General Interest Section

Maximum Likelihood Estimation and Mathematica

By IAIN D. CURRIE†

Heriot-Watt University, Edinburgh, UK

[Received February 1994. Revised September 1994]

SUMMARY

Data and outline Mathematica code are given for several examples of maximum likelihood estimation. A common approach is taken to both elementary complete data problems and more computationally demanding incomplete data problems. In teaching, this common approach brings many conceptually simple but computationally heavy problems within reach of the student. Published examples on truncation, grouping, censoring, survival analysis and generalized linear models are considered.

Keywords: Censoring; Generalized linear models; GLIM; Grouping; Incomplete data; Mathematica; Maximum likelihood; Survival analysis; Truncating

1. Introduction

We show how several maximum likelihood problems can be solved by computer algebra. We use Mathematica (Wolfram, 1991) but almost any computer algebra program would suffice. Mathematica is an attractive tool for maximum likelihood estimation since a wide range of problems can be tackled with a common method; each section of the paper gives examples that show how Mathematica is used for such estimation. Section 2 uses two elementary problems (one discrete and one continuous) to establish the basic approach that is used with Mathematica. Section 3 applies the method to problems involving censoring, truncating and grouping; an example is taken from Yule and Kendall (1968) to illustrate estimation with grouped normal data, whereas an example discussed by Brass (1958) is used to demonstrate the method for the truncated negative binomial distribution. Section 4 shows how generalized linear models can be fitted, our examples here being a binary regression with logit link and an analysis of a dilution assay; the data come from Bliss (1935) and Wetherill (1981) respectively. Section 5 looks at two research problems. The first shows how Mathematica is used to deal with a gamma model with incomplete data and presents a problem used by Meng and Rubin (1993); the second is a

†Address for correspondence: Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Riccarton, Edinburgh, EH14 4AS, UK.
E-mail: iain@ma.hw.ac.uk

© 1995 Royal Statistical Society
quadratic logistic model discussed by Dellaportas and Smith (1993). Some conclusions are given in Section 6.

2. Simple Examples

In this section we consider two elementary examples. The first shows how to deal with a discrete distribution (Poisson) whereas the second uses a continuous distribution (normal). We begin with estimation for the Poisson distribution. Table 1 gives the Bortkewitsch (1898) data on deaths from the kick of a horse in 10 Prussian Army corps over the 20-year period 1875–94. As is usual for data of this sort, the 200 observations are presented in a frequency table.

Mathematica consists of a core program, known as the kernel, and many more specialized functions organized in packages. We begin by loading two of the statistical packages with

```mathematica
Get["Statistics'DiscreteDistributions""]
Get["Statistics'ContinuousDistributions"]
```

To obtain the likelihood for the Poisson model for these data in Mathematica we proceed as follows. First, we use the discrete distributions package and obtain the probability distribution with

```mathematica
prob = Table[PDF[PoissonDistribution[t], r], {r, 0, 4}]
```

Alternatively, we could supply the probability function directly with

```mathematica
prob = Table[E^-t t^r/r!, {r, 0, 4}]
```

In this paper we use the statistical packages, but, whichever method is used, Mathematica responds with a list of the Poisson probabilities in familiar algebraic form:

```mathematica
Out[4] = \{E^{-t}, \frac{t}{E^t}, \frac{t^2}{2 E^t}, \frac{t^3}{6 E^t}, \frac{t^4}{24 E^t}\}
```

Mathematica numbers its input and output statements; Out[4] refers to the output from the fourth input line. Second, we supply the data with

```mathematica
freq = \{109, 65, 22, 3, 1\}
```

The output from

```mathematica
logl = Apply[Plus, freq Log[prob]]
```

gives the log-likelihood in algebraic form

```
```

<table>
<thead>
<tr>
<th>TABLE 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency of deaths from the kick of a horse</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>No. of deaths</td>
</tr>
<tr>
<td>Frequency</td>
</tr>
</tbody>
</table>
\[
\text{Out[6]} = 109 \log[E^{-1}] + 65 \log\left[\frac{t}{E_t}\right] + 22 \log\left[\frac{t^2}{2 E_t}\right] + 3 \log\left[\frac{t^3}{6 E_t}\right] + \log\left[\frac{t^4}{24 E_t}\right]
\]

Out[4] and Out[6] are examples of the symbolic output that characterizes Mathematica (and programs like it). The two examples give an indication of the kind of expression that Mathematica produces. From now on, we generally omit these.

