Numerical Techniques for Maximization

Many of the procedures discussed in this volume require maximizing the log likelihood or partial log likelihood function. For many models, it is impossible to perform this maximization analytically, so, numerical methods must be employed. In this appendix, we shall summarize some techniques which can be used in both univariate and multivariate cases. The reader is referred to a text on statistical computing, such as Thisted (1988), for a more detailed discussion of these techniques.

A.1 Univariate Methods

Suppose we wish to find the value $x$ which maximizes a function $f(x)$ of a single variable. Under some mild regularity conditions, $x$ maximizes $f$ if the score equation $f'(x)$ equals 0 and $f''(x) < 0$. We present three numerical methods which attempt to find the maximum of $f(x)$ by solving the score equation. Some care must be taken when using these routines because they do not ensure that the second derivative of $f$ is negative at the solution we find.

The first technique is the bisection method. Here, the algorithm starts with two initial values, $x_l$ and $x_u$, which bracket the root of $f'(x) = 0$, that is, $f'(x_l) \cdot f'(x_u) < 0$. A new guess at the root is taken to be the
midpoint of the interval \((x_L, x_U)\), namely, \(X_N = (x_L + x_U)/2\). If \(f'(x_L)\) and \(f'(x_U)\) have the same sign, \(x_L\) is replaced by \(x_N\), otherwise \(x_U\) is replaced by \(x_N\). In either case, the algorithm continues with the new values of \(x_L\) and \(x_U\) until the desired accuracy is achieved. At each step, the length of the interval \((x_U - x_L)\) is a measure of the largest possible difference between our updated guess at the root of \(f'(\cdot)\) and the actual value of the root.

A second method, of use when one has good initial starting values and a complicated second derivative of \(f(\cdot)\), is the secant method or \textit{regula falsi}. Again, we start with two initial guesses at the root, \(x_0\) and \(x_1\). These guesses need not bracket the root. After \(i\) steps of the algorithm, the new guess at the root of \(f'(\cdot)\) is given by

\[
x_{i+1} = x_i - f'(x_i)(x_i - x_{i-1})/f'(x_i) - f'(x_{i-1}).
\]

(A.1)

Iterations continue until convergence. Typical stopping criteria are

\[|x_{i+1} - x_i| < \gamma, |f'(x_{i+1})| < \gamma\]

or

\[|(x_{i+1} - x_i)/x_i| < \gamma,
\]

where \(\gamma\) is some small number.

The third method is the Newton–Raphson technique. Here, a single initial guess, \(x_0\), of the root is made. After \(i\) steps of the algorithm, the updated guess is given by

\[
x_{i+1} = x_i - f'(x_i)/f(x_i).
\]

(A.2)

Again, the iterative procedure continues until the desired level of accuracy is met. Compared to the secant method, this technique has the advantage of requiring a single starting value, and convergence is quicker than the secant method when the starting values are good. Both the secant and Newton–Raphson techniques may fail to converge when the starting values are not close to the maximum.

\section*{Example A.1}

Suppose we have the following 10 uncensored observations from a Weibull model with scale parameter \(\lambda = 1\) and shape parameter \(\alpha\), that is, \(k(t) = \alpha \lambda^{-\alpha} e^{-\alpha t}\).

Data: 2.57, 0.58, 0.82, 1.02, 0.78, 0.46, 1.04, 0.43, 0.69, 1.37

To find the maximum likelihood estimator of \(\alpha\), we need to maximize the log likelihood \(f(\alpha) = \ln L(\alpha) = n \ln(\alpha) + (\alpha - 1) \sum \ln(t_i) - \sum t_i^\alpha\).

Here, \(f'(\alpha) = \sum \ln(t_i) - \sum t_i^\alpha \ln(t_i),\) and \(f''(\alpha) = -n/\alpha^2 - \sum t_i^\alpha/\alpha \ln(t_i)^2\).

Applying the bisection method with \(\alpha_0 = 1.5\) and \(\alpha_1 = 2\) and stopping the algorithm when \(|f'(\alpha)| < 0.01\), we have the following values:

<table>
<thead>
<tr>
<th>Step</th>
<th>(\alpha_1)</th>
<th>(\alpha_2)</th>
<th>(\alpha_3)</th>
<th>(f'(\alpha_0))</th>
<th>(f'(\alpha_1))</th>
<th>(f'(\alpha_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5</td>
<td>1.75</td>
<td>1.798</td>
<td>-2.589</td>
<td>-0.387</td>
<td>-0.059</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>1.75</td>
<td>1.798</td>
<td>-0.387</td>
<td>0.697</td>
<td>-0.154</td>
</tr>
<tr>
<td>3</td>
<td>1.625</td>
<td>1.75</td>
<td>1.6875</td>
<td>0.067</td>
<td>-0.387</td>
<td>0.154</td>
</tr>
<tr>
<td>4</td>
<td>1.6875</td>
<td>1.75</td>
<td>1.71875</td>
<td>0.154</td>
<td>-0.387</td>
<td>-0.116</td>
</tr>
<tr>
<td>5</td>
<td>1.6975</td>
<td>1.71875</td>
<td>1.70333</td>
<td>0.154</td>
<td>-0.116</td>
<td>-0.049</td>
</tr>
<tr>
<td>6</td>
<td>1.70313</td>
<td>1.71875</td>
<td>1.71049</td>
<td>0.019</td>
<td>-0.116</td>
<td>-0.049</td>
</tr>
<tr>
<td>7</td>
<td>1.70313</td>
<td>1.70949</td>
<td>1.70704</td>
<td>0.019</td>
<td>-0.049</td>
<td>-0.049</td>
</tr>
<tr>
<td>8</td>
<td>1.70313</td>
<td>1.70704</td>
<td>1.70509</td>
<td>0.019</td>
<td>-0.049</td>
<td>-0.049</td>
</tr>
</tbody>
</table>

