Maximum Likelihood Estimation and Likelihood-ratio Tests

The method of maximum likelihood (ML), introduced by Fisher (1921), is widely used in human and quantitative genetics and we draw upon this approach throughout the book, especially in Chapters 13–16 (mixture distributions) and 26–27 (variance component estimation). Weir (1996) gives a useful introduction with genetic applications, while Kendall and Stuart (1979) and Edwards (1992) provide more detailed treatments.

LIKELIHOOD, SUPPORT, AND SCORE FUNCTIONS

The basic idea underlying ML is quite simple. Usually, when specifying a probability density function (say, a normal with unknown mean μ and unit variance), we treat the pdf as a function of z (the value of the random variable) with the distribution parameters $\boldsymbol{\Theta}$ assumed to be known. (While much of our discussion is in terms of a vector $\boldsymbol{\Theta}$, we use θ to indicate results for a single parameter.) With maximum likelihood estimation, we reverse the roles of the observed value and the distribution parameters by asking: Given a vector of observations $\mathbf{z}$, what can we say about $\boldsymbol{\Theta}$? To specify this alternative interpretation, the density function is denoted as $\ell(\boldsymbol{\Theta} \mid \mathbf{z})$, the **likelihood** of $\boldsymbol{\Theta}$ given the observed vector of data $\mathbf{z}$. This defines a **likelihood surface**, as $\ell(\boldsymbol{\Theta} \mid \mathbf{z})$ assigns a value to each possible point in the $\boldsymbol{\Theta}$-parameter space given the observed data $\mathbf{z}$. The **maximum likelihood estimate** (MLE) of the unknown parameters, $\widehat{\boldsymbol{\Theta}}$, is the value of $\boldsymbol{\Theta}$ corresponding to the maximum of $\ell(\boldsymbol{\Theta} \mid \mathbf{z})$, i.e., the MLE is the value of $\boldsymbol{\Theta}$ that is "most likely" to have produced the data $\mathbf{z}$. It is usually easier to find the maximum of a likelihood function by first taking its log and working with the resulting **log-likelihood**

$$L(\boldsymbol{\Theta} \mid \mathbf{z}) = \ln [\, \ell(\boldsymbol{\Theta} \mid \mathbf{z}) \,] \tag{A4.1}$$

L is also referred to as the **support**. Since the natural log is a monotonic function, $\ell(\boldsymbol{\Theta})$ has the same maxima as $\ln\big[\ell(\boldsymbol{\Theta})\big]$, so that the maximum of L also corresponds to the maximum of the likelihood function. The **score** S of a likelihood function is the first derivative of L with respect to the likelihood parameters, with $S(\theta) = \partial L(\theta)/\partial\theta$ for a single parameter likelihood function, and

$$\mathbf{S}(\boldsymbol{\Theta}) = \frac{\partial L(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = \begin{pmatrix} \partial L(\boldsymbol{\Theta})/\partial \boldsymbol{\Theta}_1 \\ \vdots \\ \partial L(\boldsymbol{\Theta})/\partial \boldsymbol{\Theta}_n \end{pmatrix} \tag{A4.2}$$

for a vector of n parameters. From elementary calculus it follows that the score evaluated at the MLE is zero, $\mathbf{S}(\widehat{\boldsymbol{\Theta}}) = \mathbf{0}$. This provides one approach for obtaining MLEs.

Example 1. Suppose n values, $z_1 \cdots z_n$, are sampled independently from an underlying normal with unknown mean μ and unit variance ($\sigma^2 = 1$). Letting $\mathbf{z} = (z_1, z_2, \cdots, z_n)$, what is the MLE for μ given $\mathbf{z}$? Since the observations are independent, the resulting probability density function for $\mathbf{z}$ is the product of n normal density functions,

$$\begin{aligned} p(\mathbf{z}, \mu) &= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\bigl[-(z_i - \mu)^2/2\bigr] \\ &= (2\pi)^{-n/2} \exp\bigl[-\sum_{i=1}^{n} (z_i - \mu)^2/2\bigr] \end{aligned} \tag{A4.3}$$

The log-likelihood (or support) becomes

$$L(\mu \,|\, \mathbf{z}) = \ln\bigl[\ell(\mu \,|\, \mathbf{z})\bigr] = -\left(\frac{n}{2}\right) \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (z_i - \mu)^2 \tag{A4.4}$$

which has the score function

$$S(\mu) = \frac{\partial\, L(\mu \,|\, \mathbf{z})}{\partial \mu} = \sum_{i=1}^{n} (\, z_i - \mu\,) = n(\overline{z} - \mu) \tag{A4.5}$$

Setting the score equal to zero and solving gives the MLE, $\widehat{\mu} = \overline{z}$.