One of the attractions of a program like Mathematica is the ease with which plots can be produced. The command

\[
\text{Plot[\logl, \{t, a, b\}]}
\]

produces a plot of \logl where the plot parameters \text{a} and \text{b} determine the range of the plot. A little experimentation with \text{a} and \text{b} will give a satisfactory plot; Fig. 1 was produced with \text{a} = 0.3 and \text{b} = 0.8.

The maximum likelihood estimate of the Poisson parameter \text{t} is obtained with

\[
\text{mle = FindMinimum[-logl, \{t, 1\}]}
\]

The \text{FindMinimum} function uses the method of steepest descent with derivatives computed algebraically. \text{FindMinimum} returns a list with two entries: the first element of the list is the minimum value of \text{-logl} and the second is the maximum likelihood estimate of \text{t}. The initial value of 1 in equations (5) can be taken from Fig. 1. The information function is found symbolically with

\[
\text{information = -D[logl, t, t]}
\]

and the standard error of the maximum likelihood estimate of \text{t} follows with

Fig. 1. Log-likelihood for the Poisson model
TABLE 2
Hours sleep of 10 patients with and without a sedative

<table>
<thead>
<tr>
<th>Patient</th>
<th>Amount of sleep (h) for the following treatments:</th>
<th>Difference (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sedative</td>
<td>None</td>
</tr>
<tr>
<td>A</td>
<td>1.3</td>
<td>0.6</td>
</tr>
<tr>
<td>B</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>C</td>
<td>6.2</td>
<td>2.5</td>
</tr>
<tr>
<td>D</td>
<td>3.6</td>
<td>2.8</td>
</tr>
<tr>
<td>E</td>
<td>4.9</td>
<td>2.9</td>
</tr>
<tr>
<td>F</td>
<td>1.4</td>
<td>3.0</td>
</tr>
<tr>
<td>G</td>
<td>6.6</td>
<td>3.2</td>
</tr>
<tr>
<td>H</td>
<td>4.5</td>
<td>4.7</td>
</tr>
<tr>
<td>I</td>
<td>4.3</td>
<td>5.5</td>
</tr>
<tr>
<td>J</td>
<td>6.1</td>
<td>6.2</td>
</tr>
</tbody>
</table>

Fig. 2. Contour plot of the log-likelihood for the normal model
where the symbol \( \div \) means replace the variable \( t \) with its value in \( \text{mle} \).

Our second example involves the normal distribution. The method follows that laid out in equations (1), (3) and (5)–(7). Table 2 gives some data provided by Snedecor and Cochran (1967), pages 43–44, on the amount of sleep of 10 patients troubled with sleeplessness. We suppose that the differences in the amount of sleep with and without a sedative follow a normal distribution \( N(m, s) \) with mean \( m \) and standard deviation \( s \). The process of finding the maximum likelihood estimates of \( m \) and \( s \) is only slightly different from that for the Poisson data. We can read in the data with

\[
data = \{0.7, 0.0, 3.7, 0.8, 2.0, -1.6, 3.4, -0.2, -1.2, -0.1\}
\]

and evaluate the normal density function at each data point by

\[
\text{pdf} = \text{PDF}[	ext{NormalDistribution}[m, s], \text{data}]
\]

The likelihood is obtained as in equation (3) by

\[
\text{logl} = \text{Apply}[\text{Plus}, \text{Log[pdf] }]
\]

The maximum likelihood estimates of the normal parameters are obtained with the two-parameter version of equation (5):

\[
\text{maxlogl} = \text{FindMinimum}[-\text{logl}, \{m, 0\}, \{s, 1\}]
\]

\[
\text{mle} = \text{maxlogl[2]}
\]

Initial values of 0 and 1 have been supplied for the mean \( m \) and standard deviation \( s \). Finally, the standard errors of \( m \) and \( s \) are obtained as in equations (6) and (7).

\[
i_{11} = -D[\text{logl}, m, m]
i_{12} = -D[\text{logl}, m, s]
i_{22} = -D[\text{logl}, s, s]
\]

\[
\text{information} = \{\{i_{11}, i_{12}\}, \{i_{12}, i_{22}\}\}
\]

\[
\text{variance} = \text{Inverse}[\text{information}] \div \text{mle}
\]

\[
\text{standarderrorofm} = \text{Sqrt[variance[1,1]]}
\]

\[
\text{standarderrorofs} = \text{Sqrt[variance[2,2]]}
\]

