously. For example, in an $\texttt{F}_2$ population, we can test both the additive and dominance effects. The null hypothesis is $H_0 : a = d = 0$ or $H_0 : b_1 = b_2 = 0$. In this case, the t-test is not a valid choice, because it is designed for testing only a single effect. The F-test, although can be used, is rarely chosen as the test statistic for genome scanning. The F-test statistic for testing two effects is defined as

$$F = \frac{1}{2}[\,\hat{b}_1\ \hat{b}_2\,]\begin{bmatrix}\text{var}(\hat{b}_1) & \text{cov}(\hat{b}_1, \hat{b}_2) \\ \text{cov}(\hat{b}_1, \hat{b}_2) & \text{var}(\hat{b}_2)\end{bmatrix}^{-1}\begin{bmatrix}\hat{b}_1 \\ \hat{b}_2\end{bmatrix} \tag{8.13}$$

The $\frac{1}{2}$ multiplier appears because we are testing two effects. If we test k effects simultaneously, the multiplier will be $\frac{1}{k}$, and the dimensionality of the effect vector and the variance matrix will be changed to $k \times 1$ and $k \times k$ accordingly. The F-test statistic follows an F-distribution with degrees of freedom k and $n - (k + 1)$ or simply k and n when n is sufficiently large.

In contrast to the test for a single genetic effect where the F-test statistic is equivalent to the W-test statistic, when testing two or more effects, the W-test statistic is

$$W = [\,\hat{b}_1\ \hat{b}_2\,]\begin{bmatrix}\text{var}(\hat{b}_1) & \text{cov}(\hat{b}_1, \hat{b}_2) \\ \text{cov}(\hat{b}_1, \hat{b}_2) & \text{var}(\hat{b}_2)\end{bmatrix}^{-1}\begin{bmatrix}\hat{b}_1 \\ \hat{b}_2\end{bmatrix} \tag{8.14}$$

The relationship between the W-test and the F-test is $W = kF$. When the sample size is sufficiently large, the W-test statistic will approach a χ^2 distribution with k degrees of freedom ($k = 2$ in this case).

The corresponding likelihood ratio test statistic for two or more effects has the same form as that for testing a single effect except that $\theta_0 = \{b_0, \sigma^2\}$ under the null model has two parameters less than the $\theta_1 = \{b_0, b_1, b_2, \sigma^2\}$ under the full model. As a result, the λ test statistic follows a χ^2 distribution with $k = 2$ degrees of freedom. Both the W-test and the likelihood ratio test statistics follow the same χ^2 distribution, and thus they can be used interchangeably with very little difference.

The W-test statistic requires calculation of the variance-covariance matrix of the estimated parameters and its inverse. However, some of the algorithms for parameter estimation, e.g., the EM algorithm, do not have an automatic way to calculate this matrix. For these methods, the likelihood ratio test statistic may be preferred because of the ease of calculating the test statistic. When both the W-test and the λ-test statistics are available, which one is better? The answer is that the λ-test statistic is more desirable if the sample size is small. For large samples sizes, these two test-statistics are virtually the same.

In summary, the W-test and the λ-test statistics are preferable for genome scanning because they can be used for testing both a single effect and multiple effects (compared to the t-test and the Z-test which are only useful for testing a single effect). The LOD score test statistic is simply a rescaled likelihood ratio test statistic, and thus they are used interchangeably without any difference at all.

8.5 Bonferroni correction

Genome scanning involves multiple tests. Sometimes the number of tests may reach hundreds or even thousands. For a single test, the critical value for any test statistic simply takes the 95% or 99% quantile of the distribution that the test statistic follows under the null hypothesis. For example, the F-test statistic follows an F distribution, the likelihood ratio test statistic follows a chi-square distribution and the W-test statistic also follows a chi-square distribution. When multiple tests are involved, the critical value used for a single test must be adjusted to make the experiment-wise Type I error at a desired level, say 0.05.

The Bonferroni correction is a multiple test correction used when multiple statistical tests are being performed in a single experiment (Dunn, 1961). While a given alpha value α may be appropriate for each individual test, it is not for the set of all tests involved in a single experiment. In order to avoid spurious positives, the alpha value needs to be lowered to account for the number of tests being performed. The Bonferroni correction sets the Type I error for the entire set of k tests equal to β by taking the alpha value for each test equal to α. The β is now called the experiment-wise Type I error rate and α is called the test-wise Type I error rate or nominal Type I error rate. The Bonferroni correction states that, in an experiment involving k tests, if you want to control the experiment-wise Type I error rate at β, the nominal Type I error rate for a single test should be

$$\alpha = \frac{\beta}{k} \tag{8.15}$$

For example, if an experiment involves 100 tests and the investigator wants to control the experiment-wise Type I error at $\beta = 0.05$, for each of the individual tests, the nominal Type I error rate should be $\alpha = \frac{\beta}{k} = \frac{0.05}{100} = 0.0005$. In other words, for any individual test the p-value should be less than 0.0005 in order to declare significance for that test. The Bonferroni correction does not require independence of the multiple tests.

When the multiple tests are independent, there is an alternative correction for the Type I error, which is called the Šidák correction (Abdi, 2007). This correction is often confused with the Bonferroni correction. If a test-wise Type I error is α, the probability of non-significance is $1 - \alpha$ for this particular test. For k independent tests and none of them is significant, the probability is $(1 - \alpha)^k$. The experiment-wise Type I error is defined as the probability that at least one of the k tests is significant. This probability is

$$\beta = 1 - (1 - \alpha)^k \tag{8.16}$$

To find the nominal α value given the experiment wise value β, we use the reverse function

$$\alpha = 1 - (1 - \beta)^{1/k} \tag{8.17}$$

This correction is the Šidák correction. The two corrections are approximately the same when β is small because $(1 - \beta)^{1/k} \approx 1 - \frac{\beta}{k}$ and thus

$$\alpha \approx \frac{\beta}{k} \tag{8.18}$$

Therefore, the Bonferroni correction is an approximation of the Šidák correction for multiple independent tests for small β.

8.6 Permutation test

When the number of tests is large, the Bonferroni and Šidák corrections tend to be over conservative. In addition, if a test statistic does not follow any standard distribution under the null model, calculation of the p-value may be difficult for each individual test. In this case, we can adopt the permutation test to draw an empirical critical value. This method was developed by Churchill and Doerge (1994) for QTL mapping. The idea is simple, but implementation can be time consuming. When the sample size n is small, we can evaluate all $n!$ different permuted samples of the original phenotypic values while keeping the marker genotype data intact. In other words, we

Table 8.1. Phenotypic values of trait y and the genotypes of five markers from ten plants (the original data set)

plant	y	M_1	M_2	M_3	M_4	M_5
1	55.0	H	H	H	H	A
2	54.2	H	H	H	H	U
3	61.6	H	H	U	A	A
4	66.6	H	H	H	H	U
5	67.4	H	H	H	B	U
6	64.3	H	H	H	H	H
7	54.0	H	A	B	B	B
8	57.2	H	B	H	H	H
9	63.7	H	H	H	H	H
10	55.0	H	H	A	H	U

only reshuffle the phenotypes, not the marker genotypes. For each permuted sample, we apply any method of genome scanning to calculate the test statistical values for all markers. In each of the permuted sample, the association of the phenotype and genotypes of markers has been (purposely) destroyed so that the distribution of the test statistics will mimic the actual distribution under the null model, from which a desirable critical value can be drawn from the empirical null distribution. The number of permuted samples can be extremely large if the sample size is large. In this case, we can randomly reshuffle the data to purposely destroy the association between the phenotype and the marker genotype. By random reshuffling the phenotypes, individual j may take the phenotypic value of individual i for $i \neq j$ while the marker genotype of individual j remains unchanged. After reshuffling the phenotypes, we analyze the data and scan the entire genome. By chance, we may find some peaks in the test statistic profile. We know that these peaks are false because we have already destroyed the association between markers and phenotypes. We record the value of the test statistic at the highest peak of the profile and denote it by λ_1. We then reshuffle the data and scan the genome again. We may find some false peaks again. We then record the highest peak and write down the value, λ_2, and put it in the data set. We repeat the reshuffling process many times to form a large sample of λ's, denoted by $\{\lambda_1, \ldots, \lambda_M\}$, where M is a large number, say 1000. These λ values will form a distribution, called the null distribution. The 95% or 99% quantile of the null distribution is the empirical critical value for our test statistic. We then compare our test statistic for each marker (from the original data analysis) against this empirical critical value. If the test statistic of a marker is larger than this critical value, we can declare this marker as being significant. Note that permutation test is time consuming, but it is realistic with the advanced computing system currently available in most laboratories.

We now provide an example to show how to use the permutation test to draw the critical value of a test statistic. Table 8.1 gives a small sample of ten plants (the original dataset).