So, after eight steps the algorithm stops with \(\hat{\alpha} = 1.705\).

For the secant method, we shall start the algorithm with \(\alpha_0 = 1\) and \(\alpha_1 = 1.5\). The results are in the following table:

<table>
<thead>
<tr>
<th>Step</th>
<th>(\alpha_{i-1})</th>
<th>(\alpha_i)</th>
<th>(f'(\alpha_{i-1}))</th>
<th>(f'(\alpha_i))</th>
<th>(\alpha_{i+1})</th>
<th>(f'(\alpha_{i+1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5</td>
<td>1.75</td>
<td>7.065</td>
<td>1.798</td>
<td>1.671</td>
<td>0.300</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>1.671</td>
<td>1.798</td>
<td>1.671</td>
<td>0.300</td>
<td>1.705</td>
</tr>
</tbody>
</table>

Here, using the same stopping rule \(f'(\alpha) < 0.01\), the algorithm stops after two steps with \(\hat{\alpha} = 1.705\).

For the Newton–Raphson procedure, we use an initial value of \(\alpha_0 = 1.5\). The results of the algorithm are in the following table.

<table>
<thead>
<tr>
<th>Step</th>
<th>(\alpha_{i-1})</th>
<th>(f'(\alpha_{i-1}))</th>
<th>(f'(\alpha_i))</th>
<th>(\alpha_i)</th>
<th>(f'(\alpha_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5</td>
<td>-9.847</td>
<td>1.701</td>
<td>0.038</td>
<td>1.705</td>
</tr>
<tr>
<td>2</td>
<td>1.701</td>
<td>0.038</td>
<td>8.655</td>
<td>1.705</td>
<td>2.9 \times 10^{-6}</td>
</tr>
</tbody>
</table>

Again, using the same stopping rule \(f'(\alpha) < 0.01\), the algorithm stops after two steps with \(\hat{\alpha} = 1.705\). Notice the first step of the Newton–Raphson algorithm moves closer to the root than the secant method.

\section*{A.2 Multivariate Methods}

We present three methods to maximize a function of more than one variable. The first is the method of steepest ascent which requires only the vector of first derivatives of the function. This method is robust to the starting values used in the iterative scheme, but may require a large number of steps to converge to the maximum. The second is the multivariate extension of the Newton–Raphson method. This method, which requires both the first and second derivatives of the function,
converges quite rapidly when the starting values are close to the root, but may not converge when the starting values are poorly chosen. The third, called Marquardt’s (1963) method, is a compromise between these two methods. It uses a blending constant which controls how closely the algorithm resembles either the method of steepest ascent or the Newton–Raphson method.

Some notation is needed before presenting the three methods. Let \( f(\mathbf{x}) \) be a function of the \( p \)-dimensional vector \( \mathbf{x} = (x_1, \ldots, x_p)^T \). Let \( \mathbf{u}(\mathbf{x}) \) be the \( p \)-vector of first order partial derivatives of \( f(\mathbf{x}) \), that is,

\[
\mathbf{u}(\mathbf{x}) = \{u_1(\mathbf{x}), \ldots, u_p(\mathbf{x})\}^T
\]

where

\[
u_j(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_j}, \quad j = 1, \ldots, p.
\]

Let \( \mathbf{H}(\mathbf{x}) \) be the \( p \times p \) Hessian matrix of mixed second partial derivatives of \( f(\mathbf{x}) \), defined by

\[
\mathbf{H}(\mathbf{x}) = (H_{ij}(\mathbf{x})), \quad i, j = 1, \ldots, p \quad \text{where} \quad H_{ij}(\mathbf{x}) = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}.
\]

The method of steepest ascent starts with an initial guess, \( \mathbf{x}_0 \), of the point which maximizes \( f(\mathbf{x}) \). At any point, the gradient vector \( \mathbf{u}(\mathbf{x}) \) points the direction of steepest ascent of the function \( f(\mathbf{x}) \). The algorithm moves along this direction by an amount \( d \) to a new estimate of the maximum from the current estimate. The step size \( d \) is chosen to maximize the function in this direction, that is, we pick \( d \) to maximize \( f(\mathbf{x} + du(\mathbf{x})) \). This requires maximizing a function of a single variable, so that any of the techniques discussed earlier can be employed.