Mathematica can produce contour plots or three-dimensional plots; Fig. 2, a contour plot of \( \text{logl} \) with specified contour levels, is produced with

\[
\text{summit} = -\text{maxlogl[1]}
\]

\[
\text{contours} = \text{summit} - \{0.5, 1, 1.5, 2, 2.5, 3, 3.5\}
\]

\[
\text{ContourPlot}[\text{logl}, \{m, -1, 2.5\}, \{s, 1, 3.5\}, \text{Contours} \rightarrow \text{contours}]
\]

These two examples show the general approach used to obtain maximum likelihood estimates with Mathematica:

(a) the elements of the likelihood function are found (equation (1) or (8));
(b) the log-likelihood is obtained (equation (3) or (9));
(c) if required, a plot is produced (equation (4) or (12));
(d) the maximum likelihood estimate(s) are obtained with the general minimization routine (equation (5) or (10));

(e) standard errors are found (equations (6) and (7) or equation (11)).

We do not suggest that maximum likelihood estimation should always be tackled in this way. Clearly, there is no attempt to make use of any special structure that a problem might have, and this puts a severe practical limit on the size of the problem that can be tackled. Nevertheless, some non-trivial problems can still be solved by using little more than the simple Mathematica code given above. From a conceptual point of view, this unified approach is attractive, particularly in teaching. In the rest of the paper we give a variety of examples which demonstrate this point.

3. Censored, Truncated and Grouped Data

Several papers appeared in the 1950s on maximum likelihood estimation where the sample came from a truncated distribution. Plackett (1953) and Moore (1954) considered the Poisson distribution whereas Sampford (1955) and Brass (1958) looked at the problem when the sampled distribution was negative binomial. The aim of these papers was to find efficient alternatives to maximum likelihood since, at that time, the exact solution of the maximum likelihood equations and the evaluation of the information matrix were daunting arithmetical tasks. Cohen has published a series of papers on estimation with censored samples; Cohen (1991), on truncated and censored samples, summarizes much of the work in this area.

Yule and Kendall (1968) presented many examples of grouped frequency distributions. They used Sheppard's corrections in the calculation of moments from grouped data. Although today's data handling capabilities do not require that data be grouped, grouped data still arise and present fitting problems similar to those that arise for censored and truncated data. In this section we show that the direct approach laid out in Section 2 can be applied to estimation with censored, truncated or grouped data.

Table 3 is taken from p. 82 of Yule and Kendall (1968) and gives the heights of 8585 adult males in the British Isles; the data were collected in 1883. We suppose that a normal distribution N[μ, σ] is to be fitted to the data. The measurements

<table>
<thead>
<tr>
<th>Height (in)</th>
<th>Frequency</th>
<th>Height (in)</th>
<th>Frequency</th>
<th>Height (in)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>57-</td>
<td>2</td>
<td>64-</td>
<td>669</td>
<td>71-</td>
<td>392</td>
</tr>
<tr>
<td>58-</td>
<td>4</td>
<td>65-</td>
<td>990</td>
<td>72-</td>
<td>202</td>
</tr>
<tr>
<td>59-</td>
<td>14</td>
<td>66-</td>
<td>1223</td>
<td>73-</td>
<td>79</td>
</tr>
<tr>
<td>60-</td>
<td>41</td>
<td>67-</td>
<td>1329</td>
<td>74-</td>
<td>32</td>
</tr>
<tr>
<td>61-</td>
<td>83</td>
<td>68-</td>
<td>1230</td>
<td>75-</td>
<td>16</td>
</tr>
<tr>
<td>62-</td>
<td>169</td>
<td>69-</td>
<td>1063</td>
<td>76-</td>
<td>5</td>
</tr>
<tr>
<td>63-</td>
<td>394</td>
<td>70-</td>
<td>646</td>
<td>77-</td>
<td>2</td>
</tr>
</tbody>
</table>
leading to Table 3 were made to the nearest 1/8th of an inch, so we can set up the class boundaries with

\[
\text{leftend } = \text{Table}[56 + i, \{i, 1, 21\}] - 1/16
\]
\[
\text{rightend } = \text{leftend } + 1
\]

The probabilities associated with each class interval are found similarly to equation (8):

\[
\text{prob } = \text{CDF[NormalDistribution[m, s], rightend]}
\]
\[
- \text{CDF[NormalDistribution[m, s], leftend]}
\]

We finish the calculation by using the method laid out in Section 2. The data are supplied with

\[
\text{freq } = \{2, 4, 14, 41, 83, 169, 394, 669, 990, 1223, 1329, 1230, 1063, 646, 392, 202, 79, 32, 16, 5, 2\}
\]