Ten randomly reshuffled samples are demonstrated in Table 8.2. We can see that the first observation of sample 1 (S_1) takes the phenotype of plant number 6 while the genotypes of the five markers remain unchanged. Another example is that the second observation of sample 2 (S_2) takes the phenotype of plant number 8 while the genotypes of the five markers are still the genotypes for plant number 2. The phenotypic values corresponding to the ten reshuffled samples are given in Table 8.3. Each sample is subject to genome scanning, i.e., five F-test statistics are calculated, one for each marker. The maximum F-test statistic value for each reshuffled sample is given in the last row of Table 8.3. For example, the maximum F-value in the first sample (S_1) is 2.12, the maximum F-value for S_2 is 1.28 and so on. We then sorted the ten F-values from the ten samples in descending order as shown in the following sequence,

Table 8.2. Plant ID's of ten randomly reshuffled samples, denoted by $S_1, S_2, \ldots, S_{10}$, respectively

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
6	5	9	10	5	6	2	10	1	10
2	8	6	8	1	2	7	6	4	5
4	1	5	7	8	4	1	9	3	9
3	10	1	4	10	3	10	8	5	3
8	3	2	1	7	8	4	5	8	1
9	4	3	5	2	9	3	3	7	8
10	7	10	2	4	10	8	2	9	2
1	9	7	6	3	1	9	1	10	6
7	6	8	3	6	7	6	7	6	7
5	2	4	9	9	5	5	4	2	4

$$\{4.51, 3.95, 3.51, 3.21, 2.33, 2.12, 1.85, 1.28, 1.11, 0.95\}$$

The ten F-values are assumed to be sampled from the null distribution. The empirical 90% quantile is 3.95, which can be used as the critical value for the F-test statistic to compare under the Type I error of 0.10. The number of reshuffled samples in the example is not sufficiently large to give 95% quantile for the Type I error of $\alpha = 0.05$. In practice, the number of randomly reshuffled samples depends on $\alpha = 0.05$ due to Monte Carlo error. Nettleton and Doerge (2000) recommended that the permutation sample size should be at least $\frac{5}{\alpha}$, where α is the experiment-wise Type I error rate. In permutation analysis, there is no such a thing as nominal Type I error. In practice, we often choose 1000 as the permutation sample size.

Permutation test is not a method for genome scanning; rather, it is only a way to draw an empirical critical value of a test statistic for us to decide statistical significance of a marker. It applies to all test statistics, e.g., the F-test, the W-test, the likelihood ratio test and so on. The phrase “permutation test” can be confusing because it is not a method for significance test. “Permutation analysis” may be a better phrase to describe this empirical approach of critical value calculation.

8.7 Piepho’s approximate critical value

Permutation analysis is perhaps the best method for drawing the empirical critical value to control the genome wise Type I error rate. However, it can be time consuming.Piepho (2001) developed an approximate method, which does not require randomly reshuffling of the data. The method simply uses exiting test statistical values of all points across the genome. The test statistic must be the likelihood ratio test statistic. If the test statistic is the LOD score, a simple conversion to the likelihood ratio test statistic is required. The W-test statistic may also be used because it also follows a chi-square distribution under the null model. Let $\beta = 0.05$ be the genome wise Type I error, $C = \chi^2_{k,1-\alpha}$ be the $(1-\alpha) \times 100\%$ quantile of the chi-square distribution and k (the degrees of freedom of the test statistic) is the number of genetic effects subject to statistical test, where $k = 1$ for a BC design and $k = 2$ for an F_2 design. The following relations provides a way to solve for $C = \chi^2_{k,1-\alpha}$, the critical value for the likelihood ratio test statistic to compare so that the genome wise Type I error is controlled at β.

$$\beta = m\Pr(\chi^2_k > C) + \frac{2^{-\frac{1}{2}k} C^{-\frac{1}{2}(1-k)} e^{-\frac{1}{2}C}}{\Gamma(\frac{k}{2})} \sum_{i=1}^{m} v_i \tag{8.19}$$

where m is the number of chromosomes and $\Gamma(\frac{k}{2})$ is the Gamma function. The v_i for the ith chromosome is defined as

Table 8.3. The corresponding phenotypic values and the maximum F-test statistical values (last row) in the ten randomly reshuffled samples

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
	64.3	67.4	63.7	55.0	67.4	64.3	54.2	55.0	55.0	55.0
	54.2	57.2	64.3	57.2	55.0	54.2	54.0	64.3	66.6	67.4
	66.6	55.0	67.4	54.0	57.2	66.6	55.0	63.7	61.6	63.7
	61.6	55.0	55.0	66.6	55.0	61.6	55.0	57.2	67.4	61.6
	57.2	61.6	54.2	55.0	54.0	57.2	66.6	67.4	57.2	55.0
	63.7	66.6	61.6	67.4	54.2	63.7	61.6	61.6	54.0	57.2
	55.0	54.0	55.0	54.2	66.6	55.0	57.2	54.2	63.7	54.2
	55.0	63.7	54.0	64.3	61.6	55.0	63.7	55.0	55.0	64.3
	54.0	64.3	57.2	61.6	64.3	54.0	64.3	54.0	64.3	54.0
	67.4	54.2	66.6	63.7	63.7	67.4	67.4	66.6	54.2	66.6
F-test	2.12	1.28	3.51	1.85	2.33	4.51	3.21	3.95	1.11	0.95

$$v_i = \left|\sqrt{\lambda_1} - \sqrt{\lambda_2}\right| + \left|\sqrt{\lambda_2} - \sqrt{\lambda_3}\right| + ... + \left|\sqrt{\lambda_{m_i-1}} - \sqrt{\lambda_{m_i}}\right| \tag{8.20}$$

where λ_l for $l = 1, ..., m_i$ is the likelihood ratio test statistic for marker l in chromosome i, and m_i is the total number of markers in chromosome i. Once β is given, the above equation is simply a function of C. A numerical solution can be found using the bi-section algorithm. Once C is found, the Type I error for an individual marker can be obtained using the inverse function of the chi-square distribution function. The Gamma function $\Gamma(\frac{k}{2})$ depends on k, the number of genetic effects. For the common designs of experiments, k only takes 1, 2 or 3. For example, $k = 1$ for BC, DH (double haploid) and RIL (recombinant inbred line) designs. If the additive effect is the only one to be tested in the F_2 design, k also equals 1. If both the additive and dominance effects are tested in the F_2 design, $k = 2$. In a four-way cross design, k equals 3. Therefore, we only need the value of $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, $\Gamma(\frac{2}{2}) = \Gamma(1) = 1$ and $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$.

8.8 Theoretical consideration

Genome scanning described in this chapter refers to marker analysis. Since a marker is not a QTL, the estimated marker effect only reflects a fraction of the QTL effect. Take a BC design as an example. Assume that a QTL with effect a is d cM away from a marker. If we use this marker to estimate the QTL effect, the marker effect will not be equal to a; rather, it will be $(1 - 2r)a$, where $r = \frac{1}{2}(1 - e^{-d/(2\times100)})$ is the recombination fraction between the marker and the QTL. The marker effect is only a fraction of the QTL effect. This fraction is $(1 - 2r)$, which is the correlation coefficient between the marker and the QTL genotype indicator variables. When $r = 0$, a situation where the marker overlaps with the QTL, the marker effect is identical to the QTL effect. On the other hand, if $r = 0.5$, a situation where the marker is not linked to the QTL, the marker effect equals zero, regardless how large the QTL effect is. When $0 < r < 0.5$, what we estimate for the marker is a confounded effect between the QTL effect and the linkage parameter. A small marker effect may be due to a large QTL effect but weakly linked to the marker or a small QTL effect with a strong linkage. There is no way to tell the actual QTL effect unless more markers are taken into account simultaneously, which is the topic to be addressed in the next chapter when interval mapping is introduced.

We now prove that the correlation between the marker and the QTL is $1 - 2r$. Let X be the indicator variable for the QTL genotype, i.e., $X = 1$ for A_1A_1 and $X = 0$ for A_1A_2. Let M be the corresponding indicator variable for the marker genotype, i.e., $M = 1$ and $M = 0$, respectively, for the two genotypes of the marker. The joint distribution of X and M is given in Table 8.4. This joint distribution table is symmetrical, meaning that both M and X have the same marginal distribution. First, let us look at the marginal distribution of variable

		M		
		1	0	
X	1	$(1-r)/2$	$r/2$	1/2
	0	$r/2$	$(1-r)/2$	1/2
		1/2	1/2	

Table 8.4. Joint distribution of X (QTL genotype) and M (marker genotype)

M. From the joint distribution table, we get $\Pr(M=1) = \frac{1-r}{2} + \frac{r}{2} = \frac{1}{2}$ and $\Pr(M=0) = \frac{r}{2} + \frac{1-r}{2} = \frac{1}{2}$. Therefore, the variance of M is

$$\begin{aligned} \mathrm{var}(M) =& E(M^2) - E^2(M) \\ =& \left(\frac{1}{2} \times 1^2 + \frac{1}{2} \times 0^2\right) - \left(\frac{1}{2} \times 1 + \frac{1}{2} \times 0\right)^2 \\ =& \frac{1}{4} \end{aligned} \tag{8.21}$$

Similarly, $\mathrm{var}(X) = \frac{1}{4}$, due to the symmetrical nature. We now evaluate the covariance between M and X.

$$\begin{aligned} \mathrm{cov}(M, X) =& E(MX) - E(M)E(X) \\ =& \frac{1-r}{2} \times 1 \times 1 - \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}\left[2(1-r) - 1\right] \\ =& \frac{1}{4}(1-2r) \end{aligned} \tag{8.22}$$

The correlation coefficient between the two variables is

$$\rho_{MX} = \frac{\mathrm{cov}(M, X)}{\sqrt{\mathrm{var}(M)\mathrm{var}(X)}} = \frac{\frac{1}{4}(1-2r)}{\frac{1}{4}} = 1 - 2r \tag{8.23}$$