The updated guess at the point which maximizes \( f(\mathbf{x}) \) is given by

\[
\mathbf{x}_{k+1} = \mathbf{x}_k + du(\mathbf{x}_k).
\]

The second method is the Newton–Raphson method which, like the method of steepest ascent, starts with an initial guess at the point which maximizes \( f(\mathbf{x}) \). After \( k \) steps of the algorithm, the updated estimate of the point which maximizes \( f(\mathbf{x}) \) is given by

\[
\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}(\mathbf{x}_k)^{-1}\mathbf{u}(\mathbf{x}_k).
\]

The Newton–Raphson algorithm converges quite rapidly when the initial guess is not too far from the maximum. When the initial guess is poor, the algorithm may move in the wrong direction or may take a step in the correct direction, but overshoot the root. The value of the function should be computed at each step to ensure that the algorithm is moving in the correct direction. If \( f(\mathbf{x}_k) \) is smaller than \( f(\mathbf{x}_{k+1}) \), one option is to cut the step size in half and try \( \mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}(\mathbf{x}_k)^{-1}\mathbf{u}(\mathbf{x}_k)/2 \). This procedure is used in SAS and BMDP in the Cox regression procedure.

The third method is Marquardt’s (1963) compromise between the method of steepest ascent and the Newton–Raphson method. This method uses a constant, \( \gamma \), which blends the two methods together. When \( \gamma \) is zero, the method reduces to the Newton–Raphson method, and, as \( \gamma \to 0 \), the method approaches the method of steepest ascent. Again, the method starts with an initial guess, \( \mathbf{x}_0 \). Let \( \mathbf{S} \) be the \( p \times p \) diagonal scaling matrix with diagonal element \( (\mathbf{H}(\mathbf{x}_k)^{-1})^{-1} \). The updated estimate of the maximum is given by

\[
\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{S}_k \mathbf{H}(\mathbf{x}_k)\mathbf{S}_k + \gamma \mathbf{I}^{-1}\mathbf{u}(\mathbf{x}_k),
\]

where \( \mathbf{I} \) is the identity matrix. Typically, the algorithm is implemented with a small value of \( \gamma \) for the first iteration. If \( f(\mathbf{x}_k) < f(\mathbf{x}_0) \), then we are having difficulty approaching the maximum and the value of \( \gamma \) is increased until \( f(\mathbf{x}_k) > f(\mathbf{x}_0) \). This procedure is iterated until convergence is attained. For the final step of the algorithm, a "Newton–Raphson" step with \( \gamma = 0 \) is taken to ensure convergence.

In the multivariate maximization problem, there are several suggestions for declaring convergence of these algorithms. These include stopping when \( f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < \varepsilon \) or \( ||f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)|| < \varepsilon \) or when \( \sum |u_j(\mathbf{x}_{k+1})|^2 < \varepsilon \) or \( \max|u_j(\mathbf{x}_{k+1})| < \varepsilon \) or when \( \sum (x_{j+1} - x_j)^2 < \varepsilon \) or \( \max(x_{j+1} - x_j) < \varepsilon \).

**EXAMPLE A.2**

We shall fit a two-parameter Weibull model with survival function \( S(t) = \exp(-\lambda t^\alpha) \) to the ten observations in Example A.2. Here the log likelihood function is given by

\[
L(\lambda, \alpha) = n ln \lambda + n ln \alpha + (\alpha - 1) \sum ln t_i - \lambda \sum \eta_i.
\]

The score vector \( \mathbf{u}(\lambda, \alpha) \) is expressed by

\[
u_\lambda(\lambda, \alpha) = \frac{\partial L(\lambda, \alpha)}{\partial \lambda} = \frac{n}{\lambda} - \sum \eta_i
\]

and

\[
u_\alpha(\lambda, \alpha) = \frac{\partial L(\lambda, \alpha)}{\partial \alpha} = \frac{n}{\alpha} + \sum ln t_i - \lambda \sum \eta_i ln t_i
\]

The Hessian matrix is

\[
\mathbf{H}(\lambda, \alpha) = \begin{pmatrix}
-\frac{n^2}{\lambda^3} & -\sum \eta_i t_i \\
-\sum \eta_i t_i & -\frac{n}{\alpha^3} - \lambda \sum \eta_i (ln \eta_i)
\end{pmatrix}
\]

To apply the method of steepest ascent, we must find the value of \( \lambda \) which maximizes \( L(\lambda + d\lambda, \alpha_k, \alpha) \), \( \alpha_k \) and \( d\lambda \) are the initial guess for \( \lambda \).