The likelihood function \( \log l \) is found with equation (3) and the maximum likelihood estimates of \( m \) and \( s \) similarly to equation (10). Yule and Kendall (1968) gave the mean of these data as 67.458 in and the Sheppard corrected standard deviation as 2.556 in; the maximum likelihood estimates are also 67.458 in and 2.556 in respectively. We can also find the standard errors of our estimates; formulae (11) apply without alteration and give 0.028 in and 0.020 in as the standard errors of the estimates of \( m \) and \( s \) respectively. We can now also give a measure of the loss of information that has resulted from the grouping. We assume that the population standard deviation is 2.556 in and ask, if there is no grouping, what sample size would give a standard error of the mean of 0.028 in? This gives a sample size of 8477; the loss of information is very small, about 1%. A similar calculation for \( s \) gives an equivalent sample size of 8371; the loss of information is slightly greater at 2.5%, but still very small.

For our second example in this section we use the data in Table 4. These data were used by Brass (1958) to illustrate his simplified method for fitting the truncated negative binomial distribution to data. The observations are the number of children ever born to a particular sample of mothers over 40 years old.

The fitting method is almost identical with that described in Section 2 for a sample from the (complete) Poisson distribution. It is convenient to define

\[
\text{negbin } = \text{NegativeBinomialDistribution[r, p]}
\]

and then the truncated probabilities are simply given by

\[
\text{prob } = \text{Table[PDF[negbin, i], \{i, 1, 12\}]/(1 - PDF[negbin, 0])}
\]

The list \( \text{freq} \) is obtained from Table 4 and the maximum likelihood estimates are found with equations (3) and (10); initial estimates of 1 for \( r \) and 0.5 for \( p \) are

<table>
<thead>
<tr>
<th>TABLE 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of children of 340 African mothers</td>
</tr>
<tr>
<td>No. of children</td>
</tr>
<tr>
<td>No. of mothers</td>
</tr>
</tbody>
</table>
sufficient. A contour plot of the log-likelihood can be obtained with equation (12) and standard errors follow as in equations (11). The values found by Brass in 1958 are confirmed.

4. Generalized Linear Models

Generalized linear models (GLMs) are a large and important class of models that are conveniently fitted by GLIM (Francis et al., 1993) (and many other statistical packages). The GLIM algorithm takes full advantage of the structure of GLMs and models with many parameters are efficiently fitted. The maximum likelihood aspect of the model fitting process is scarcely noticed by the user since it only appears in the $\texttt{error}$ statement. Mathematica allows GLMs with small numbers of parameters to be fitted in the same way as the models in Sections 2 and 3. We demonstrate by looking at two models. The first is a standard binary regression with logit link. The second example is a dilution assay which is fitted with the help of the $\texttt{offset}$ directive in GLIM; in Mathematica the standard method applies directly.

The data in Table 5 were described by Bliss (1935) and give the number of beetles killed after exposure to various concentrations of carbon disulphide ($\text{CS}_2$). A preliminary plot (which could easily be done in Mathematica) suggests that a standard binary regression with the logit link describes the data very well. We set up the data lists as follows: dose $= \{1.6907, 1.7242, \ldots\}$, exposed $= \{59, 60, \ldots\}$ and killed $= \{6, 13, \ldots\}$. The probabilities of a death at each dose are given by

$$p = \exp(b_0 + b_1 \text{dose})/(1 + \exp(b_0 + b_1 \text{dose}))$$  

(13)

The contribution to the likelihood at each dose level is given by

$$\text{prob} = \text{Table}[
\text{PDF}[
\text{BinomialDistribution}[
\text{exposed}[[k]], \ p[[k]]], \ \text{killed}[[k]]],
\{k, 1, 8\}\]$$  

(14)

As before, we find the log-likelihood with equation (3). The maximum likelihood estimates and their standard errors are found with equations (10) and (11); initial values for $b_0$ and $b_1$ can both be taken as 0.

The data in Table 6 are from a dilution assay described by Wetherill (1981), p. 95. The concentrations in the suspensions $S_0$, $S_1$, $S_2$, $S_3$ and $S_4$ are taken to be $\rho$.