Because of the symmetry of X and M, this correlation is also equal to the regression coefficient, i.e.,

$$\beta_{XM} = \frac{\mathrm{cov}(M, X)}{\mathrm{var}(M)} = \frac{\frac{1}{4}(1-2r)}{\frac{1}{4}} = 1 - 2r \tag{8.24}$$

Recall that when a marker is used to estimate the effect of a linked QTL, the QTL effect will be biased by a factor $(1-2r)$. This fraction is the correlation between X and M. In fact, it is the regression coefficient of X on M. Because $\rho_{MX} = \beta_{XM}$, we say $(1-2r)$ is the correlation coefficient. We now show why the factor of reduction is β_{XM}. Recall that the QTL model is

$$y_j = b_0 + E(X_j|M_j)b_1 + e_j \tag{8.25}$$

where $E(X_j|M)$ is the conditional mean of X_j given M_j. We use marker genotype M_j to infer the QTL genotype X_j. The conditional mean can be expressed as the predicted value of X from M using the following regression equation,

$$E(X_j|M_j) = E(X_j) + M_j\beta_{XM} = \frac{1}{2} + M_j(1 - 2r) \tag{8.26}$$

Substituting this equation (8.26) into the above model (equation 8.25), we get

$$\begin{aligned} y_j &= \left(b_0 + \tfrac{1}{2}\right) + M_j(\beta_{XM}b_1) + e_j \\ &= b_0^* + M_j b_1^* + e_j \end{aligned} \tag{8.27}$$

where $b_0^* = b_0 + \frac{1}{2}$ and $b_1^* = \beta_{XM}b_1 = (1 - 2r)a$. Note that $b_1 = a$ is the genetic effect and $\beta_{XM} = 1 - 2r$ as given in equation (8.24).

9

Interval Mapping

Interval mapping is an extension of the individual marker analysis so that two markers are analyzed at a time. In the marker analysis (Chapter 8), we cannot estimate the exact position of a QTL. With interval mapping, we use two markers to determine an interval, within which a putative QTL position is proposed. The genotype of the putative QTL is not observable but can be inferred with a certain probability using the three-point or multipoint method introduced in Chapter 4. Once the genotype of the QTL is inferred, we can estimate and test the QTL effect at that particular position. We divide the interval into many putative positions of QTL with one or two cM apart and investigate every putative position within the interval. Once we have searched the current interval, we move on to the next interval and so on until all intervals have been searched. The putative QTL position (not necessarily at a marker) that has the maximum test statistical value is the estimated QTL position. Figure 9.1 demonstrates the process of genome scanning for markers only (panel a), for markers and virtual markers (panel b) and for every point of the chromosome (panel c).

Interval mapping was originally developed by Lander and Botstein (1989) and further modified by numerous authors. Interval mapping has revolutionized genetic mapping because we can really pinpoint the exact location of a QTL. In each of the four sections that follow, we will introduce one specific statistical method of interval mapping based on the $\mathtt{F}_2$ design. Methods of interval mapping for a BC design is straightforward and thus will not be discussed in this chapter. Maximum likelihood (ML) method of interval mapping (Lander and Botstein, 1989) is the optimal method for interval mapping. Least squares (LS) method (Haley and Knott, 1992) is a simplified approximation of Lander and Botstein method. The iteratively reweighted least squares (IRLS) method (Xu, 1998a,b) is a further improved method over the least squares method. Recently Feenstra et al. (2006) developed an estimating equation (EE) method for QTL mapping, which is an extension of the IRLS with improved performance. Han and Xu (2008) developed a Fisher scoring algorithm (FISHER) for QTL mapping. Both the EE and FISHER algorithms maximize the same likelihood function and thus they generate identical result. In this chapter, we introduce the methods based on their simplicity rather than their chronological orders of development. Therefore, the methods will be introduced in the following order: LS, IRLS, FISHER and ML. Bayesian method will be discussed in a later chapter where multiple QTL mapping is addressed.

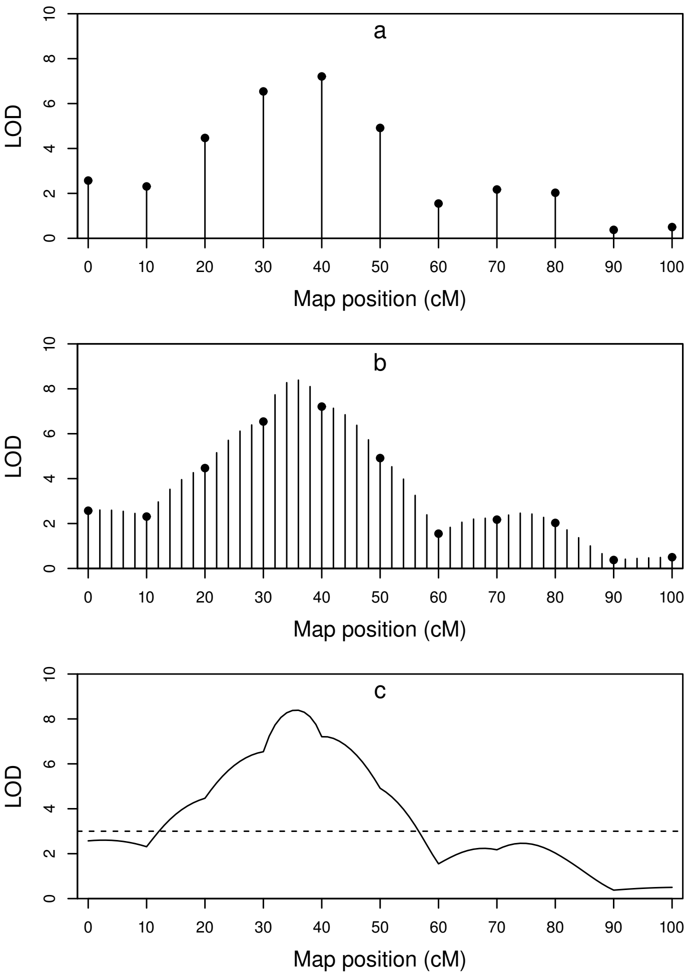

Fig. 9.1. The LOD test statistics for (a) marker effects (top panel), (b) virtual marker effects (panel in the middle), and (c) every point of a simulated chromosome (bottom panel).

9.1 Least squares method

The LS method was introduced by Haley and Knott (1992) aiming to improving the computational speed. The statistical model for the phenotypic value of the jth individual is

$$y_j = X_j\beta + Z_j\gamma + \varepsilon_j \tag{9.1}$$

where β is a $p \times 1$ vector for some model effects that are irrelevant to QTL effects, X_j is a $1 \times p$ known design vector, $\gamma = \{a, d\}$ is a 2×1 vector for QTL effects of a putative locus (a for additive effect and d for dominance effect), Z_j is a 1×2 vector for the genotype indicator variable defined as

$$Z_j = \begin{cases} H_1 & \text{for } A_1A_1 \\ H_2 & \text{for } A_1A_2 \\ H_3 & \text{for } A_2A_2 \end{cases} \tag{9.2}$$

where H_k for $k = 1, 2, 3$ is the kth row of matrix

$$H = \begin{bmatrix} +1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{9.3}$$

The residual error ε_j is assumed to be a $N(0, \sigma^2)$ variable. Although normal distribution for ε_j is not a required assumption for the LS method, it is required for the ML method. It is important to include non-QTL effects β in the model to control the residual error variance as small as possible. For example, location and year effects are common in replicated experiments. These effects are not related to QTL but will contribute to the residual error if not included in the model. If there is no such a non-QTL effect to consider in a nice designed experiment, β will be a single parameter (intercept) and X_j will be unity across all $j = 1, \ldots, n$.

With interval mapping, the QTL genotype is never known unless the putative QTL position overlaps with a fully informative marker. Therefore,Haley and Knott (1992) suggested to replace the unknown Z_j by the expectation of Z_j conditional on flanking marker genotype. Let $p_j(1)$, $p_j(0)$ and $p_j(-1)$ be the conditional probabilities for the three genotypes given flanking marker information (see Chapter 4 for the method of calculating conditional probability). The LS model of Haley and Knott (1992) is

$$y_j = X_j\beta + U_j\gamma + e_j \tag{9.4}$$

where

$$U_j = E(Z_j) = p_j(+1)H_1 + p_j(0)H_2 + p_j(-1)H_3 \tag{9.5}$$

is the conditional expectation of Z_j. The residual error e_j (different from ε_j) remains normal with mean zero and variance σ^2, although this assumption has been violated (see next section). The least squares estimate of β and γ is

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} X_j^T X_j & \sum_{j=1}^{n} X_j^T U_j \\ \sum_{j=1}^{n} U_j^T X_j & \sum_{j=1}^{n} U_j^T U_j \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n} X_j^T y_j \\ \sum_{j=1}^{n} U_j^T y_j \end{bmatrix} \tag{9.6}$$

and the estimated residual error variance is

$$\hat{\sigma}^2 = \frac{1}{n-p-2} \sum_{j=1}^{n} (y_j - X_j\hat{\beta} - U_j\hat{\gamma})^2 \tag{9.7}$$

The variance-covariance matrix of the estimated parameters is

$$\operatorname{var} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} X_j^T X_j & \sum_{j=1}^{n} X_j^T U_j \\ \sum_{j=1}^{n} U_j^T X_j & \sum_{j=1}^{n} U_j^T U_j \end{bmatrix}^{-1} \hat{\sigma}^2 \tag{9.8}$$

which is a $(p+2) \times (p+2)$ matrix. Let

$$\operatorname{var}(\hat{\gamma}) = V = \begin{bmatrix} \operatorname{var}(\hat{a}) & \operatorname{cov}(\hat{a}, \hat{d}) \\ \operatorname{cov}(\hat{a}, \hat{d}) & \operatorname{var}(\hat{d}) \end{bmatrix} \tag{9.9}$$

be the 2×2 lower diagonal bock of matrix (9.8). The standard errors of the estimated additive and dominance effects are the square roots of the diagonal elements of matrix (9.9).