Large-sample Properties of MLEs

MLEs have several important features when the sample size is large:

1. **Consistency**: As the sample size increases, the MLE converges to the true parameter value, e.g., $\widehat{\boldsymbol{\Theta}} \to \boldsymbol{\Theta}$.
2. **Invariance**: If $f(\boldsymbol{\Theta})$ is a function of the unknown parameters of the distribution, then the MLE of $f(\boldsymbol{\Theta})$ is $f(\widehat{\boldsymbol{\Theta}})$, i.e., the MLE of a function of the parameters is simply that function evaluated at the MLE. For example, the MLE of $\sqrt{\theta} = (\,\widehat{\theta}\,)^{1/2}$.
3. **Asymptotic normality and efficiency**: As the sample size increases, the sampling distribution of the MLE converges to a normal and (generally)

no other estimation procedure has a smaller variance. Hence, for sufficiently large sample sizes, estimates obtained via maximum likelihood typically have the smallest confidence intervals.

4. **Variance**: For large sample sizes, the variance of an MLE (assuming a single unknown parameter) is approximately the negative of the reciprocal of the second derivative of the log-likelihood function evaluated at the MLE $\widehat{\theta}$,

$$\sigma^2\left(\widehat{\theta}\,\right) \simeq -\left(\frac{\partial^2\, L(\theta \,|\, \mathbf{z})}{\partial\, \theta^2}\,\bigg|_{\theta=\widehat{\theta}} \right)^{-1} \tag{A4.6}$$

This is just the reciprocal of the curvature of the log-likelihood surface at the MLE. The flatter the likelihood surface around its maximum value (the MLE), the larger the variance; the steeper the surface, the smaller the variance. The minus sign appears because the second derivative is negative (downward curvature) at the maximum of the likelihood function.

Example 2. What is the large-sample variance of the MLE for μ from Example 1?

$$\frac{\partial^2\, L(\mu \,|\, \mathbf{z})}{\partial \mu^2} = \frac{\partial S(\mu \,|\, \mathbf{z})}{\partial \mu} = \frac{\partial \left(\sum_{i=1}^{n} (z_i - \mu) \right)}{\partial \mu} = -n$$

Applying Equation A4.6,

$$\sigma^2\left(\,\widehat{\mu}\,\right) \simeq \frac{1}{n}$$

Using the asymptotic normality of MLEs, the approximate distribution of the MLE is $\widehat{\mu} \sim \mathrm{N}(\mu, n^{-1})$, and the resulting 95 percent confidence interval for μ is $\widehat{\mu} \pm 1.96/\sqrt{n}$.

The Fisher Information Matrix

When estimating a vector of parameters, Equation A4.6 can be generalized by using the **Hessian matrix**, **H**, the matrix of second partials of the log-likelihood, whose ijth element is given by

$$\mathbf{H}_{ij} = \frac{\partial^2 L(\boldsymbol{\Theta} \,|\, \mathbf{z})}{\partial \boldsymbol{\Theta}_i \, \partial \boldsymbol{\Theta}_j} \tag{A4.7a}$$

$\mathbf{H}(\boldsymbol{\Theta}_o)$ refers to the Hessian matrix evaluated at the point $\boldsymbol{\Theta}_o$ and provides a measure of the local curvature of L around that point. The **Fisher information matrix** ($\mathbf{F}$), the negative of expected value of the Hessian matrix for L,

$$\mathbf{F}(\boldsymbol{\Theta}) = -E\left[\,\mathbf{H}(\boldsymbol{\Theta})\,\right] \tag{A4.7b}$$

provides a measure of the multidimensional curvature of the log-likelihood surface. Alternately, $\mathbf{F}$ can be computed as the expected value of the outer product of the score vector,

$$\mathbf{F}(\boldsymbol{\Theta}) = E\left[\mathbf{S}(\boldsymbol{\Theta})\,\mathbf{S}(\boldsymbol{\Theta})^T\right] \tag{A4.7c}$$

The covariance matrix for the MLEs is simply the inverse of the information matrix, with

$$\sigma\left(\widehat{\boldsymbol{\Theta}}_i, \widehat{\boldsymbol{\Theta}}_j\right) = \left[\mathbf{F}(\boldsymbol{\Theta})^{-1}\right]_{ij} \tag{A4.7d}$$

As in the univariate case, if the likelihood surface is highly curved (very peaked) around the MLE, then the standard errors (being the inverse of the local curvature) are small, while if the likelihood is very flat, the sampling variance is large. For large sample sizes, $\mathbf{F}$ is often approximated by the Hessian matrix evaluated at the MLE,

$$\mathbf{F}(\boldsymbol{\Theta}) \simeq -\mathbf{H}(\widehat{\boldsymbol{\Theta}}) \tag{A4.7e}$$

