A Bootstrap Evaluation of the EM Algorithm for Censored Survival Data

George W. Pasdirtz, Hazleton Laboratories

Introduction
At Hazleton Laboratories, we are looking for new ways to analyze survival data from long-term rodent carcinogenicity studies. The available parametric and nonparametric techniques in SAS® rely on assumptions that are sometimes difficult to justify scientifically. In this paper, the EM algorithm is investigated as an alternative to SAS PROC LIFEREG.

In a typical rodent carcinogenicity study, we record time to death and, for selected palpable and rapidly lethal tumors, time to tumor (Thakur, Wetzel, and Stevens, 1990) for animals given various doses of a compound. The current methods for analyzing survival data using PROC LIFEREG are specially designed to handle a common feature of lifetime data, censored observations. Censoring enters either due to withdrawal of animals or termination of the study. Some animals die on test for reasons unrelated to treatment or are serially sacrificed to obtain tumor onset times. At the end of the study, the remaining animals are sacrificed. In both cases, the complete lifetimes (and thus tumor incidences) for the censored animals are unobserved.

Accelerated Failure Time Model
The accelerated failure time model assumes that the independent variables act as multipliers on the event time (death or tumor onset).

\[ T = \exp(\mathbf{X}\beta)T_0 \]
\[ \ln(T) = \mathbf{X}\beta + \ln(T_0) \]
\[ \ln(T) = \mathbf{X}\beta + \mathbf{e} \]

where \( T \) is an event time, \( T_0 \) is an event time from a baseline distribution, \( \mathbf{X} \) is a matrix of independent variables, and \( \mathbf{e} \) is an error term, equivalent to \( \ln(T_0) \).

ML Algorithm
The unknown parameters \( \beta \) and the variance \( \sigma \) in the accelerated failure time model are estimated in PROC LIFEREG using an iterative maximum likelihood (ML) algorithm. The likelihood for any observation

\[ L(\beta, \sigma) = \delta \ln(f(e)) + (1 - \delta) \ln(S(e)) \]

is written as a function of a probability density function \( f(e) \) and a survival function \( S(e) \) depending on whether the observation is censored \( \delta = 1 \) or not.

EM Algorithm
The EM (Expectation-Maximization) algorithm (Efron, 1967, Dempster et al., 1977, Turnbull and Mitchell, 1978, and Schmee and Hahn, 1979) drops the explicit likelihood function and simply treats the censored observations as if they were missing data. The lifetime of an animal that died accidentally on test or that was still alive when the experiment was terminated is considered to be a missing observation. The algorithm alternates between estimating the missing data (the E-step) and maximizing or minimizing (the M-step) some criterion, say the root mean square error (RMSE).

The estimation proceeds by first obtaining

\[ \beta_1 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ln(T) \]

ignoring censoring. Then:

**E-step**: Replace all censored event times with the expectation

\[ \ln(T_{s+1}) = \mathbf{X}\beta_1 \]

**M-step**: Reestimate with censored event times now taken at their expectation

\[ \beta_{s+1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ln(T_s) \]

The EM steps are then repeated until the change in RMSE is small. A SAS macro that computes the EM algorithm is presented in Appendix 1.

Hazleton Data
PROC LIFEREG and the EM algorithm were used to estimate group differences in time-to-death using data from a 106-week rodent carcinogenicity study conducted at Hazleton Laboratories. A compound was administered at four different dose levels and compared with a control group using planned comparisons coded in the data step (Freund, 1989):  

1277
Notice from a comparison of the results in Appendices 2 and 3 that, while the estimated coefficients are similar in sign to those produced by PROC LIFEREG, the EM standard errors are smaller, resulting in larger t-statistics.

Also note that

- Survival is lower for the low-dose groups, a result that often indicates the reduction in diet-related tumors as a result of side effects at high doses.
- Over 40% of the sample is right-censored.
- The effect of the EM algorithm was to increase mean survival and decrease the standard deviation, but only in the high-dose groups. For example, Group 3 went from a mean of 96.7 weeks and a standard deviation of 6.413 to 105.49 weeks and 0.51, respectively.