We can use either the F-test or the W-test statistic to test the hypothesis of $H_0 : \gamma = 0$. The W-test statistic is

$$W = \hat{\gamma}^T V^{-1} \hat{\gamma} = \begin{bmatrix} \hat{a} & \hat{d} \end{bmatrix} \begin{bmatrix} \operatorname{var}(\hat{a}) & \operatorname{cov}(\hat{a}, \hat{d}) \\ \operatorname{cov}(\hat{a}, \hat{d}) & \operatorname{var}(\hat{d}) \end{bmatrix}^{-1} \begin{bmatrix} \hat{a} \\ \hat{d} \end{bmatrix} \tag{9.10}$$

The likelihood ratio test statistic can also be applied if we assume that $e_j \sim N(0, \sigma^2)$ for all $j = 1, \ldots, n$. The log likelihood function for the full model is

$$\begin{aligned} L_1 &= -\frac{n}{2} \ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \sum_{j=1}^{n} (y - X_j\hat{\beta} - U_j\hat{\gamma})^2 \\ &\approx -\frac{n}{2} \left[\ln(\hat{\sigma}^2) + 1 \right] \end{aligned} \tag{9.11}$$

The reduced model under $H_0 : \gamma = 0$ is

$$\begin{aligned} L_0 &= -\frac{n}{2} \ln(\hat{\hat{\sigma}}^2) - \frac{1}{2\hat{\hat{\sigma}}^2} \sum_{j=1}^{n} (y - X_j\hat{\hat{\beta}})^2 \\ &\approx -\frac{n}{2} \left[\ln(\hat{\hat{\sigma}}^2) + 1 \right] \end{aligned} \tag{9.12}$$

where

$$\hat{\hat{\beta}} = \left[\sum_{j=1}^{n} X_j^T X_j \right]^{-1} \left[\sum_{j=1}^{n} X_j^T y_j \right] \tag{9.13}$$

and

$$\hat{\hat{\sigma}}^2 = \frac{1}{n-p} \sum_{j=1}^{n} (y_j - X_j \hat{\hat{\beta}})^2 \tag{9.14}$$

The likelihood ratio test statistic is

$$\lambda = -2(L_0 - L_1) \tag{9.15}$$

9.2 Weighted least squares

Xu (1995) realized that the LS method is flawed because the residual variance is heterogeneous after replacing X_j by its conditional expectation U_j. The conditional variance of X_j given marker information varies from one individual to another and it will contribute to the residual variance. Xu (1998a,b) modified the exact model

$$y_j = X_j \beta + Z_j \gamma + \varepsilon_j \tag{9.16}$$

by

$$y_j = X_j \beta + U_j \gamma + (Z_j - U_j)\gamma + \varepsilon_j \tag{9.17}$$

which differs from the Haley and Knott's (1992) model by $(Z_j - U_j)\gamma$. Since Z_j is not observable, this additional term is merged into the residual error if ignored. Let

$$e_j = (Z_j - U_j)\gamma + \varepsilon_j \tag{9.18}$$

be the new residual error. The Haley and Knott's (1992) model can be rewritten as

$$y_j = X_j \beta + U_j \gamma + e_j \tag{9.19}$$

Although we assume $\varepsilon_j \sim N(0, \sigma^2)$, this does not validate the normal assumption of e_j. The expectation for e_j is

$$E(e_j) = [E(Z_j) - U_j]\,\gamma + E(\varepsilon_j) = 0 \tag{9.20}$$

The variance of e_j is

$$\mathrm{var}(e_j) = \sigma_j^2 = \gamma^T \mathrm{var}(Z_j)\gamma + \sigma^2 = \left(\frac{1}{\sigma^2} \gamma^T \Sigma_j \gamma + 1 \right) \sigma^2 \tag{9.21}$$

where $\Sigma_j = \mathrm{var}(Z_j)$, which is defined as a conditional variance-covariance matrix given flanking marker information. The explicit forms of Σ_j is

$$\Sigma_j = E(Z_j^T Z_j) - E(Z_j^T)E(Z_j), \tag{9.22}$$

where

$$E(Z_j^T Z_j) = p_j(1)H_1^T H_1 + p_j(0)H_2^T H_2 + p_j(-1)H_3^T H_3 \tag{9.23}$$

and

$$E(Z_j) = U_j = p_j(1)H_1 + p_j(0)H_2 + p_j(-1)H_3. \tag{9.24}$$

Let

$$\sigma_j^2 = \left(\frac{1}{\sigma^2}\gamma^T \Sigma_j \gamma + 1\right)\sigma^2 = \frac{1}{W_j}\sigma^2 \tag{9.25}$$

where

$$W_j = \left(\frac{1}{\sigma^2}\gamma^T \Sigma_j \gamma + 1\right)^{-1} \tag{9.26}$$

is the weight variable for the jth individual. The weighted least squares estimate of the parameters are

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \sum\limits_{j=1}^n X_j^T W_j X_j & \sum\limits_{j=1}^n X_j^T W_j U_j \\ \sum\limits_{j=1}^n U_j^T W_j X_j & \sum\limits_{j=1}^n U_j^T W_j U_j \end{bmatrix}^{-1} \begin{bmatrix} \sum\limits_{j=1}^n X_j^T W_j y_j \\ \sum\limits_{j=1}^n U_j^T W_j y_j \end{bmatrix} \tag{9.27}$$

and

$$\hat{\sigma}^2 = \frac{1}{n-p-2}\sum_{j=1}^n W_j(y_j - X_j\hat{\beta} - U_j\hat{\gamma})^2 \tag{9.28}$$

Since W_j is a function of σ^2, iterations are required. The iteration process is demonstrated as below.

1. Initialize γ and σ^2
2. Update β and γ using equation 9.27
3. Update σ^2 using equation 9.28
4. Repeat Step 2 to Step 3 until a certain criterion of convergence is satisfied.

The iteration process is very fast, usually taking less than 5 iterations to converge. Since the weight is not a constant (it is a function of the parameters), repeatedly updating the weight is required. Therefore, the weighted least squares method is also called iteratively reweighted least squares (IRLS). The few cycle of iterations make the results of IRLS very close to that of the maximum likelihood method (to be introduce later). A nice property of the IRLS is that the variance-covariance matrix of the estimated parameters is automatically given as a by-product of the iteration process. This matrix is

$$\text{var}\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \sum\limits_{j=1}^n X_j^T W_j X_j & \sum\limits_{j=1}^n X_j^T W_j U_j \\ \sum\limits_{j=1}^n U_j^T W_j X_j & \sum\limits_{j=1}^n U_j^T W_j U_j \end{bmatrix}^{-1} \hat{\sigma}^2 \tag{9.29}$$

As a result, the F- or W-test statistic can be used for significance test. Like the least squares method, a likelihood ratio test statistic can also be established for significance test. The L_0 under the null model is the same as that described in the section of least squares method. The L_1 under the alternative model is

$$\begin{aligned} L_1 &= -\frac{n}{2}\ln(\hat{\sigma}^2) + \frac{1}{2}\sum_{j=1}^{n}\ln(W_j) - \frac{1}{2\hat{\sigma}^2}\sum_{j=1}^{n}W_j(y - X_j\hat{\beta} - U_j\hat{\gamma})^2 \\ &\approx -\frac{n}{2}\left[\ln(\hat{\sigma}^2) + 1\right] + \frac{1}{2}\sum_{j=1}^{n}\ln(W_j) \end{aligned} \tag{9.30}$$

9.3 Fisher scoring

The weighted least squares solution described in the previous section does not maximize the log likelihood function (9.30). We can prove that it actually maximizes equation (9.30) if W_j is treated as a constant. The fact that W_j is a function of parameters makes the above weighted least squares estimates suboptimal. The optimal solution should be obtained by maximizing equation (9.30) fully without assuming W_j being a constant.