This needs to be done numerically and this example uses a Newton–Raphson algorithm. Convergence of the algorithm is declared when the maximum of \( \eta_1 \) and \( \eta_4 \) is less than 0.1. Starting with an initial guess of \( \alpha = 1 \) and \( \lambda = 10/\sum t_i = 1.024 \), which leads to a log likelihood of \(-9.757\), we have the following results:

\[
x_{k+1} = x_k - S_k \mathbf{H}(x_k) S_k + \gamma I^{-1} \mathbf{u}(x_k),
\]

where \( I \) is the identity matrix.
Thus the method of steepest ascent yields maximum likelihood estimates of $\hat{\theta} = 0.839$ and $\hat{\varphi} = 1.792$ after 5 iterations of the algorithm. Applying the Newton-Raphson algorithm with the same starting values and convergence criterion yields

This method yields maximum likelihood estimates of $\hat{\lambda} = 0.832$ and $\hat{\varphi} = 1.796$ after three iterations.

Using $\gamma = 0.5$ in Marquardt's method yields

Here, the algorithm converges in five steps to estimates of $\hat{\lambda} = 0.845$ and $\hat{\varphi} = 1.777$.

Large-Sample Tests Based on Likelihood Theory

Many of the test procedures used in survival analysis are based on the asymptotic properties of the likelihood or the partial likelihood. These test procedures are based on either the maximized likelihood itself (likelihood ratio tests), on the estimators standardized by use of the information matrix (Wald tests), or on the first derivatives of the log likelihood (score tests). In this appendix, we will review how these tests are constructed. See Chapter 9 of Cox and Hinkley (1974) for a more detailed reference.

Let $Y$ denote the data and $\theta = (\theta_1, \ldots , \theta_p)$ be the parameter vector. Let $L(\theta : Y)$ denote either the likelihood or partial likelihood function. The maximum likelihood estimator of $\theta$ is the function of the data which maximizes the likelihood, that is, $\hat{\theta}(Y) = \hat{\theta}$ is the value of $\theta$ which maximizes $L(\theta : Y)$ or, equivalently, maximizes $\log L(\theta : Y)$.

Associated with the likelihood function is the efficient score vector $U(\theta) = (U_1(\theta), \ldots , U_p(\theta))$ defined by

$$U_j(\theta) = \frac{\delta}{\delta \theta_j} \ln L(\theta : Y).$$

In most regular cases, the maximum likelihood estimator is the solution to the equation $U(\theta) = 0$. The efficient score vector has the property
that its expected value is zero when the expectation is taken with respect to the true value of $\theta$.

A second key quantity in large-sample likelihood theory is the Fisher information matrix defined by

$$ I(\theta) = E_\theta \left[ \frac{\partial^2}{\partial \theta \partial \theta^T} \ln L(\theta) \right] $$

$$ = \left\{ -E_\theta \left[ \frac{\partial^2}{\partial \theta \partial \theta^T} \ln L(\theta \mid Y) \right] \right\}, \quad j, k = 1, \ldots, p. \quad (B.2) $$

Computation of the expectation in (B.2) is very difficult in most applications of likelihood theory, so a consistent estimator of $I(\theta)$ is used. This estimator is the observed information, $I(\hat{\theta})$, whose $(j, k)$th element is given by

$$ I_{jk}(\hat{\theta}) = -\frac{\partial^2}{\partial \theta_j \partial \theta_k} \ln L(\hat{\theta} \mid Y), \quad j, k = 1, \ldots, p. \quad (B.3) $$

The first set of tests based on the likelihood are for the simple null hypothesis, $H_0: \theta = \theta_0$. The first test is the likelihood ratio test based on the statistic

$$ X^2_{LR} = -2 \ln L(\theta_0 \mid Y) - \ln L(\hat{\theta} \mid Y) $$

This statistic has an asymptotic chi-squared distribution with $p$ degrees of freedom under the null hypothesis.

A second test, called the Wald test, is based on the large-sample distribution of the maximum likelihood estimator. For large samples, $\theta$ has a multivariate normal distribution with mean $\theta_0$ and covariance matrix $I(\theta_0)$, so the quadratic form

$$ (\theta - \theta_0)^T I^{-1}(\theta_0) (\theta - \theta_0) $$

has a chi-squared distribution with $p$ degrees of freedom for large samples. Using the observed information as an estimator of the Fisher information, the Wald statistic is expressed as

$$ X^2_W = (\theta - \theta_0)^T I^{-1}(\hat{\theta}) (\theta - \theta_0) \quad (B.5) $$

which has a chi-squared distribution with $p$ degrees of freedom for large samples when $H_0$ is true.

The third test, called the score or Rao test, is based on the efficient score statistics. When $\theta = \theta_0$, the score vector $U(\theta_0)$ has a large-sample multivariate normal distribution with mean 0 and covariance matrix $I(\theta_0)$. This leads to a test statistic given by

$$ X^2_S = (\theta - \theta_0)^T I^{-1}(\theta_0) (\theta - \theta_0) $$

As for the Wald test, the Fisher information is replaced in most applications by the observed information, so the test statistic is given by

$$ X^2_S = (\theta - \theta_0)^T I^{-1}(\hat{\theta}) (\theta - \theta_0) \quad (B.6) $$

Again, this statistic has an asymptotic chi-squared distribution with $p$ degrees of freedom when $H_0$ is true. The score test has an advantage in many applications in that the maximum likelihood estimates need not be calculated.