Example 3. Suppose n values are sampled independently from a normal with unknown mean and variance. What are the MLEs and their sampling variances? Here $\boldsymbol{\Theta} = (\mu, \sigma)^T$. Noting that $\sum_{i=1}^{n}(z_i - \mu)^2 = n(\,\overline{z^2} - 2\,\overline{z}\,\mu + \mu^2)$, the same logic leading to Equation A4.3 shows that the log-likelihood function is

$$L(\mu, \sigma^2 \,|\, \mathbf{z}) = -\left(\frac{n}{2}\right)\ln(2\pi) - \left(\frac{n}{2}\right)\ln(\sigma^2) - \frac{n\,(\,\overline{z^2} - 2\,\overline{z}\,\mu + \mu^2\,)}{2\,\sigma^2} \tag{A4.8a}$$

Taking derivatives, the score vector becomes

$$\mathbf{S}(\boldsymbol{\Theta}) = \begin{pmatrix} \partial L(\boldsymbol{\Theta})/\partial\mu \\ \\ \partial L(\boldsymbol{\Theta})/\partial\sigma^2 \end{pmatrix} = \left(\frac{n}{\sigma^2}\right)\begin{pmatrix} \overline{z} - \mu \\ \\ \dfrac{\overline{z^2} - 2\,\overline{z}\,\mu + \mu^2}{2\,\sigma^2} - \dfrac{1}{2} \end{pmatrix} \tag{A4.8b}$$

Solving $\mathbf{S}(\widehat{\boldsymbol{\Theta}}) = \mathbf{0}$ gives the MLEs as

$$\widehat{\boldsymbol{\Theta}} = \begin{pmatrix} \widehat{\mu} \\ \\ \widehat{\sigma}^{\,2} \end{pmatrix} = \begin{pmatrix} \overline{z} \\ \\ \overline{z^2} - \overline{z}^{\,2} \end{pmatrix} \tag{A4.8c}$$

As the first step towards computing the Hessian and Fisher matrices, the second partials are found to be

$$\frac{\partial L^2}{(\partial\mu)^2} = -\frac{n}{\sigma^2}, \qquad \frac{\partial L^2}{\partial\mu\,\partial\sigma^2} = -\frac{n(\,\overline{z}-\mu\,)}{\sigma^4} \tag{A4.8d}$$

$$\frac{\partial L^2}{(\partial\sigma^2)^2} = \frac{n}{2\,\sigma^4}\left(1 - \frac{2\,(\,\overline{z^2} - 2\,\overline{z}\,\mu + \mu^2\,)}{\sigma^2}\right) \tag{A4.8e}$$

Since $E(\,\overline{z}\,) = \mu$, the first two derivatives have expected values of $-n/\sigma^2$ and 0. Likewise, since $E(\,\overline{z^2}\,) = \mu^2+\sigma^2$, the expected value of Equation A4.8e becomes

$$E\left(\frac{\partial L^2}{(\partial\sigma^2)^2}\right) = \frac{n}{2\,\sigma^4}\left(1 - \frac{2\,(\,\mu^2+\sigma^2-2\,\mu^2+\mu^2\,)}{\sigma^2}\right) = -\frac{n}{2\,\sigma^4}$$

With the above results, the Fisher matrix becomes

$$\mathbf{F} = -E(\mathbf{H}) = \begin{pmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{n}{2\,\sigma^4} \end{pmatrix}$$

Alternatively, evaluating the derivatives at the MLE, $\widehat{\boldsymbol{\Theta}} = (\,\overline{z},\,\widehat{\sigma}^{\,2})^T$, Equation A4.8d gives values of $-n/\widehat{\sigma}^{\,2}$ and 0, while Equation A4.8e gives $-n/(2\,\widehat{\sigma}^{\,4})$, so that the value of the Hessian matrix evaluated at the MLE becomes

$$\mathbf{H}(\widehat{\boldsymbol{\Theta}}) = -\begin{pmatrix} \dfrac{n}{\widehat{\sigma}^{\,2}} & 0 \\ 0 & \dfrac{n}{2\,\widehat{\sigma}^{\,4}} \end{pmatrix}$$

Applying Equation A4.7d gives the large-sample variances and covariance for the MLEs as

$$\sigma^2(\,\widehat{\mu}\,) = \sigma^2/n \simeq \widehat{\sigma}^{\,2}/n, \quad \sigma^2(\,\widehat{\sigma}^{\,2}) = 2\,\sigma^4/n \simeq 2\,\widehat{\sigma}^{\,4}/n, \quad \sigma(\,\widehat{\mu},\widehat{\sigma}^{\,2}) = 0$$