Recall that the linear model for y_j is

$$y_j = X_j\beta + U_j\gamma + e_j \tag{9.31}$$

where the residual error $e_j = (Z_j - U_j)\gamma + \varepsilon_j$ has a zero mean and variance

$$\sigma_j^2 = \left(\frac{1}{\sigma^2}\gamma^T\Sigma_j\gamma + 1\right)\sigma^2 = \frac{1}{W_j}\sigma^2 \tag{9.32}$$

If we assume that $e_j \sim N(0, \sigma_j^2)$, we can construct the following log likelihood function,

$$L(\theta) = -\frac{n}{2}\ln(\sigma^2) + \frac{1}{2}\sum_{j=1}^{n}\ln(W_j) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}W_j(y - X_j\beta - U_j\gamma)^2 \tag{9.33}$$

where $\theta = \{\beta, \gamma, \sigma^2\}$ is the vector of parameters. The maximum likelihood solution for the above likelihood function is hard to obtain because W_j is not a constant but a function of the parameters. The Newton-Raphson algorithm may be adopted but it requires the second partial derivative of the log likelihood function with respect to the parameter, which is very complicated. In addition, the Newton-Raphson algorithm often misbehaves when the dimensionality of θ is high. We now introduce the Fisher scoring algorithm for finding the MLE of θ. The method requires the first partial derivative of $L(\theta)$ with respect to the parameters, called the score vector and denoted by $S(\theta)$, and the information matrix, denoted by $I(\theta)$. The score vector has the following form,

$$S(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} \sum\limits_{j=1}^{n} X_j^T W_j (y_j - \mu_j) \\ \frac{1}{\sigma^2} \sum\limits_{j=1}^{n} U_j^T W_j (y_j - \mu_j) - \frac{1}{\sigma^2} \sum\limits_{j=1}^{n} W_j \Sigma_j \gamma + \frac{1}{\sigma^4} \sum\limits_{j=1}^{n} (y_j - \mu_j)^2 W_j^2 \Sigma_j \gamma \\ \frac{1}{2\sigma^4} \sum\limits_{j=1}^{n} W_j^2 (y_j - \mu_j)^2 - \frac{1}{2\sigma^2} \sum\limits_{j=1}^{n} W_j \end{bmatrix} \tag{9.34}$$

where

$$\mu_j = X_j \beta + U_j \gamma \tag{9.35}$$

The information matrix is given below

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} \sum\limits_{j=1}^{n} X_j^T W_j X_j & \frac{1}{\sigma^2} \sum\limits_{j=1}^{n} X_j^T W_j U_j & 0 \\ \frac{1}{\sigma^2} \sum\limits_{j=1}^{n} U_j W_j X_j, & \frac{1}{\sigma^2} \sum\limits_{j=1}^{n} U_j^T W_j U_j + \frac{2}{\sigma^4} \sum\limits_{j=1}^{n} W_j^2 \Sigma_j \gamma \gamma^T \Sigma_j, & \frac{1}{\sigma^4} \sum\limits_{j=1}^{n} W_j^2 \Sigma_j \gamma \\ 0 & \frac{1}{\sigma^4} \sum\limits_{j=1}^{n} W_j^2 \gamma^T \Sigma_j & \frac{1}{2\sigma^4} \sum\limits_{j=1}^{n} W_j^2 \end{bmatrix} \tag{9.36}$$

The Fisher scoring algorithm is implemented using the following iteration equation,

$$\theta^{(t+1)} = \theta^{(t)} + I^{-1}(\theta^{(t)}) S(\theta^{(t)}) \tag{9.37}$$

where $\theta^{(t)}$ is the parameter value at iteration t and $\theta^{(t+1)}$ is the updated value. Once the iteration process converges, the variance-covariance matrix of the estimated parameters is automatically given, which is

$$\text{var}(\hat{\theta}) = I^{-1}(\hat{\theta}) \tag{9.38}$$

The detailed expression of this matrix is

$$\text{var} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \sum\limits_{j=1}^{n} X_j^T W_j X_j & \sum\limits_{j=1}^{n} X_j^T W_j U_j & 0 \\ \sum\limits_{j=1}^{n} U_j W_j X_j, & \sum\limits_{j=1}^{n} U_j^T W_j U_j + \frac{2}{\hat{\sigma}^2} \sum\limits_{j=1}^{n} W_j^2 \Sigma_j \hat{\gamma} \hat{\gamma}^T \Sigma_j, & \frac{1}{\sigma^2} \sum\limits_{j=1}^{n} W_j^2 \Sigma_j \hat{\gamma} \\ 0 & \frac{1}{\hat{\sigma}^2} \sum\limits_{j=1}^{n} W_j^2 \hat{\gamma}^T \Sigma_j & \frac{1}{2\hat{\sigma}^2} \sum\limits_{j=1}^{n} W_j^2 \end{bmatrix}^{-1} \hat{\sigma}^2 \tag{9.39}$$

which can be compared with the variance-covariance matrix of the iteratively reweighted least squares estimate given in the previous section (equation 9.29).

We now give the derivation of the score vector and the information matrix. We can write the log likelihood function as

$$L(\theta) = \sum_{j=1}^{n} L_j(\theta) \tag{9.40}$$

where

$$L_j(\theta) = -\frac{1}{2}\ln(\sigma^2) + \frac{1}{2}\ln W_j - \frac{1}{2\sigma^2}W_j(y_j - \mu_j)^2 \tag{9.41}$$

and

$$\mu_j = X_j\beta + U_j\gamma \tag{9.42}$$

The score vector is a vector of the first partial derivatives, as shown below

$$S(\theta) = \sum_{j=1}^{n} S_j(\theta) \tag{9.43}$$

where

$$S_j(\theta) = \begin{bmatrix} \frac{\partial}{\partial\beta}L_j(\theta) \\ \frac{\partial}{\partial\gamma}L_j(\theta) \\ \frac{\partial}{\partial\sigma^2}L_j(\theta) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}X_j^T W_j(y_j - \mu_j) \\ \frac{1}{\sigma^2}U_j^T W_j(y_j - \mu_j) - \frac{1}{\sigma^2}W_j\Sigma_j\gamma + \frac{1}{\sigma^4}(y_j - \mu_j)^2 W_j^2\Sigma_j\gamma \\ \frac{1}{2\sigma^4}W_j^2(y_j - \mu_j)^2 - \frac{1}{2\sigma^2}W_j \end{bmatrix} \tag{9.44}$$

Therefore, we only need to take the sum of the first partial derivatives across individuals to get the score vector. Note that when deriving $S_j(\theta)$ we need the following derivatives,

$$\frac{\partial W_j}{\partial\theta} = \begin{bmatrix} \frac{\partial W_j}{\partial\beta} \\ \frac{\partial W_j}{\partial\gamma} \\ \frac{\partial W_j}{\partial\sigma^2} \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{2}{\sigma^2}W_j^2\Sigma_j\gamma \\ \frac{1}{\sigma^2}W_j(1 - W_j) \end{bmatrix} \tag{9.45}$$

and

$$\frac{\partial\mu_j}{\partial\theta} = \begin{bmatrix} \frac{\partial\mu_j}{\partial\beta} \\ \frac{\partial\mu_j}{\partial\gamma} \\ \frac{\partial\mu_j}{\partial\sigma^2} \end{bmatrix} = \begin{bmatrix} X_j^T \\ U_j^T \\ 0 \end{bmatrix} \tag{9.46}$$

The information matrix is

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta) = \sum_{j=1}^{n} -E[H_j(\theta)] \tag{9.47}$$

where

$$H_j(\theta) = \frac{\partial^2 L_j(\theta)}{\partial\theta\partial\theta^T} = \begin{bmatrix} \frac{\partial^2 L_j(\theta)}{\partial\beta\partial\beta^T} & \frac{\partial^2 L_j(\theta)}{\partial\beta\partial\gamma^T} & \frac{\partial^2 L_j(\theta)}{\partial\beta\partial\sigma^2} \\ \frac{\partial^2 L_j(\theta)}{\partial\gamma\partial\beta^T} & \frac{\partial^2 L_j(\theta)}{\partial\gamma\partial\gamma^T} & \frac{\partial^2 L_j(\theta)}{\partial\gamma\partial\sigma^2} \\ \frac{\partial^2 L_j(\theta)}{\partial\sigma^2\partial\beta^T} & \frac{\partial^2 L_j(\theta)}{\partial\sigma^2\partial\gamma^T} & \frac{\partial^2 L_j(\theta)}{\partial\sigma^2\partial\sigma^2} \end{bmatrix} \tag{9.48}$$

is the second partial derivative of $L_j(\theta)$ with respect to the parameters and called the Hessian matrix. Derivation of this matrix is very tedious, but the negative expectation of the Hessian matrix is identical to the expectation of the product of the score vector (Wedderburn, 1974),

$$-E[H_j(\theta)] = E[S_j(\theta)S_j^T(\theta)] \tag{9.49}$$

Using this identity, we can avoid the Hessian matrix. Therefore, the information matrix is

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta) = \sum_{j=1}^{n} E[S_j(\theta)S_j^T(\theta)] \tag{9.50}$$

where

$$E[S_j(\theta)S_j^T(\theta)] = \begin{bmatrix} E\left(\frac{\partial L_j(\theta)}{\partial\beta}\frac{\partial L_j(\theta)}{\partial\beta^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial\beta}\frac{\partial L_j(\theta)}{\partial\gamma^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial\beta}\frac{\partial L_j(\theta)}{\partial\sigma^2}\right) \\ E\left(\frac{\partial L_j(\theta)}{\partial\gamma}\frac{\partial L_j(\theta)}{\partial\beta^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial\gamma}\frac{\partial L_j(\theta)}{\partial\gamma^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial\gamma}\frac{\partial L_j(\theta)}{\partial\sigma^2}\right) \\ E\left(\frac{\partial L_j(\theta)}{\partial\sigma^2}\frac{\partial L_j(\theta)}{\partial\beta^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial\sigma^2}\frac{\partial L_j(\theta)}{\partial\gamma^T}\right) & E\left(\frac{\partial L_j(\theta)}{\partial\sigma^2}\frac{\partial L_j(\theta)}{\partial\sigma^2}\right) \end{bmatrix} \tag{9.51}$$