<table>
<thead>
<tr>
<th>Dose (log$_{10} \text{CS}_2\text{ mg l}^{-1}$)</th>
<th>No. of insects</th>
<th>No. killed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6907</td>
<td>59</td>
<td>6</td>
</tr>
<tr>
<td>1.7242</td>
<td>60</td>
<td>13</td>
</tr>
<tr>
<td>1.7552</td>
<td>62</td>
<td>18</td>
</tr>
<tr>
<td>1.7842</td>
<td>56</td>
<td>28</td>
</tr>
<tr>
<td>1.8113</td>
<td>63</td>
<td>52</td>
</tr>
<tr>
<td>1.8369</td>
<td>59</td>
<td>53</td>
</tr>
<tr>
<td>1.8610</td>
<td>62</td>
<td>61</td>
</tr>
<tr>
<td>1.8839</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>
\(p/2, p/4, p/8\) and \(p/16\) respectively. In GLIM the standard dilution assay model is fitted with a binomial error and a complementary log-log-link (Francis et al. (1993), p. 572). The regressor variable corresponding to the dilution appears in the linear predictor with a known coefficient which GLIM handles with the \$offset directive. The maximum likelihood estimate of \(p\) can be obtained in Mathematica in a more transparent way. If \(X_k\) is the number of samples out of \(n_k\) which contain particles then

\[X_k \sim B(n_k, 1 - \exp(-2^{-k}p)), \quad k = 0, 1, 2, 3, 4.\]

With \(x = \{10, 9, \ldots\}\) and \(n = \{10, 12, \ldots\}\) the binomial probabilities are given as in equation (13) by

\[p = \text{Table}[1 - \text{Exp}[-2^r(-k)\rho]], \{k, 0, 4\}\]

and the contributions to the likelihood function by

\[\text{prob} = \text{Table}[	ext{PDF}[\text{BinomialDistribution}[n[\{r\}], p[\{r\}]], x[\{r\}]], \{r, 1, 5\}\]

The model is fitted by following the recipe laid out in equations (3) and (5)-(7).

5. Further Examples

This section looks at two examples from recent published research. The first shows how Mathematica is used to deal with a gamma model with incomplete data. This example was used by Meng and Rubin (1993) to illustrate the expectation-conditional maximization (ECM) algorithm. The ECM algorithm is an extension of the EM algorithm (Dempster et al., 1977) and can be used when the maximization step of the EM algorithm does not have an explicit solution. The ECM algorithm might be used in survival analysis when the lifetimes follow a gamma distribution. The standard parametric models in survival analysis are exponential, Weibull and extreme value (Aitkin and Clayton, 1980); these models give rise to proportional hazards and can be fitted by using GLIM. Gamma models do not fit into the proportional hazards framework and it is difficult to see how they could be fitted in GLIM. With Mathematica we can fit both the standard proportional hazards models and gamma models with the same direct approach.

The data in Table 7 were discussed by Gehan (1965) and used by Aitkin and

<table>
<thead>
<tr>
<th>Suspension</th>
<th>No. of unit volumes observed</th>
<th>No. containing particles</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>S1</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>S2</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>S3</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>S4</td>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>

TABLE 6
Dilution assay
TABLE 7
Times of remission of leukaemia patients

<table>
<thead>
<tr>
<th>Times of remission (weeks) for the following samples:</th>
<th>Sample 1 (drug)</th>
<th>Sample 2 (placebo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6†</td>
<td>10†</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>11†</td>
<td>23†</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>25†</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>32†</td>
</tr>
<tr>
<td>7</td>
<td>17†</td>
<td>32†</td>
</tr>
<tr>
<td>9†</td>
<td>19†</td>
<td>34†</td>
</tr>
<tr>
<td>10</td>
<td>20†</td>
<td>35†</td>
</tr>
</tbody>
</table>

† Censored.

Clayton (1980) to illustrate their method. The likelihood function for data of the kind given in Table 7 is

\[
\prod_{i=1}^{n} f_i(t_i) \prod_{i=n+1}^{n+m} S_i(t_i) \]

(15)

where \(t_1, \ldots, t_{n+m}\) are the survival times of the \(n + m\) individuals, the last \(m\) of whom are right censored, and \(f_i(t)\) and \(S_i(t) = 1 - F_i(t)\) are the density function and survivor function respectively of the \(i\)th individual. In Mathematica, both \(f_i(t)\) and \(S_i(t)\) are available for the exponential, Weibull, extreme value and gamma distributions and simple censored survival models can be fitted by using the method laid out earlier. For example, one of the models fitted by Aitkin and Clayton (1980) to the data in Table 7 is a Weibull model with hazard functions \(\alpha t^{\gamma-1} \exp(\beta_0)\) and \(\alpha t^{\gamma-1} \exp(\beta_0 + \beta_1)\). The Mathematica functions in expression (15) are supplied with functions like

\[
\text{tailW1}[t_] := 1 - \text{CDF}[\text{WeibullDistribution}[a, \text{E}^{-b0/a}], t] 
\]