LIKELIHOOD-RATIO TESTS

Maximum likelihood provides for extremely convenient tests of hypotheses in the form of **likelihood-ratio**, or LR, tests (reviewed in Chapter 24 of Kendall and Stuart 1979) that examine whether a reduced model provides the same fit as a full model. The likelihood-ratio test statistic is given by

$$LR = 2\ln\left(\frac{\ell(\widehat{\boldsymbol{\Theta}}\,|\,\mathbf{z})}{\ell(\widehat{\boldsymbol{\Theta}}_r\,|\,\mathbf{z})}\right) = -2\ln\left(\frac{\ell(\widehat{\boldsymbol{\Theta}}_r\,|\,\mathbf{z})}{\ell(\widehat{\boldsymbol{\Theta}}\,|\,\mathbf{z})}\right) = -2\left[\,L(\widehat{\boldsymbol{\Theta}}_r\,|\,\mathbf{z}) - L(\widehat{\boldsymbol{\Theta}}\,|\,\mathbf{z})\right] \tag{A4.9}$$

where $\ell(\widehat{\boldsymbol{\Theta}} \,|\, \mathbf{z})$ is the likelihood evaluated at the MLE and $\ell(\widehat{\boldsymbol{\Theta}}_r \,|\, \mathbf{z})$ is the maximum of the likelihood function, subject to the restriction that r parameters unconstrained in the full likelihood analysis are assigned fixed values. For sufficiently large sample size, *the LR test statistic is χ^2_r-distributed*, a χ^2 with r degrees of freedom (Wald 1943).

Example 4. Suppose we wish to test the hypothesis that $\mu = 0$ in Example 1. Here the MLE is $\widehat{\mu} = \overline{z}$ and the LR test statistic becomes

$$-2\ln\left(\frac{\ell(\,0\,|\,\mathbf{z})}{\ell(\,\widehat{\mu}\,|\,\mathbf{z})}\right) = -2\ln\left(\frac{(2\pi)^{-n/2}\exp\bigl(-\sum_{i=1}^{n}(z_i-0)^2/2\bigr)}{(2\pi)^{-n/2}\exp\bigl(-\sum_{i=1}^{n}(z_i-\overline{z})^2/2\bigr)}\right)$$

$$= \sum_{i=1}^{n}\left[\, z_i^2 - (z_i-\overline{z})^2 \,\right] = n\,\overline{z}^{\,2}$$

This test statistic is distributed as a χ^2 with one degree of freedom, as one parameter (μ) was assigned a fixed value in the reduced model. Since Prob($\chi^2_1 > 3.84$) = 0.05, the hypothesis $\mu = 0$ is rejected at the 5% level if the test statistic exceeds 3.84.

Now suppose we wish to test this hypothesis under the conditions of Example 3, where the variance is also unknown and hence must also be estimated. Here the MLEs for the full model are given by Equation A4.8c. Substituting $\mu = 0$ into Equation A4.8b gives the score function for the restricted model as

$$\frac{\partial L(\sigma^2)}{\partial \sigma^2} = \frac{n}{\sigma^2}\left(\frac{\overline{z^2}}{2\,\sigma^2} - \frac{1}{2}\right)$$

giving the MLE for σ^2 under this restriction as $\widehat{\sigma}^2_r = \overline{z^2}$. Substituting the MLEs into the likelihood functions, and once again using the identity $\sum(z_i-\mu)^2 = n(\,\overline{z^2} - 2\,\overline{z}\,\mu + \mu^2)$ gives the LR test statistic as

$$-2\ln\left(\frac{\ell(\,0,\widehat{\sigma}^2_r\,|\,\mathbf{z}\,)}{\ell(\,\widehat{\mu},\widehat{\sigma}^{\,2}\,|\,\mathbf{z}\,)}\right)$$

$$= -2\ln\left(\frac{(\,\overline{z^2}\,)^{-n/2}\cdot\exp\bigl[\,-n\,\overline{z^2}\,/\,(\,2\,\overline{z^2}\,)\,\bigr]}{(\,\overline{z^2}-\overline{z}^{\,2}\,)^{-n/2}\cdot\exp\bigl[\,-n\,(\,\overline{z^2}-\overline{z}^{\,2})\,/\,2\,(\,\overline{z^2}-\overline{z}^{\,2})\,\bigr]}\right)$$

$$= -n\,\ln\left(1 - \frac{(\,\overline{z}\,)^2}{\overline{z^2}}\right)$$

Again, for large samples this follows a χ^2_1 distribution as the value of one parameter is assigned a fixed value.