Note that the expectation is taken with respect to the phenotypic value y_j . In other words, after taking the expectation, variable y_j will disappear from the expressions. There are six different blocks in the above matrix. We will only provide the derivation for one block as an example. The derivations of the remaining five blocks are left to students for practice. The result can be found in Han and Xu (2008). We now show the derivation of the first block of the matrix. The product (before taking the expectation) is

$$\begin{aligned}\frac{\partial L_j(\theta)}{\partial\beta}\frac{\partial L_j(\theta)}{\partial\beta^T} &= \left[\frac{1}{\sigma^2}X_j^T W_j(y_j-\mu_j)\right]\left[\frac{1}{\sigma^2}X_j^T W_j(y_j-\mu_j)\right]^T \\ &= \frac{1}{\sigma^4}X_j^T W_j^2 X_j^T (y_j-\mu_j)^2\end{aligned} \tag{9.52}$$

The expectation of it is

$$\begin{aligned}E\left(\frac{\partial L_j(\theta)}{\partial\beta}\frac{\partial L_j(\theta)}{\partial\beta^T}\right) &= \frac{1}{\sigma^4}X_j^T W_j^2 X_j^T E\left[(y_j-\mu_j)^2\right] \\ &= \frac{1}{\sigma^2}X_j^T W_j X_j^T\end{aligned} \tag{9.53}$$

The second line of the above equation requires the following identity,

$$E\left[(y_j-\mu_j)^2\right] = \frac{1}{W_j}\sigma^2 \tag{9.54}$$

Taking the sum of equation (9.53) across individuals, we get

$$I_{11}(\theta) = \frac{1}{\sigma^2}\sum_{j=1}^{n} X_j^T W_j X_j \tag{9.55}$$

which is the first block of the information matrix. When deriving the expectations for the remaining five blocks, we need the following expectations,

$$E[(y_j-\mu_j)^k] = \begin{cases} 0 & \text{for odd } k \\ W_j^{-1}\sigma^2 & \text{for } k=2 \\ 3W_j^{-2}\sigma^4 & \text{for } k=4 \end{cases} \tag{9.56}$$

The above expectations requires the assumption of $y_j \sim N(\mu_j, \sigma_j^2)$ where $\sigma_j^2 = W_j^{-1}\sigma^2$.

9.4 Maximum likelihood method

The maximum likelihood method (Lander and Botstein, 1989) is the optimal one compared to all other methods described in this chapter. The linear model for the phenotypic value of y_j is

$$y_j = X_j\beta + Z_j\gamma + \varepsilon_j \tag{9.57}$$

where $\varepsilon_j \sim N(0, \sigma^2)$ is assumed. The genotype indicator variable Z_j is a missing value because we cannot observe the genotype of a putative QTL. Rather than replacing Z_j by U_j as done in the least squares and the weighted least squares methods, the maximum likelihood method takes into consideration the mixture distribution of y_j. We have learned the mixture distribution in Chapter 7 when we deal with segregation analysis of quantitative traits. We now extend the mixture model to interval mapping. When the genotype of the putative QTL is observed, the probability density of y_j is

$$\begin{aligned} f_k(y_j) &= \Pr(y_j|Z_j = H_k) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_j - X_j\beta + H_k\gamma)^2\right] \end{aligned} \tag{9.58}$$

When flanking marker information is used, the conditional probability that $Z_j = H_k$ is

$$p_j(k) = \Pr(Z_j = H_k), \forall k = 1, 2, 3 \tag{9.59}$$

for the three genotypes, A_1A_1, A_1A_2 and A_2A_2. These probabilities are different from the Mendelian segregation ratio $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ as described in the segregation analysis. They are the conditional probabilities given marker information and thus vary from one individual to another because different individuals may have different marker genotypes. Using the conditional probabilities as weights, we get the mixture distribution

$$f(y_j) = \sum_{k=1}^{3} p_j(2-k) f_k(y_j) \tag{9.60}$$

where

$$p_j(2-k) = \begin{cases} p_j(-1) & \text{for } k = 1 \\ p_j(0) & \text{for } k = 2 \\ p_j(+1) & \text{for } k = 3 \end{cases} \tag{9.61}$$

is a special notation for the conditional probability and should not be interpreted as p_j times $(2-k)$. The log likelihood function is

$$L(\theta) = \sum_{j=1}^{n} L_j(\theta) \tag{9.62}$$

where $L_j(\theta) = \ln f(y_j)$.

9.4.1 EM algorithm

The MLE of θ can be obtained using any numerical algorithms but the EM algorithm is generally more preferable than others because we can take advantage of the mixture distribution. Derivation of the EM algorithm has been given in Chapter 7 when segregation analysis was introduced. Here we simply give the result of the EM algorithm. Assuming that the genotypes of all individuals are observed, the maximum likelihood estimates of parameters would be

$$\begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} X_j^T X_j & \sum_{j=1}^{n} X_j^T Z_j \\ \sum_{j=1}^{n} Z_j^T X_j & \sum_{j=1}^{n} Z_j^T Z_j \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n} X_j^T y_j \\ \sum_{j=1}^{n} Z_j^T y_j \end{bmatrix} \tag{9.63}$$

and

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^{n} (y_j - X_j\beta - Z_j\gamma)^2 \tag{9.64}$$

The EM algorithm takes advantage of the above explicit solutions of the parameters by substituting all entities containing the missing value Z_j by their posterior expectations, i.e.,

$$\begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} X_j^T X_j & \sum_{j=1}^{n} X_j^T E(Z_j) \\ \sum_{j=1}^{n} E(Z_j^T) X_j & \sum_{j=1}^{n} E(Z_j^T Z_j) \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n} X_j^T y_j \\ \sum_{j=1}^{n} E(Z_j^T) y_j \end{bmatrix} \tag{9.65}$$

and

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^{n} E\left[(y_j - X_j\beta - Z_j\gamma)^2\right] \tag{9.66}$$

where the expectations are taken using the posterior probabilities of QTL genotypes, which is defined as

$$p_j^*(2-k) = \frac{p_j(2-k) f_k(y_j)}{\sum_{k'=1}^{3} p_j(2-k') f_{k'}(y_j)}, \forall k = 1, 2, 3 \tag{9.67}$$

The posterior expectations are

$$\begin{aligned} E(Z_j) &= \sum_{k=1}^{3} p_j^*(2-k)H_k \\ E(Z_j^T Z_j) &= \sum_{k=1}^{3} p_j^*(2-k)H_k^T H_k \\ E\left[(y_j - X_j\beta - Z_j\gamma)^2\right] &= \sum_{k=1}^{3} p_j^*(2-k)(y_j - X_j\beta - H_k\gamma)^2 \end{aligned} \tag{9.68}$$

Since $f_k(y_j)$ is a function of parameters and thus $p_j^*(2-k)$ is also a function of the parameters. However, the parameters are unknown and they are the very quantities we want to find out. Therefore, iterations are required. Here is the iteration process,

1. Initialize $\theta = \theta^{(t)}$ for $t = 0$
2. Calculate the posterior expectations using equations (9.67) and (9.68)
3. Update parameters using equations (9.65) and (9.66)
4. Increment t by 1 and repeat Step 2 to Step 3 until a certain criterion of convergence is satisfied.

Once the iteration converges, the MLE of the parameters is $\hat{\theta} = \theta^{(t)}$, where t is the number of iterations required for convergence.

9.4.2 Variance-covariance matrix of $\hat{\boldsymbol{\theta}}$

Unlike the weighted least squares and the Fisher scoring algorithms where the variance-covariance matrix of the estimated parameters is automatically given as a by-product of the iteration process, the EM algorithm requires an additional step to calculate this matrix. The method was developed by Louis (1982) and it requires the score vectors and the Hessian matrix for the complete-data log likelihood function rather than the actual observed log likelihood function. The complete-data log likelihood function is the log likelihood function as if Z_j were observed, which is

$$L(\theta, Z) = \sum_{j=1}^{n} L_j(\theta, Z) \tag{9.69}$$

where

$$L_j(\theta, Z) = -\frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y_j - X_j\beta - Z_j\gamma)^2 \tag{9.70}$$

The score vector is

$$S(\theta, Z) = \sum_{j=1}^{n} S_j(\theta, Z) \tag{9.71}$$

where

$$S_j(\theta, Z) = \begin{bmatrix} \frac{\partial}{\partial \beta} L_j(\theta, Z) \\ \frac{\partial}{\partial \gamma} L_j(\theta, Z) \\ \frac{\partial}{\partial \sigma^2} L_j(\theta, Z) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} X_j^T (y_j - X_j \beta - Z_j \gamma) \\ \frac{1}{\sigma^2} Z_j^T (y_j - X_j \beta - Z_j \gamma) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} (y_j - X_j \beta - Z_j \gamma)^2 \end{bmatrix} \tag{9.72}$$