(16)

where the ‘\(^{-}\)’ character indicates that \(\text{tailW1}[t]\) is a user-defined function with dummy argument \(t\); the parameterization chosen matches the proportional hazards parameterization used in Aitkin and Clayton (1980). The maximum likelihood estimates of \(\alpha, \beta_0\) and \(\beta_1\) are now obtained by minimizing the log-likelihood in the usual way. One advantage of Mathematica is that the whole log-likelihood is considered and hence the standard errors of \(\alpha, \beta_0\) and \(\beta_1\) are obtained directly as in equations (11). In GLIM there is the minor complication that the estimation of \(\beta_0\) and \(\beta_1\) is conditional on \(\alpha\), and hence additional work is required to obtain the correct standard errors.

The fitting of the exponential, Weibull and extreme value models to the data in Table 7 is straightforward in Mathematica but there is a complication with the gamma model; we give some details of the solution for the gamma models since this not only lays out the method for the other models but also shows some other features of Mathematica. We suppose that the data are supplied in the data lists \(\text{times} = \{6, 6, \ldots\}\), \(\text{group} = \{1, 1, \ldots, 2, 2\}\) and \(\text{censor} = \{0, 1, \ldots\}\) where 0 codes for a censored value and 1 for a value observed exactly. We suppose that we wish to fit a model with gamma survival times where the distributions of the survival
times in the two populations have a common shape parameter $a$ and mean survival times $m_1$ and $m_2$ respectively. The density and survival functions in expression (15) are given by

\[
\begin{align*}
t\text{ailG}_1(t) & := 1 - \text{CDF}[	ext{GammaDistribution}[a, m_1/a], t] \\
p\text{dfG}_1(t) & := \text{PDF}[	ext{GammaDistribution}[a, m_1/a], t] \\
t\text{ailG}_2(t) & := 1 - \text{CDF}[	ext{GammaDistribution}[a, m_2/a], t] \\
p\text{dfG}_2(t) & := \text{PDF}[	ext{GammaDistribution}[a, m_2/a], t]
\end{align*}
\] (17)

We now define a function which will find the contribution to the likelihood function for an individual survival time, $z$, according to the group value $y$ and the censoring value $x$:

\[
e\text{lement}[x_, y_, z_] := \text{Which}[x == 1 \&\& y == 1, p\text{dfG}_1[z], x == 0 \&\& y == 1, t\text{ailG}_1[z], x == 1 \&\& y == 2, p\text{dfG}_2[z], x == 0 \&\& y == 2, t\text{ailG}_2[z]]
\]

Finally, we use the Mathematica function \text{MapThread} to apply the function \text{element} to the three data lists \text{times}, \text{group} and \text{censor}; the components of the likelihood function result:

\[
\text{components} = \text{MapThread}[\text{element}, \{\text{censor}, \text{group}, \text{times}\}]
\]

The log-likelihood function is given as usual by

\[
\logl = \text{Apply}[\text{Plus}, \text{Log}[\text{components}]]
\]

The log-likelihood appropriate for exponential, Weibull and extreme value models can be obtained by making the obvious changes to expressions (17) and the \text{FindMinimum} function can then be used to estimate the parameters in the models in the usual way (equations (5) and (10)); the standard errors follow with a simple extension of equations (11). For the gamma distribution the \text{FindMinimum} function will not work in the form of equation (5) or (10) since Mathematica cannot compute the symbolic derivative of the functions \text{tailG}_1 and \text{tailG}_2, but \text{FindMinimum} has a numerical form where two initial values are supplied for each parameter. Some care is needed in the choice of these values. The \text{Plot} function is invaluable here. We define the Mathematica function \text{l}[a, m_1, m_2] by

\[
l[a_, m_1, m_2] = \logl
\]

We can now find approximations to the maximum likelihood estimates by a succession of plots of $l$ by varying one variable while the other two variables are held fixed. For example, we might start with a fixed at 1 (an exponential distribution) and $m_2$ fixed at 8.7 (the mean of the second group). We issue

\[
\text{Plot}[l[1, m_1, 8.7], \{m_1, 10, 50\}]
\]