The *G*-test

A common likelihood-ratio based test is the ***G*-test** for goodness of fit. Consider n observations that have been apportioned into a set of N different categories, and denote these by the vector $\mathbf{n} = (n_1, n_2, \cdots, n_N)$. Likewise, let p_i represent the true population frequency of the ith category and let $\mathbf{p} = (p_1, p_2, \cdots, p_N)$. From the multinomial distribution, the likelihood of $\mathbf{p}$ given the observations $\mathbf{n}$ is

$$\ell(\mathbf{p} \,|\, \mathbf{n}) = k\, p_1^{n_1}\, p_2^{n_2} \cdots p_N^{n_N} \tag{A4.10a}$$

where k is the appropriate multinomial coefficient (which is independent of the p_i). It can be shown that the values of p_i that maximize Equation A4.10a (and hence are the MLE's) are $\widehat{p}_i = n_i/n$. This gives the value of the maximum of the likelihood function as

$$\ell(\widehat{\mathbf{p}} \,|\, \mathbf{n}) = k \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2} \cdots \left(\frac{n_N}{n}\right)^{n_N} \tag{A4.10b}$$

In order to test whether the observed data are consistent with a specified vector $\mathbf{q}$ of population frequencies, we need the value of the likelihood function under this constraint. Denoting the expected value for the number of individuals in category i by $\widehat{n}_i = q_i n$, we can write $q_i = \widehat{n}_i/n$. Substitution into Equation A4.10a gives the likelihood under $\mathbf{q}$ as

$$\ell(\mathbf{q} \,|\, \mathbf{n}) = k \left(\frac{\widehat{n}_1}{n}\right)^{n_1} \left(\frac{\widehat{n}_2}{n}\right)^{n_2} \cdots \left(\frac{\widehat{n}_N}{n}\right)^{n_N} \tag{A4.10c}$$

Applying Equation A4.9 yields the likelihood-ratio test (in this case, it is also called the *G*-test, for goodness of fit) that the observed data are consistent with $\mathbf{q}$,

$$G = -2 \ln \left(\frac{\ell(\mathbf{q} \,|\, \mathbf{n})}{\ell(\widehat{\mathbf{p}} \,|\, \mathbf{n})}\right) = -2 \sum_{i=1}^{N} n_i \ln \left(\frac{\widehat{n}_i}{n_i}\right) = -2 \sum_{i=1}^{N} n_i \ln \left(\frac{q_i}{\widehat{p}_i}\right) \tag{A4.11}$$

Since the N frequencies sum to one, there are $N - 1$ unconstrained parameters in the full likelihood, implying that G is asymptotically distributed as a χ^2_{N-1} random variable. Since large sample sizes are required to give the likelihood-ratio test a χ^2 distribution, caution should be exercised in employing this test whenever any expected quantity is less than five (e.g., any $q_i < 5/n$), a problem that can sometimes be avoided by pooling cells. Sokal and Rohlf (1995) provide a thorough overview of these and other matters.

Likelihood-ratio Tests for the General Linear Model

As a final example of likelihood-ratio tests, consider the general linear model (Chapters 8, 26, 27), $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where we assume that the $n \times 1$ vector of residual errors $\mathbf{e}$ is multivariate normal, with mean vector zero and covariance matrix $\mathbf{V}$, i.e., $\mathbf{e} \sim \text{MVN}(\mathbf{0}, \mathbf{V})$. From Equation 8.24, the density function for $\mathbf{e}$ is

$$(2\pi)^{-n/2} |\, \mathbf{V} \,|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{e}^T\mathbf{V}^{-1}\mathbf{e}\right)$$

Writing the vector of residuals as $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ gives the resulting likelihood for $\boldsymbol{\beta}$ and $\mathbf{V}$, conditional on the observed data ($\mathbf{y}$, $\mathbf{X}$), as

$$\ell(\boldsymbol{\beta}, \mathbf{V} \,|\, \mathbf{y},\, \mathbf{X}) = (2\pi)^{-n/2} |\, \mathbf{V} \,|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)$$

which has log-likelihood

$$L = \ln \ell = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln |\, \mathbf{V} \,| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \qquad \text{(A4.12)}$$

Here $\boldsymbol{\beta}$ is a vector of fixed effects and the matrix $\mathbf{V}$ is a function of k variance components, with $\mathbf{V} = \sum_{i=1}^{k} \mathbf{R}_i\, \sigma_i^2$ where the $\mathbf{R}_i$ are matrices of known constants. Thus, the parameters to be estimated are the vector $\boldsymbol{\beta}$ of fixed effects and the k variances, σ_i^2.

Suppose we wish to compare the relative fit of two models that assume the same covariance structure (i.e., the same $\mathbf{V}$), but have different vectors of fixed effects, a vector $\boldsymbol{\beta}_f$ for the full model vs. a vector $\boldsymbol{\beta}_r$ for the reduced model that assumes fewer factors. The resulting likelihood-ratio test statistic is

$$\begin{aligned} LR &= -2\Big[\, L(\widehat{\boldsymbol{\beta}}_r \,|\, \mathbf{y},\, \mathbf{X}_r) - L(\widehat{\boldsymbol{\beta}}_f \,|\, \mathbf{y},\, \mathbf{X}_f)\Big] \\ &= \Big[(\mathbf{y} - \widehat{\mathbf{y}}_r)^T\widehat{\mathbf{V}}^{-1}(\mathbf{y} - \widehat{\mathbf{y}}_r) - (\mathbf{y} - \widehat{\mathbf{y}}_f)^T\widehat{\mathbf{V}}^{-1}(\mathbf{y} - \widehat{\mathbf{y}}_f)\Big] \end{aligned} \qquad \text{(A4.13)}$$