The second partial derivative (Hessian matrix) is

$$H(\theta, Z) = \sum_{j=1}^{n} H_j(\theta, Z) \tag{9.73}$$

where

$$H_j(\theta, Z) = \begin{bmatrix} \frac{\partial^2 L_j(\theta,Z)}{\partial \beta \partial \beta^T} & \frac{\partial^2 L_j(\theta,Z)}{\partial \beta \partial \gamma^T} & \frac{\partial^2 L_j(\theta,Z)}{\partial \beta \partial \sigma^2} \\ \frac{\partial^2 L_j(\theta,Z)}{\partial \gamma \partial \beta^T} & \frac{\partial^2 L_j(\theta,Z)}{\partial \gamma \partial \gamma^T} & \frac{\partial^2 L_j(\theta,Z)}{\partial \gamma \partial \sigma^2} \\ \frac{L_j(\theta,Z)}{\partial \sigma^2 \partial \beta^T} & \frac{L_j(\theta,Z)}{\partial \sigma^2 \partial \gamma^T} & \frac{L_j(\theta,Z)}{\partial \sigma^2 \partial \sigma^2} \end{bmatrix} \tag{9.74}$$

The six different blocks of the above matrix are

$$\begin{aligned} \frac{\partial^2 L_j(\theta)}{\partial \beta \partial \beta^T} &= -\frac{1}{\sigma^2} X_j^T X_j \\ \frac{\partial^2 L_j(\theta)}{\partial \beta \partial \gamma^T} &= -\frac{1}{\sigma^2} X_j^T Z_j \\ \frac{\partial^2 L_j(\theta)}{\partial \beta \partial \sigma^2} &= -\frac{1}{\sigma^4} X_j^T (y_j - X_j \beta - Z_j \gamma) \\ \frac{\partial^2 L_j(\theta)}{\partial \gamma \partial \gamma^T} &= -\frac{1}{\sigma^2} Z_j^T Z_j \\ \frac{\partial^2 L_j(\theta)}{\partial \gamma \partial \sigma^2} &= -\frac{1}{\sigma^4} Z_j^T (y_j - X_j \beta - Z_j \gamma) \\ \frac{\partial^2 L_j(\theta)}{\partial \sigma^2 \partial \sigma^2} &= \frac{1}{2\sigma^4} - \frac{1}{\sigma^6} (y_j - X_j \beta - Z_j \gamma)^2 \end{aligned} \tag{9.75}$$

We now have the score vector and the Hessian matrix available for the complete-data log likelihood function. The Louis information matrix is

$$I(\theta) = -E\left[H(\theta, Z)\right] - E\left[S(\theta, Z) S^T(\theta, Z)\right] \tag{9.76}$$

where the expectations are taken with respect the missing value (Z_j) using the posterior probabilities of QTL genotypes. At the MLE of parameters, $E\left[S(\hat{\theta}, Z)\right] = 0$. Therefore,

$$\begin{aligned} E\left[S(\theta, Z) S^T(\theta, Z)\right] &= \mathrm{var}\left[S(\theta, Z)\right] + E\left[S(\theta, Z)\right] E\left[S^T(\theta, Z)\right] \\ &= \mathrm{var}\left[S(\theta, Z)\right] \end{aligned} \tag{9.77}$$

As a result, an alternative expression of the Louis information matrix is

$$
\begin{aligned}
I(\theta) &= -E\left[H(\theta, Z)\right] - \operatorname{var}\left[S(\theta, Z)\right] \\
&= -\sum_{j=1}^{n} E\left[H_j(\theta, Z)\right] - \sum_{j=1}^{n} \operatorname{var}\left[S_j(\theta, Z)\right] \qquad (9.78) \\
& \qquad (9.79)
\end{aligned}
$$

The expectations are

$$
E\left[H_j(\theta, Z)\right] = \begin{bmatrix} E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\beta\partial\beta^T}\right) & E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\beta\partial\gamma^T}\right) & E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\beta\partial\sigma^2}\right) \\ E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\gamma\partial\beta^T}\right) & E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\gamma\partial\gamma^T}\right) & E\left(\frac{\partial^2 L_j(\theta,Z)}{\partial\gamma\partial\sigma^2}\right) \\ E\left(\frac{L_j(\theta,Z)}{\partial\sigma^2\partial\beta^T}\right) & E\left(\frac{L_j(\theta,Z)}{\partial\sigma^2\partial\gamma^T}\right) & E\left(\frac{L_j(\theta,Z)}{\partial\sigma^2\partial\sigma^2}\right) \end{bmatrix} \qquad (9.80)
$$

The six different blocks of the above matrix are

$$
\begin{aligned}
E\left(\frac{\partial^2 L_j(\theta)}{\partial\beta\partial\beta^T}\right) &= -\frac{1}{\sigma^2} X_j^T X_j \\
E\left(\frac{\partial^2 L_j(\theta)}{\partial\beta\partial\gamma^T}\right) &= -\frac{1}{\sigma^2} X_j^T E(Z_j) \\
E\left(\frac{\partial^2 L_j(\theta)}{\partial\beta\partial\sigma^2}\right) &= -\frac{1}{\sigma^4} X_j^T \left[y_j - X_j\beta - E(Z_j)\gamma\right] \\
E\left(\frac{\partial^2 L_j(\theta)}{\partial\gamma\partial\gamma^T}\right) &= -\frac{1}{\sigma^2} E(Z_j^T Z_j) \\
E\left(\frac{\partial^2 L_j(\theta)}{\partial\gamma\partial\sigma^2}\right) &= -\frac{1}{\sigma^4} E\left[Z_j^T(y_j - X_j\beta - Z_j\gamma)\right] \\
E\left(\frac{\partial^2 L_j(\theta)}{\partial\sigma^2\partial\sigma^2}\right) &= \frac{1}{2\sigma^4} - \frac{1}{\sigma^6} E\left[(y_j - X_j\beta - Z_j\gamma)^2\right] \qquad (9.81)
\end{aligned}
$$

Again, all the expectations are taken with respect to the missing value Z_j, not the observed phenotype y_j. This is very different from the information matrix of the Fisher scoring algorithm. The variance-covariance matrix of the score vector is

$$
\operatorname{var}\left[S(\theta, Z)\right] = \sum_{j=1}^{n} \operatorname{var}\left[S_j(\theta, Z)\right] \qquad (9.82)
$$

where var$[S_j(\theta, Z)]$ is a symmetric matrix as shown below,

$$
\begin{bmatrix} \operatorname{var}\left(\frac{\partial L_j(\theta,Z)}{\partial\beta}\right) & \operatorname{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\beta}, \frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right) & \operatorname{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\beta}, \frac{\partial L_j(\theta,Z)}{\partial\sigma^2}\right) \\ \operatorname{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\gamma}, \frac{\partial L_j(\theta,Z)}{\partial\beta^T}\right) & \operatorname{var}\left(\frac{\partial L_j(\theta,Z)}{\partial\gamma}\right) & \operatorname{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\gamma}, \frac{\partial L_j(\theta,Z)}{\partial\sigma^2}\right) \\ \operatorname{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\sigma^2}, \frac{\partial L_j(\theta,Z)}{\partial\beta^T}\right) & \operatorname{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\sigma^2}, \frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right) & \operatorname{var}\left(\frac{\partial L_j(\theta,Z)}{\partial\sigma^2}\right) \end{bmatrix} \qquad (9.83)
$$

The variances are calculated with respect to the missing value Z_j using the posterior probabilities of QTL genotypes. We only provide the detailed expression of one block of the above matrix. The remaining blocks are left to students for practice. The block that is used as an example is the (1,2) block.

$$\operatorname{cov}\left(\frac{\partial L_j(\theta,Z)}{\partial\beta},\frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right)= \\ E\left[\frac{\partial L_j(\theta,Z)}{\partial\beta}\frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right]-E\left[\frac{\partial L_j(\theta,Z)}{\partial\beta}\right]E\left[\frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right] \quad (9.84)$$

where

$$\begin{aligned} E\left[\frac{\partial L_j(\theta,Z)}{\partial\beta}\right] &= \frac{1}{\sigma^2}X_j^T\left[y_j-X_j\beta-E(Z_j)\gamma\right] \\ E\left[\frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right] &= \frac{1}{\sigma^2}E\left[(y_j-X_j\beta-Z_j\gamma)Z_j\right] \\ E\left[\frac{\partial L_j(\theta,Z)}{\partial\beta}\frac{\partial L_j(\theta,Z)}{\partial\gamma^T}\right] &= \frac{1}{\sigma^4}E\left[X_j^T(y_j-X_j\beta-Z_j\gamma)^2Z_j\right] \end{aligned} \quad (9.85)$$

We already learned how to calculate $E(Z_j)$ using the posterior probability of QTL genotype. The other expectations are

$$\begin{aligned} E\left[(y_j-X_j\beta-Z_j\gamma)Z_j\right] &= \sum_{k=1}^{3}p_j^*(2-k)(y_j-X_j\beta-H_k\gamma)H_k \\ E\left[X_j^T(y_j-X_j\beta-Z_j\gamma)^2Z_j^T\right] &= \sum_{k=1}^{3}p_j^*(2-k)X_j^T(y_j-X_j\beta-H_k\gamma)^2H_k^T \end{aligned} \quad (9.86)$$

When calculating the information matrix, the parameter θ is substituted by $\hat{\theta}$, the MLE of θ . Therefore, the observed information matrix is

$$I(\hat{\theta})=-E\left[H(\hat{\theta},Z)\right]-\operatorname{var}\left[S(\hat{\theta},Z)\right] \quad (9.87)$$

and the variance-covariance matrix of the estimated parameters is $\operatorname{var}(\hat{\theta})=I^{-1}(\hat{\theta})$.