**EXAMPLE B.1**

Suppose we have a censored sample of size $n$ from an exponential population with hazard rate $\lambda$. We wish to test the hypothesis that $\lambda = 1$. Let $(T_i, \delta_i), i = 1, \ldots, n$, so that the likelihood, $L(\lambda; T_i, \delta_i), i = 1, \ldots, n$, is given by

$$ \prod_{i=1}^n \lambda^{\delta_i} e^{-\lambda T_i} = \lambda^{\delta_i} e^{-\lambda T_i} $$

where $D = \sum_{i=1}^n \delta_i$ is the observed number of deaths and $S = \sum_{i=1}^n T_i$ is the total time on test (see Section 3.5). Thus,

$$ \ln L(\lambda) = D \ln \lambda - S, \quad (B.7) $$

$$ U(\lambda) = \frac{d}{d\lambda} \ln L(\lambda) = \frac{D}{\lambda} - S, \quad (B.8) $$

and

$$ I(\lambda) = \frac{d^2}{d\lambda^2} \ln L(\lambda) = \frac{D}{\lambda^2}. \quad (B.9) $$

Solving B.8 for $\lambda$ gives us the maximum likelihood estimator, $\hat{\lambda} = D/S$. Using these statistics,

$$ X^2_S = \left( \frac{D}{S} - 1 \right)^T \frac{D}{S} \left( \frac{D}{S} - 1 \right) = \frac{D^2 - S^2}{S} $$

$$ = 2S - 2D + 2S \ln(D/S) $$

In this case, note that the Wald and Rao tests are identical. All three of these statistics have asymptotic chi-squared distributions with one degree of freedom.

All three test statistics can be used to test composite hypotheses. Suppose the parameter vector $\theta$ is divided into two vectors $\theta_1$ and $\theta_2$, of lengths $p_1$ and $p_2$, respectively. We would like to test the hypothesis

$$ H_0: \theta_1 = \theta_1^0. $$

Here $\theta_1$ is a nuisance parameter. Let $\hat{\theta}(\theta_2)$ be the maximum likelihood estimates of $\theta_1$ obtained by maximizing the likelihood with respect to $\theta_2$, with $\theta_1$ fixed at $\theta_1^0$. That is, $\hat{\theta}(\theta_2)$ maximizes $L(\theta_1, \theta_2) \mid Y$ with respect to $\theta_2$. We also partition the information matrix $I$ into

$$ I = \begin{pmatrix} I_{\theta_1 \theta_1} & I_{\theta_1 \theta_2} \\ I_{\theta_2 \theta_1} & I_{\theta_2 \theta_2} \end{pmatrix}, \quad (B.10) $$

where $I_{\theta_1 \theta_1}$ is of dimension $p_1 \times p_1$, $I_{\theta_2 \theta_2}$ is of dimension $p_2 \times p_2$, and $I_{\theta_1 \theta_2}$ is $p_1 \times p_2$, and $I_{\theta_1 \theta_1}$ is $p_1 \times p_1$. Notice that a partitioned information matrix
has an inverse which is also a partitioned matrix with
\[ I^{-1} = \begin{pmatrix} \mathbf{I}^p & \mathbf{0} \\ \mathbf{0} & \mathbf{I}^q \end{pmatrix} \]  

(B.11)

With these refinements, the three statistics for testing \( H_0: \phi = \phi_0 \) are given by:

- Likelihood ratio test:
  \[ X^2_{LR} = -2 \left( \ln L(\phi_0, \hat{\phi}(\phi_0); \gamma) - \ln L(\hat{\theta}; \gamma) \right), \]
  (B.12)

- Wald test:
  \[ X^2_{W} = (\hat{\phi} - \phi_0) [\mathbf{II}^{p,q} \hat{\phi}(\phi_0)]^{-1} (\hat{\phi} - \phi_0)' , \]
  (B.13)

- Score test:
  \[ X^2_{S} = \mathbf{U}(\phi_0, \hat{\phi}(\phi_0)) [\mathbf{II}^{p,q} \hat{\phi}(\phi_0)]^{-1} \mathbf{U}^t(\phi_0, \hat{\phi}(\phi_0)). \]
  (B.14)

All three statistics have an asymptotic chi-squared distribution with \( m \) degrees of freedom when the null hypothesis is true.