where $\widehat{\mathbf{y}}_f = \mathbf{X}_f\widehat{\boldsymbol{\beta}}_f$ and $\widehat{\mathbf{y}}_r = \mathbf{X}_r\widehat{\boldsymbol{\beta}}_r$ are the predicted means under the full and reduced models. For large sample sizes, this test statistic follows a χ^2 distribution with $n_f - n_r$ degrees of freedom, where n_f and n_r are the degrees of freedom for the full and reduced models, respectively.

Example 5. Suppose the y_i values are the means of n different populations, e.g., data from a series of populations being used in a line-cross analysis (Chapter 9). Assuming the means are independent but with potentially different variances (due to differences in sample sizes, among other things), $\mathbf{V}$ is a diagonal matrix

whose ith element is the variance of the ith mean. Denoting the variance of the ith mean by $\text{Var}(y_i)$, then recalling Equation A3.11b, the quadratic product in the LR test reduces to

$$(\mathbf{y} - \widehat{\mathbf{y}})^T \widehat{\mathbf{V}}^{-1} (\mathbf{y} - \widehat{\mathbf{y}}) = \sum_{i=1}^{n} \frac{(y_i - \widehat{y}_i)^2}{\text{Var}(y_i)}$$

Hence, the likelihood-ratio test statistic for comparing a full model with a reduced model assuming fewer effects is given by

$$\sum_{i=1}^{n} \frac{[\, y_i - \widehat{y}_i(r)\,]^2}{\text{Var}(y_i)} - \sum_{i=1}^{n} \frac{[\, y_i - \widehat{y}_i(f)\,]^2}{\text{Var}(y_i)} \qquad \text{(A4.14)}$$

which is just the difference in the χ^2 values for the fit of the full and reduced models. This test follows a χ^2 distribution with degrees of freedom given by the difference in degrees of freedom for the full and reduced models.

ITERATIVE METHODS FOR SOLVING ML EQUATIONS

While ML estimation and hypothesis testing with likelihood ratios is conceptually straightforward, in practice it can be quite difficult to accomplish due to the complexities associated with having to find the maximum of the likelihood function. Ideally, closed-form solutions to the MLEs can be obtained by deriving the score vector, setting it equal to zero, and solving. However, in many cases this is impractical and numerical approaches must be used. In very simple cases with one or two parameters, a brute force approach relying upon a **grid search** can be used, where one computes a one- or two-dimensional plot of the likelihood surface as a function of the unknown parameters. With more than two variables, this is impractical and a variety of iterative methods have been suggested as alternatives. We discuss two of these here, Newton-Raphson and EM methods (Chapter 27 discusses these methods further in the context of variance-component estimation). A potential problem with all iterative methods is that they may not converge to the true MLEs if the likelihood surface contains several local maxima. Iterative methods require an initial starting value, and a poor choice can result in the iteration converging to a solution that is a local, but not a global, maximum. Hence, when applying iterative methods, several starting points should be used.

Newton-Raphson Methods

Recall from elementary calculus that one can approximate a function $f(x)$ by expanding it in a power series around a point x_o,

$$f(x) \simeq f(x_0) + (x - x_o) f'(x_o)$$

This suggests one approach for finding roots of the equation $f(x) = 0$. Given some initial guess x_0, an improved value is obtained by solving

$$f(x) = 0 \simeq f(x_0) + (x - x_o)f'(x_o)$$

for x, or

$$x \simeq x_0 - \frac{f(x_0)}{f'(x_0)} \tag{A4.15a}$$

Noting that the score function is zero at the MLE [$S(\widehat{\theta}) = 0$], this suggests one approach for obtaining an iterative solution of the MLE. Applying Equation A4.15a to the score function, so that $f = S$ and $f' = \partial S(\theta)/\partial\theta = \partial L^2(\theta)/\partial^2\theta$, an updated estimate, $\widehat{\theta}^{(k+1)}$, of a current estimate $\widehat{\theta}^{(k)}$ is given by

$$\widehat{\theta}^{(k+1)} = \widehat{\theta}^{(k)} - \left(\frac{\partial L^2(\theta)}{\partial^2\theta}\bigg|_{\theta=\widehat{\theta}^{(k)}}\right)^{-1} S\big[\widehat{\theta}^{(k)}\big] \tag{A4.15b}$$