The plot suggests that $m_1$ is about 32. We continue with

\[
\text{Plot}[l[a, 32, 8.7], \{a, 0.5, 2\}]
\]

This leads quickly to good approximations to the maximum likelihood estimates. \text{FindMinimum} can now be used to obtain more accurate values. In the present case, we used
max = FindMinimum[-logl, {a, 1.4, 1.5}, {m1, 32, 33}, {m2, 8, 9}]
mleG = max[[2]]

The hazard functions for the two groups and the ratio of the hazards are given by

hazardG1 = pdfG1[t]/tailG1[t]. mleG
hazardG2 = pdfG2[t]/tailG2[t]. mleG
ratioG = hazardG2/hazardG1

A plot of the ratios of the hazards for the exponential, Weibull and gamma models can now be obtained with

Plot[{ratioE, ratioW, ratioG}, {t, 1, 30}]

and is given in Fig. 3 (we have assumed that ratioE and ratioW have been obtained, and we have omitted the graphics options that control the detailed appearance of the graph). Fig. 3 emphasizes the proportional hazards nature of the exponential and Weibull models, and the non-proportional hazards nature of the gamma model is apparent.

The information matrix must also be found numerically. One simple way of doing this is to use finite differences. First, we obtain the maximum likelihood estimates amax, m1max and m2max as follows:

amax = a/. mleG; m1max = m1/. mleG; m2max = m2/. mleG

where the semicolon allows more than one command on a single input line. The finite

Fig. 3. Ratios of hazards: ---, exponential model; ---, Weibull model; ------, gamma model
difference approximation to $\frac{\partial^2(\log l)}{\partial a^2}$ at the point $(a_{\text{max}}, m_{1\text{max}}, m_{2\text{max}})$ with increment $h$ is

$$da^2 = (l[a + h, m_{1\text{max}}, m_{2\text{max}}] - 2l[a, m_{1\text{max}}, m_{2\text{max}}] + l[a - h, m_{1\text{max}}, m_{2\text{max}}])/h^2$$

and a few trials with $h$ give a satisfactory estimate of $\frac{\partial^2(\log l)}{\partial a^2}$. The remaining entries of the information matrix can be found in a similar fashion; Mathematica functions could be defined to streamline this process.

The second example is taken from Dellaportas and Smith (1993), who showed how the Gibbs sampler can be used to make Bayesian inferences for GLMs. They gave data from Knuiman and Speed (1988) on the connection between duration of diabetes and retinopathy, an eye disease. Table 8 is a modification of Knuiman and Speed's original data and is used by Dellaportas and Smith to illustrate the fitting of the quadratic logistic model

$$\log \left( \frac{\pi_{1j}}{\pi_{2j}} \right) = \beta_1 + \beta_2 Z_j + \beta_3 Z_j^2 = \eta_j$$

where $(\pi_{1j}, \pi_{2j})$ are the probabilities of being with or without retinopathy with duration in the $j$th category and $Z = (Z_j) = (1, 4, 7, 10, 13, 16, 19, 24)$ is the vector of mid-durations. In this example, Dellaportas and Smith assumed that no prior data are available and take as the prior $p(\beta_1, \beta_2, \beta_3) = $ constant. There is no difficulty in obtaining the posterior summary statistics of Knuiman and Speed since these are the maximum likelihood estimates. We set yes = \{1, 1, ..., 1\}, no = \{6, 4, ...\} and $z = \{1, 4, ..., 10\}$, where the scaling of $Z$ gives a more convenient scale for the estimation of the regression parameters $\beta_1, \beta_2$ and $\beta_3$. The linear predictor for the quadratic logistic model is given by

$$\eta_j = \beta_1 + \beta_2 z_j + \beta_3 z_j^2$$

and the log-likelihood follows (Dellaportas and Smith (1993), equation (2)) as

$$logl = \text{Apply}[\text{Plus}, \text{yes} \eta_j - (\text{yes} + \text{no}) \text{Log}[1 + \text{Exp}(\eta)]]$$