**Example B.2**

Consider the problem of comparing two treatments, where the time to event in each group has an exponential distribution. For population one, we assume that the hazard rate is \( \lambda \) whereas for population two, we assume that the hazard rate is \( \beta \). We shall test \( H_0: \beta = 1 \) treating \( \lambda \) as a nuisance parameter. The likelihood function is given by

\[ L(\lambda, \beta; D_i, D_{ii}, S_i, S_{ii}) \equiv \lambda^{D_i} \beta^{D_{ii}} \exp(-\lambda S_i - \beta S_{ii}) \]

(B.15)

where \( D_i \) is the number of events and \( S_i \) is the total time on test in the \( i \)th sample, \( i = 1, 2 \). From (B.15),

\[ \ln L(\beta, \lambda) = (D_1 + D_2) \ln \lambda + D_2 \ln \beta - \lambda S_1 - \beta S_2, \]

(B.16)

\[ U_\lambda(\beta, \lambda) = \frac{\partial}{\partial \lambda} \ln L(\beta, \lambda) = \frac{D_1 + D_2}{\lambda} - S_1 - \beta S_2, \]

(B.17)

\[ U_\beta(\beta, \lambda) = \frac{\partial}{\partial \beta} \ln L(\beta, \lambda) = \frac{D_1 + D_2}{\beta} - S_1 - \lambda S_2, \]

(B.18)

\[ U_{\lambda \beta}(\beta, \lambda) = -\frac{\partial^2 \ln L(\beta, \lambda)}{\partial \beta \partial \lambda} = \frac{D_2}{\beta^2}, \]

(B.19)

\[ U_{\lambda \lambda}(\beta, \lambda) = -\frac{\partial^2 \ln L(\beta, \lambda)}{\partial \lambda^2} = \frac{D_1 + D_2}{\lambda^2}, \]

(B.20)

and

\[ U_{\beta \beta}(\beta, \lambda) = -\frac{\partial^2 \ln L(\beta, \lambda)}{\partial \beta^2} = S_1. \]

(B.21)

Solving the system of equations \( U_\lambda(\beta, \lambda) = 0 \) and \( U_{\lambda \beta}(\beta, \lambda) = 0 \) yields the global maximum likelihood estimators \( \hat{\beta} = S_1 D_2 / (S_1 S_2) \) and \( \hat{\lambda} = D_1 / S_1. \)

Appendix B Large-Sample Tests Based on Likelihood Theory

Solving \( U_\lambda(\beta, \lambda) = 0 \), for \( \beta \) fixed at its value under \( H_0 \), yields \( \hat{\lambda}(\beta = 1) \), denoted by \( \hat{\lambda}(1) = (D_1 + D_2)/(S_1 + S_2) \). Thus, we have from B.12 a likelihood ratio test statistic of

\[ X^2_{LR} = -2 \left[ \ln L(D_1 + D_2) - \ln L(S_1 + S_2) \right] - \left[ (D_1 + D_2) \ln \hat{\lambda} + D_1 \ln \beta - \hat{\lambda} S_1 - \hat{\beta} S_2 \right]. \]

(B.22)

\[ = 2D_1 \ln \left[ \frac{D_1(S_1 + S_2)}{S_1(D_1 + D_2)} \right] + 2D_2 \ln \left[ \frac{D_2(S_1 + S_2)}{S_2(D_1 + D_2)} \right]. \]

From B.19-B.21,

\[ f^{\text{LR}}(\beta, \lambda) = \frac{\partial^2 L(D_1 + D_2)}{\partial \beta^2} \left( \hat{\beta}(1) \right) \]

so the Wald test is given by

\[ X^2_{W} = (\hat{\beta} - 1)^2 \left[ \frac{\partial^2 L(D_1 + D_2)}{\partial \beta^2} \left( \hat{\beta}(1) \right) \right]^{-1} \]

\[ = \frac{D_2(S_1 + S_2) - (\hat{\lambda}(1) S_2)^2}{D_2 S_1^2 (D_1 + D_2)} \]

The score test is given by

\[ X^2_{S} = (\hat{\lambda}(1) S_2)^2 \left( \frac{D_1 + D_2}{D_2 D_1 + D_2 - (\hat{\lambda}(1) S_2)^2} \right) \]

\[ = \frac{(D_1 + D_2)(S_1 + S_2) - (D_1 + D_2) S_2^2}{D_2 S_1^2 + D_1 S_2^2 - (D_1 + D_2) S_2^2}. \]

If, for example, \( D_1 = 10, D_2 = 12, S_1 = 25, \) and \( S_2 = 27 \), then \( X^2_{LR} = 0.0607, X^2_{W} = 0.0545 \) and \( X^2_{S} = 0.0448 \), all nonsignificant when compared to a chi-square with one degree of freedom.
Solutions to Chapter 13

13.1  
\[ T = 14.8, \ V = 107.5, \ Z = 1.4, \ \rho = 0.1539. \]  
No evidence of random effect.

13.3  
(a) Standard Cox model \( b = -1.035, \ \text{SE}(b) = 0.44, \ \rho = 0.0187. \)

(b) Gamma Frailty Model \( b = -1.305, \ \text{SE}(b) = 0.528, \ \rho = 0.0133. \)  
Estimate of \( \theta = 0.713, \ \text{SE} = 0.622, \ \text{Wald } p\text{-value of test of } \theta = 0 = 0.2517, \ \text{likelihood ratio } p\text{-value} = 0.1286.

13.5  
(a) See 13.1.

(b) Adjusted \( SE = 0.3852, \) test statistic \( = -1.035/0.3852, \ \rho = 0.0072.