which is interated until $|\widehat{\theta}^{(k+1)} - \widehat{\theta}^{(k)}|$ is sufficiently small to declare convergence. This is the **Newton-Raphson** method, a member of a class of **quadratic methods**. Such methods involve second partial derivatives of the likelihood function and have a quadratic convergence rate. The same logic when applied to a multivariate Taylor series implies that a vector $\widehat{\boldsymbol{\Theta}}^{(k)}$ of current estimates is updated by using

$$\widehat{\boldsymbol{\Theta}}^{(k+1)} = \widehat{\boldsymbol{\Theta}}^{(k)} - \mathbf{H}^{-1}\big(\widehat{\boldsymbol{\Theta}}^{(k)}\big)\,\mathbf{S}\big(\widehat{\boldsymbol{\Theta}}^{(k)}\big) \tag{A4.16}$$

where $\mathbf{S}$ and $\mathbf{H}$ are the vector of scores and the Hessian matrix, respectively, both evaluated at the current estimate.

One variant of this approach is **Fisher's scoring**, where the Hessian matrix $\mathbf{H}$ is replaced by its expected value, the negative of Fisher's information matrix $\mathbf{F}$ (Equation A4.7b),

$$\widehat{\boldsymbol{\Theta}}^{(k+1)} = \widehat{\boldsymbol{\Theta}}^{(k)} + \mathbf{F}^{-1}\big(\widehat{\boldsymbol{\Theta}}^{(k)}\big)\,\mathbf{S}\big(\widehat{\boldsymbol{\Theta}}^{(k)}\big) \tag{A4.17}$$

One advantage of Fisher's scoring is that $\mathbf{F}$ is usually of a simpler form than $\mathbf{H}$, often containing elements equal to zero that are non-zero in $\mathbf{H}$. This can make $\mathbf{F}$ easier to compute and invert (e.g., compare Equations 27.34 and 27.35b). Further, Fisher's scoring appears to be more robust to poor initial starting choices than the strict Newton-Raphson method (Jennrich and Sampson 1986). In addition to the advantage of quadratic convergence, both Newton-Raphson and Fisher's scoring yield the covariance matrix of MLE estimates from $\mathbf{H}$ (or $\mathbf{F}$) using the final interation values of $\boldsymbol{\Theta}$ and applying Equation A4.7. Additional quadratic methods are discussed by Kennedy and Gentle (1980).

Expectation-maximization (EM) Methods

Newton-Raphson and related methods require the first and second derivatives of the likelihood function, which can be difficult to obtain and/or computationally demanding (e.g., requiring the repeated inversion of large matrices). An alternative strategy is to use expectation-maximization (EM) methods, which were introduced by Dempster et al. (1977) as a very general iterative approach for data sets with missing (or incomplete) data. The idea is that, in many cases, if we had more information about certain observations, MLEs are easily obtained. For example, if observations are drawn from a mixture distribution (Chapter 13), obtaining the MLEs for the means and variances of the underlying distributions is trival *provided* we know from which distribution each individual observation is drawn. Thus the original data set is treated as incomplete data, missing additional information (e.g., for a mixture model, which distribution a specific observation is drawn from). Using a current estimate of the unknown parameter values, the expected value of the incomplete data is computed (e.g., for a mixture model, the category identity of each individual is estimated). This is the **expectation**, or **E step**. The result is a set of likelihood equations that are considerably easier to solve than the full likelihood (the **maximization**, or **M step**). The new estimates obtained from the M step are then used to update the expected values, and this approach is iterated until convergence. The EM method refers to a general class of approaches, and there can be several EM versions for solving the same problem.

While EM methods often have fairly simple forms and hence are easy to program, they can be extremely slow to converge to a solution. EM methods offer computational advantages over Newton-Raphson methods, as they do not have to compute second derivatives of the likelihood function and they do not directly evaluate the full likelihood function. However, this is a disadvantage in terms of constructing confidence intervals and LR tests, as other approaches must be used to obtain the standard errors of the MLEs and to compute the likelihoods needed for LR tests. Chapter 27 discusses an EM method for computing unknown variance components in linear models, while our focus here is on the other broad class of likelihood models used throughout this book, mixture models (introduced in Chapter 13).

EM for Mixture Model Likelihoods

Mixture models naturally appear in a number of quantitative-genetic settings, wherein the observed distribution is really a weighted sum of a number of underlying distributions. For example, when a major diallelic locus is segregating in a population, the phenotypic distribution is a weighted sum of the three distributions representing each major locus genotype (Chapter 13). The general likelihood function for a single observation z from the kinds of mixture models considered

in this book has the form

$$\ell(\boldsymbol{\Theta} \mid z) = \sum_{k=1}^{N} \pi_k \cdot \varphi(z, \mu_k, \sigma^2) \tag{A4.18a}$$

where the distribution is assumed to result from N underlying normals, the kth of which has frequency π_k, mean μ_k, and common variance σ^2. We assume that the number N of underlying distributions is known and wish to estimate the $2N \times 1$ vector $\boldsymbol{\Theta}$ of the N means, the common variance, and the $N - 1$ independent mixing proportions (the π_k). With n individuals independently drawn from this distribution, the full likelihood is

$$\ell(\boldsymbol{\Theta} \mid \mathbf{z}) = \prod_{i=1}^{n} \ell(\boldsymbol{\Theta} \mid z_i) \tag{A4.18b}$$

While appearing rather simple, the full likelihood function is complicated to work with analytically, and numerical approaches are usually employed.