<table>
<thead>
<tr>
<th>Duration of diabetes (years)</th>
<th>Retinopathy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>0-2</td>
<td>1</td>
</tr>
<tr>
<td>3-5</td>
<td>1</td>
</tr>
<tr>
<td>6-8</td>
<td>1</td>
</tr>
<tr>
<td>9-11</td>
<td>1</td>
</tr>
<tr>
<td>12-14</td>
<td>1</td>
</tr>
<tr>
<td>15-17</td>
<td>1</td>
</tr>
<tr>
<td>18-20</td>
<td>1</td>
</tr>
<tr>
<td>21+</td>
<td>1</td>
</tr>
</tbody>
</table>
The maximum likelihood estimate $\beta$ is found by equation (10) and its covariance matrix by equations (11). Mathematica provides a variety of integration routines and these can be used to find the posterior mean and covariance matrix of the Bayesian estimate of $\beta$. We obtained

$$
\beta^* = \begin{pmatrix}
-2.469 \\
2.457 \\
-0.493
\end{pmatrix}, \quad D^* = \begin{pmatrix}
1.517 & -2.594 & 0.892 \\
-2.594 & 6.117 & -2.403 \\
0.892 & -2.403 & 1.039
\end{pmatrix},
$$

in close agreement with the simulated values of Dellaportas and Smith. One of the advantages of the Gibbs sampler approach is that, once the sample of $\beta$-values has been assembled, further statistics (like means and covariances) are easily found. Of course, Mathematica must compute any integrals from scratch, but its powerful language and simple graphics have other advantages. Dellaportas and Smith (1993) compared the maximum likelihood and Bayesian solutions by graphing the marginal distributions of $\beta_1$, $\beta_2$ and $\beta_3$ which they readily obtained from their sample of $\beta$-values. Instead we compare the two solutions by plotting the linear predictor $\eta$ together with approximate (pointwise) 95% confidence or Bayesian intervals. In Mathematica we proceed in a direct fashion: let $\text{betabayes} = \{-2.469, 2.457, -0.493\}$ and $\text{covbayes} = \{\{1.517, -2.594, 0.892\}, \ldots\}$ be $\beta^*$ and $D^*$ respectively. For the Bayesian plot the functions to be computed are given by

$$
\text{etabayes} = \text{betabayes}[\{1\}] + \text{betabayes}[\{2\}] \, xx + \text{betabayes}[\{3\}] \, xx^2 \\
\text{sterror} = \text{Sqrt}[\{1, xx, xx^2\}.\text{covbayes}.\{\{1\}, \{xx\}, \{xx^2\}\}] \\
\text{upper} = \text{etabayes} + 2 \, \text{sterror} \\
\text{lower} = \text{etabayes} - 2 \, \text{sterror} \\
\text{Plot}[\{\text{etabayes}, \text{upper}, \text{lower}\}, \{xx, 0, 2.5\}]
$$

In these expressions the generic variable $xx$ is used, not the data list $z$; the output from the line $\text{etabayes} = \ldots$ is the algebraic expression

$$\text{Out}[n] = -2.469 + 2.457 \, xx - 0.493 \, xx^2$$

Fig. 4 also includes the plot from the maximum likelihood estimation. Although $\beta^* = 1.1\beta$ the predicted values of $\eta$ under the two methods are quite close over the range of $Z$. The Bayesian interval is noticeably wider than the maximum likelihood interval for small and large values of $Z$.

6. Concluding Remarks

This paper has used Mathematica to solve problems in maximum likelihood estimation, which range from the elementary to the computationally demanding. A feature of the solutions described has been the common approach to these diverse problems. We have targeted the paper at anyone with little or no experience of Mathematica, and for that reason we have used a very restricted set of Mathematica commands. Three omissions are worth mentioning. First, Mathematica has extensive algebraic simplification commands. The output provided by equations (3), (9) and similar expressions can be greatly simplified; we have omitted such simplifications since they refer mainly to the efficiency of computations and do not materially affect the method. Second, we have used the FindMinimum command. A different
approach involves differentiating the log-likelihood and solving the resulting equation(s). For example, instead of equation (5) we could proceed with

\[
\begin{align*}
d\log l &= D[\log l, t] \\
mle &= \text{Solve}[d\log l == 0, t]
\end{align*}
\]

The Solve command will find exact solutions, if they exist; a numerical version of Solve (NSolve) exists for equations without exact solutions. In this paper we have ignored this avenue, although this alternative approach could provide a more efficient solution for many problems. Third, we have ignored any particular structure that a problem may have. This leads to a conceptually simple approach but which in practice will be severely limited. Even a program as powerful and sophisticated as Mathematica will need help; a serious user could supply such help as he or she saw fit.

We see Mathematica as having two roles in statistics. In teaching, the approach outlined in this paper brings many conceptually simple but computationally heavy problems within easy reach of the student. As a research instrument, Mathematica provides a symbolic computational tool of immense power, which complements the data handling and model fitting abilities of any statistical package.

References