9.4.3 Hypothesis test

The hypothesis that $H_0:\gamma=0$ can be tested using several different ways. If $\operatorname{var}(\hat{\theta})$ is already calculated, we can use the F- or W-test statistic, which requires $\operatorname{var}(\hat{\gamma})$, the variance-covariance matrix of the estimated QTL effects. It is a submatrix of $\operatorname{var}(\hat{\theta})$. The W-test statistic is

$$W=\hat{\gamma}^T\operatorname{var}^{-1}(\hat{\gamma})\hat{\gamma} \quad (9.88)$$

Alternatively, the likelihood ratio test statistic can be applied to test H_0. We have presented two log likelihood functions, one is the complete-data log likelihood function, denoted by $L(\theta,Z)$, and the other is the observed log likelihood function, denoted by $L(\theta)$. The log likelihood function used to construct the likelihood ratio test statistic is $L(\theta)$, not $L(\theta, Z)$. This complete-data log likelihood function, $L(\theta, Z)$, is only used to derive the EM algorithm and the observed information matrix. The likelihood ratio test statistic is

$$\lambda = -2(L_0 - L_1)$$

where $L_1 = L(\hat{\theta})$ is the observed log likelihood function evaluated at $\hat{\theta} = \{\hat{\beta}, \hat{\gamma}, \hat{\sigma}^2\}$ and L_0 is the log likelihood function evaluated at $\hat{\hat{\theta}} = \{\hat{\hat{\beta}}, 0, \hat{\hat{\sigma}}^2\}$ under the restricted model. The estimated parameter $\hat{\hat{\theta}}$ under the restricted model and L_0 are the same as those given in the section of the least squares method.

9.5 Remarks on the four methods of interval mapping

The LS method (Haley and Knott, 1992) is an approximation of the ML method, aiming to improve the computational speed. The method has been extended substantially to many other situations, e.g., multiple trait QTL mapping (Knott and Haley, 2000) and QTL mapping for binary traits (Visscher et al., 1996). When used for binary and other non-normal traits, the method is no longer called LS. Because of the fast speed, the method remains a popular

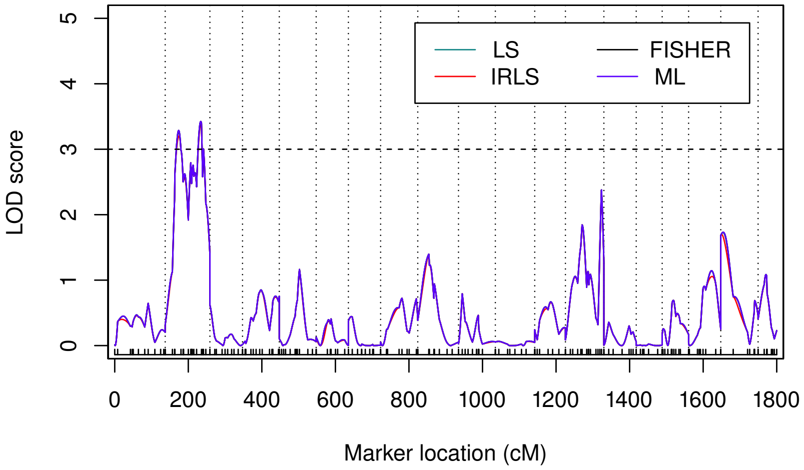

Fig. 9.2. The LOD test statistic profiles for four methods of interval mapping (LS - least square, IRLS - iteratively reweighted least square, FISHER - Fisher scoring, and ML - maximum likelihood). The mouse data were obtained from Lan et al. (2006). The trait investigated is the 10th week body weight. The 19 chromosomes (excluding the sex chromosome) are separated by the vertical dotted lines. The unevenly distributed black ticks on the horizontal axis indicate the marker locations.

method, even though the computer power has increased by many orders of magnitude since the LS was developed. In some literature (e.g., Feenstra et al. (2006)), the LS method is also called the H-K method in honor of the authors, Haley and Knott (1992). Xu (1995) noticed that the LS method, although a good approximation to ML in terms of estimates of QTL effects and test statistic, may lead to a biased (inflated) estimate for the residual error variance. Based on this work, Xu (1998a,b) eventually developed the iteratively reweighted least squares (IRLS) method. In these works (Xu, 1998a,b), the iteratively reweighted least squares was abbreviated IRWLS. Xu (1998b) compared LS, IRLS and ML in a variety of situations and conclude that IRLS is always better than LS and as efficient as ML. When the residual error does not have a normal distribution, which is required by the ML method, LS and IRLS can be better than ML. In other words, LS and IRLS are more robust than ML to the departure from normality. Kao (2000) and Feenstra et al. (2006) conducted more comprehensive investigation on LS, IRLS and ML and found that when epistatic effects exist, LS can generate unsatisfactory results, but IRLS and ML usually map QTL better than LS. In addition, Feenstra et al. (2006) modified the weighted least square method by using the estimating equations (EE) algorithm. This algorithm further improved the efficiency of the weighted least squares by maximizing an approximate likelihood function. Most recently, Han and Xu (2008) developed a Fisher scoring (FISHER) algorithm to maximize the approximate likelihood function. Both the EE and Fisher algorithm maximize the same likelihood function, and thus they produce identical results.

The LS method ignores the uncertainty of the QTL genotype. The IRLS, FISHER (or EE) and ML methods use different ways to extract information from the uncertainty of QTL genotype. If the putative location of QTL overlaps with a fully informative marker, all four methods produce identical result. Therefore, if the marker density is sufficiently high, there is virtually no difference for the four methods. For low marker density, when the putative position is far away from either flanking marker, the four methods will show some difference. This difference will be magnified by large QTL. Han and Xu (2008) compared the four methods in a simulation experiment and showed that when the putative QTL position is fixed in the middle of a 10 cM interval, the four methods generated almost identical results. However, when the interval expands to 20 cM, the differences among the four methods become noticeable.

Interval mapping with a 1 cM increment for the mouse 10th week body weight data were conducted using all the four methods by Han and Xu (2008). The LOD test statistic profiles are shown in Figure 9.2 for the four methods of interval mapping (LS, IRLS, FISHER and ML). There is virtually no differences for the four methods. The difference in LOD profiles is noticeable when the marker density is low. Comparisons for the estimated QTL effects were also conducted for the mouse data. Figure 9.3 shows the estimated QTL effect profiles along the genome for the four methods. Again the difference is barely noticeable.

A final remark on interval mapping is the way to infer the QTL genotype using flanking markers. If only flanking markers are used to infer the genotype of a putative position bracketed by the two markers, the method is called interval mapping. Strictly speaking, interval mapping only applies to fully informative markers because we always use flanking markers to infer the QTL genotype. However, almost all datasets obtained from real life experiments contain missing, uninformative or partially informative markers. To extract maximum information from markers, people always use the multipoint method (Jiang and Zeng, 1997) to infer a QTL genotype. The multipoint method uses more markers or even all markers of the entire chromosome (not just flanking markers) to infer the genotype of a putative position. With the multipoint analysis, we no longer have the notion of interval, and thus interval mapping is no longer an appropriate phrase to describe QTL mapping. Unfortunately, a more appropriate phrase has not been proposed and people are used to the phrase of interval mapping. Therefore, the so called interval mapping in the current literature means QTL mapping under a single QTL model, regardless whether the genotype of a putative QTL position is inferred from flanking markers or all markers.

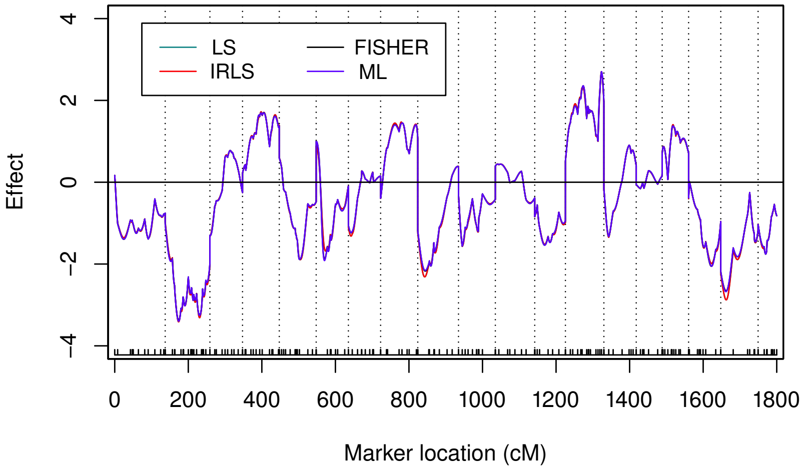

Fig. 9.3. The QTL effect profiles for four methods of interval mapping (LS - least square, IRLS - iteratively reweighted least square, FISHER - Fisher scoring, and ML - maximum likelihood). The mouse data were obtained from Lan et al. (2006). The trait investigated is the 10th week body weight. The 19 chromosomes (excluding the sex chromosome) are separated by the vertical dotted lines. The unevenly distributed black ticks on the horizontal axis indicate the marker locations.