Bibliography


Bibliography


Peto, R. and Pike, M. C. Conservation of the Approximation $\int_{-\infty}^{0} - e^{-\frac{x^2}{2}} \, dx$ in the Log Rank Test for Survival Data or Tumor Incidence Data. Biometrica 29 (1973): 579-584.

Author Index

Aalen, O. O., 79, 94, 100, 215, 529, 343, 345, 346, 427, 515
Aisbett, C. W., 427, 439, 522
Akaike, H., 277, 515
Altman, S. A., 162, 525
Amato, D. A., 436, 427, 437, 521
American Joint Committee for Cancer Staging, 9, 515
Anbari, R., 2, 519
Anderson, J., 6, 7, 523, 524
Anderson, K. M., 74, 523
Anja, E., 368, 516
Aikin, K., 4, 517
Aitken, R. R., 4, 516, 517
Ayres, L. W., 7, 520
Bain, L. J., 44, 524
Barlow, R. E., 380, 516
Barnell, L. M., 19, 515, 521
Batchelor, J. R., 438, 516
Bay, W. H., 6, 523
Beadle, G. F., 2, 519
Bemler, J. N., 45, 517
Bey, W. H., 39, 516
Bie, O., 106, 108, 117, 516
Bigs, J. C., 4, 517
Billingsley, P., 227, 516
Bohdar, R. N., 429, 522
Bolnick, L., 18, 516
Breslow, N. E., 183, 205, 208, 259, 516
Brodsky, L., 4, 516, 517
Brooksmeier, R., 120, 122, 231, 234, 517
Brown, J. B. W., 100, 517
Buckley, J. D., 183, 517
Bulova, S. I., 4, 517
Burgert, E. O., 2, 519
Campo, R., 380, 516
Chang, C. L., 4, 140, 517
Chang, Y.-C., 305, 522
Chevtelovsky, D., 214, 521
Christensen, E., 307, 517
Christensen, R., 197, 520
Chung, C. F., 117, 517
Cimina, J. L., 429, 522
Clausen, R., 45, 517
Clayton, D., 426, 436, 517
Cleveland, W. S., 361, 517
Collett, D., 419, 517
Conn, S., 18, 516
Commenges, D., 427, 517
Contal, C., 273, 274, 517
Censoring, See Random Right Censoring
Conditional Probability Function, 128
Crude Probabilities, See Cumulative Incidence Functions
Cumulative Incidence Functions
Definition, 52-53
Estimation, 127-128
Definition, 50
Identifiability, 51-52
Net Probabilities, 52-53
Partial Crude Probabilities, 52-53
Confidence Bands for the Cumulative Hazard Rate, 114-115, 116-118
for the Survival Function, 109-114, 116-117
Confidence Intervals for the Cumulative Hazard Rate, 106-108
for the Mean, 119
for the Median, 121-123
for the Survival Function, 104-108
Conditional Failure Rate, See Hazard Function
Constant Excess Mortality Model, 185-186
Counting Process, 79-86, 205, 215, 220
Counting Process Martingale, 81-82
Cox Regression Model, See Proportional Hazards Regression
Cox-Snell Residuals, 354-358, 413-418
Cumulative Hazard Function, 27, 31-32
Confidence Bands, 114-117
Confidence Interval (Pointwise), 106-108
Point Estimator, See Nelson-Aalen Estimator
Cumulative Intensity Process, 77
Cramer-Von Mises Test, See Tests of Hypotheses
Delayed Entry, See Left Truncation
Delta Method, 396, 398-401
Deviance Residuals, 354, 381-385, 416-419
Dirichlet Process, See Bayes Methods
Double Censoring, 17, 71, 146-148
Dummy Variables, 246
Efficient Scores, 254-255
EM Algorithm, 432-433, 435-436
EP Bands, 109-111, 114-118
Epanechnikov Kernel, See Kernel Smoothing
Excess Mortality, 177-187
Exponential Distribution, 33, 37-38, 45-46, 49, 406-410
Likelihood Function, 79-77
Exponential Power Distribution, 37, 44, 49
Extreme Value Distribution, 39, 47, 394-395
Filtration, See History
Fleming and Harrington Tests, See Tests of Hypotheses
Force of Mortality, See Hazard Function
Frailty Model, 425-427
Estimation for the Gamma Frailty Model, 430-436
Gamma Distribution, 37, 41-43, 59
Gastric Cancer, 224-225
Generalized Gamma Distribution, 37, 44, 394, 406-409
Gompertz Distribution, 37, 44, 58-59
Greenwood’s Formula, 92, 100-101, 124, 125
Guarantee Time, 44, 59
Hall-Wellner Bands, 111-117
Harrington and Fleming Tests, See Tests of Hypotheses
Hazard Function, 27-32, 35-36
Kernel Smoothed Estimators, See Kernel Smoothing
Life Table Estimator, 153-157
Hazard Plotting, 31, 43-44, 335, 355, 395, 409-414
Heart Disease, 72
History, 80, 314-315
Indicator Variables, 246
Information Matrix, 254-255
Intensity Process, 81
Interaction, 251, 268, 269
Interval Censoring, 18, 71-72, 75, 143-146
Inverse Gaussian Distribution, 37, 44
Kaplan-Meier Estimator, See Product Limit Estimator
Kernel Smoothing, 166-177
for the Additive Hazard Model, 332, 339-340, 343
Bicweight Kernel, 167-168, 171-172, 174-175, 337-338, 341
Cross Validation, 172-175
Epanechnikov Kernel, 167-176
Mean Integrated Squared Error, 172-175, 177
Tail Adjustment, 167-169
Uniform Kernel, 166, 171-172, 174
Kidney Dialysis, 6
Kidney Transplantation, 8-9
Kolmogorov-Smirnov Tests, See Tests of Hypotheses, Bayes Tests
Laryngeal Cancer, 9-10
Left Censoring, 70-71, 74, 140-141
Left Truncation, 16, 63-64, 72-74, 122-126, 212, 312-315
Leukemia Alternative, 48, 215
Leukemia, 2-5, 10-12
Lexis Diagram, 64-66
Life Tables, 60, 140, 178-180
Bayes Method, 197
Construction, 152-158
Likelihood Construction, 70-75, 86-87
Likelihood Ratio Tests
Definition, See Tests of Hypotheses for Proportional Hazards Model, See Proportional Hazards Regression
Linear Model Regression, See Accelerated Failure Time Model
Linear Models, 37, 39-41, 58-59, 394, 405-409, 413
Log Normal Distribution, 37, 39-41, 58-59, 394, 405-409, 413
Log Rank Test, See Tests of Hypotheses
Main Effects, 268
Marginal Model, 436-438
Marijuana Use, 17
Martingale, 81-82
Mantingale Central Limit Theorem, 85-86
Martingale Residual, 354, 359-363, 362-384, 416-418
Matched Pairs, See Tests of Hypotheses, Stratified Tests
Maximization, Numerical Methods for Multivariate, 445-446
Univariate, 443-445
Mean Residual Life, 52-56
Mean Survival Time, 33-36
Estimator, 117-119, 123
Median, 33
Estimator, 119-123
Median Residual Life, 35
Life Table Estimator, 154-157
Median Test, See Tests of Hypotheses
Mill’s Ratio, See Hazard Function
Model Checking for the Accelerated Failure Time, Model, See Accelerated Failure, Time Model, for the Exponential Distribution, See Hazard Plotting, for the Log Logistic Distribution, See Hazard Plotting, for the Log Normal Distribution, See Hazard Plotting, for the Proportional Hazards Model, See Proportional Hazards Model, for the Weibull Distribution, See Hazard Plotting, Mont Carlo Bayesian Method, See Bayes Methods, Gibbs Sampler
Multiplicative Hazards Model
Definition, 47-48
Multistate Modeling, See Proportional Hazards Regression, Survival Synthesis