When we observe a particular value, we don't know which underlying distribution (or category) it was drawn from. If we knew the category identity for each observation, the ML solutions for the mean and variance of the underlying distributions are easily computed. For example, if a single diallelic QTL is segregating, if we could determine whether individuals had QTL genotype *QQ*, *Qq*, or *qq*, then the mean for each genotype and the common variance could easily be estimated. This is the basis of the EM method. We start with some initial guess as to the category identity of each observation, which then allows us to easily compute an ML estimate of the means and variance of the underlying distribution. This guess is in the form of a weight vector for each individual, whose k element, $w(k \mid z)$, is the probability that an individual has the kth QTL genotype given they have trait value z. Using these mean and variance estimates, updated weights can be computed using **Bayes' theorem** (Equation 13.24) for conditional probabilities. Since $w(k \mid z) = \Pr(k \mid z)$, applying Bayes' theorem gives

$$w(k \mid z) = \frac{\Pr(k) \cdot \Pr(z \mid k)}{\Pr(z)} = \frac{\pi_k \cdot \varphi(z, \mu_k, \sigma^2)}{\Pr(z)} = \frac{\pi_k \cdot \varphi(z, \mu_k, \sigma^2)}{\displaystyle\sum_{j=1}^{N} \pi_j \cdot \varphi(z, \mu_j, \sigma^2)} \tag{A4.19}$$

These updated weights are then used to obtain new estimates of the category-specific means and the variance, and this procedure is repeated until convergence. Formally, this EM approach proceeds as follows (Aitkin and Wilson 1980):

(1) **Initial step**. Choose initial starting values for the MLEs of the variance $\widehat{\sigma}^{2\,(0)}$ and the vectors of mixture proportions and means,

$$\widehat{\boldsymbol{\pi}}^{(0)} = (\widehat{\pi}_1^{(0)}, \cdots, \widehat{\pi}_N^{(0)}), \qquad \widehat{\boldsymbol{\mu}}^{(0)} = (\widehat{\mu}_1^{(0)}, \cdots, \widehat{\mu}_N^{(0)}) \tag{A4.20}$$

(2) **E-step**. Define the weight $w^{(1)}(k \mid z_i)$ as the probability that observation z_i is drawn from distribution k given the current estimates $\widehat{\sigma}^{\,2\,(0)}$, $\widehat{\boldsymbol{\pi}}^{\,(0)}$, and $\widehat{\boldsymbol{\mu}}^{\,(0)}$. From Bayes' theorem,

$$w^{(1)}(k \mid z_i) = \frac{\widehat{\pi}_k^{(0)} \cdot \varphi(z_i, \widehat{\mu}_k^{(0)}, \widehat{\sigma}^{\,2\,(0)})}{\sum_{j=1}^{N} \widehat{\pi}_j^{(0)} \cdot \varphi(z_i, \widehat{\mu}_j^{(0)}, \widehat{\sigma}^{\,2\,(0)})} \tag{A4.21}$$

where $\varphi(z_i, \widehat{\mu}_k^{(0)}, \widehat{\sigma}^{\,2\,(0)})$ is the normal distribution evaluated at the value z_i using mean $\widehat{\mu}_k^{(0)}$ and variance $\widehat{\sigma}^{\,2\,(0)}$.

(3) **M-step**. Assuming these weights are correct, the updated estimates of the MLEs are obtained as follows:

(a) **Mixing proportions**: Given by the average probability of being in category k,

$$\widehat{\pi}_k^{(1)} = \overline{w}_k^{(1)} = \frac{1}{n} \sum_{i=1}^{n} w^{(1)}(k \mid z_i) \tag{A4.22a}$$

(b) **Means**: Given by the weighted average of the observations,

$$\widehat{\mu}_k^{(1)} = \frac{1}{n} \sum_{i=1}^{n} z_i \left(\frac{w^{(1)}(k \mid z_i)}{\overline{w}_k^{(1)}} \right) \tag{A4.22b}$$

(c) **Variance**: Given by the weighted variance of the observations,

$$\widehat{\sigma}^{\,2\,(1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{N} \left(z_i - \widehat{\mu}_k^{(1)} \right)^2 w^{(1)}(k \mid z_i) \tag{A4.22c}$$

These updated estimates are then used to compute new weights, and the whole procedure continues until the interations converge.