Subject Index

Pointwise Comparisons, 234-237
Tarone and Ware Test, 208-211, 213-214, 219
Large Sample Inference
Likelihood Ratio Tests, 450-453
Score Tests, 450-453
Wald Tests, 450-453
One Sample Tests, 202-205
Log Rank Test, 202-205
Hanington Fleming Test, 203
Wicoxon Test, 205
Renyi Tests, 223-228
Stratified Tests, 219-223
Two Sample Tests
Based on the Additive Hazards Regression Model, 347-348, 350
Cramer-Von Mises Test, 227-228, 229
Median Test, 231-233, 234-235
Truncation, 219
Weighted Difference in Kaplan Meier Estimator Test, 228-230, 233-234
Trend Tests, 216-219, 262
Time Dependent Covariates, See Proportional Hazards Regression

Likelihood, Bayesian and MCMC Methods in Quantitative Genetics
DANIEL SORENSEN and DANIEL GIANOLA
This book provides the foundations of likelihood, Bayesian and MCMC methods in the context of genetic analysis of quantitative traits. Although a number of excellent texts in these areas have become available in recent years, the basic ideas and tools are typically described in a technically demanding style. Considerably more detail is offered than is usual. The book reviews probability and distribution theory. Part II presents methods of inference and MCMC methods. Part IV discusses several models that can be applied in quantitative genetics, primarily from a Bayesian perspective.
STATISTICS FOR BIOLOGY AND HEALTH

ActiveEpi
DAVID G. KLEINBAUM
ActiveEpi is a complete multimedia presentation of the material commonly found in an introductory epidemiology course on CD-ROM. ActiveEpi includes 15 lessons and over 50 hours of content. It can be used in a variety of teaching formats: distance learning, self-paced learning, on-campus courses, and short courses. It uses video to instill real-world motivation for learning and applying epidemiologic concepts; narrated exposition to teach the concepts using high-quality audio synchronized with text, pictures and narration; options to provide self-assessment of basic terms and concepts; data desk to provide practice applying the methods learned to real data; navigational tools such as a dynamic table of contents, index and glossary, to allow easy-to-use methods for working through the content.
SPRINGER TEXTS IN STATISTICS