Yet, both techniques identify all groups as being significantly different. Which of the estimates is the right answer?

Bootstrap Estimates of Error and Bias

To decide which of the two estimators is more accurate we will need a way to establish true or population values against which to compare the sample estimates. A nonparametric, computer-intensive technique called the bootstrap (Efron, 1979, 1982) can be used to produce synthetic population estimates that then allow comparative measures of bias, error, and relative efficiency.

The bias of an estimator refers to the deviation of the expected value from the true value.

$$
\text{Bias}(S) = E[S] - \mu
$$

The mean squared error (MSE) is the expected value of the square of the deviation of the estimator from the true value.

$$
\text{MSE}(S) = \sigma^2 + [\text{Bias}(S)]^2
$$

Relative efficiency is the ratio of MSEs between two similar methods. Estimators should have low bias and minimum MSE.

Measures of bias and MSE, however, are not scale-free and thus cannot be used to compare dissimilar techniques. We can construct a scale-free number, the standard t-statistic

$$
t = \frac{\beta}{\sigma},
$$

and from it construct measures of Bias(t) and comparative bias

$$
\frac{\text{Bias}(t)_{EM}}{\text{Bias}(t)_{ML}}.
$$

To compute bias and MSE, the bootstrap draws a random sample, with replacement, from the original observations (X,Y). Each model is then estimated on the new, synthetic sample, and the regression estimates are stored in a file. The process is repeated many times to develop the sampling distribution of the estimators from which the “population” mean and variance can be computed. The bootstrap estimators are the nonparametric ML estimates of the population values. For regression problems (Freedman, 1981), the bootstrap replicates the standard analytical solutions.

To compute the bootstrap, we use a modification of a SAS macro written by Carson (1985). Some initial exploratory work indicated that the normal distribution (which is not the default value) produced the best estimates when using PROC LIFEREG. The synthetic population estimates are presented in Appendix 4 and Appendix 5. Notice that the sample standard errors for both the ML and the EM algorithm are smaller than the population values.

Measures of relative bias (Appendix 8) indicate a mixed conclusion when comparing the EM with the ML algorithm. The EM algorithm works best only for the intercept and the comparison of groups 1 and 2. The ML algorithm has lower relative bias for the other comparisons. The EM-RT algorithm produces estimates which are an order of magnitude greater in bias than either of the others.

Testing Assumptions

The standard methods for testing assumptions in survival models are either graphical or based on the empirical distribution function (EDF) (D’Agostino and Stephens, 1986).

The results of inspecting probability plots (omitted) are ambiguous. If a few early deaths and the large number of censored animals in the low-dose groups can be ignored, the normal distribution appears appropriate.

A Shapiro-Wilk test (Shapiro and Wilk, 1965) for normality (PROC UNIVARIATE) on the residuals from the EM algorithm was significant (p ≤ 0.0001). However, the same test on the bootstrap parameter distributions for both the ML and EM algorithms was not significant.
EM Followed by RT

In a production environment where results must be submitted to a regulatory agency, it may not be acceptable to bootstrap the EM algorithm. The failure of the Shapiro-Wilk test for the EM residuals would suggest converting to a nonparametric model based on rank transformations (RT) of the EM-corrected failure times (Conover and Iman, 1981; Conover, 1982).

The results of an EM-RT estimation are presented in Appendix 6 and produce the same conclusions as the EM algorithm. The bias of the EM-RT model, however, is greater than either the ML or the EM estimators (Appendix 8).

Results from the Shapiro-Wilk test on the bootstrap distribution of EM-RT parameters (omitted) indicate that, except for the RMSE, the normal distribution would be inappropriate. Bootstrap of a pure RT model (omitted) had larger bias than the EM-RT model, indicating some loss of efficiency when using the RT estimator.

Summary

Three techniques (ML, EM, and EM-RT) were compared for use with censored survival data.

- No one technique was clearly superior.

- As an automatic technique, bootstrapping the EM estimator and constructing appropriate null distributions for t-statistics would perform very well, but is not currently acceptable to regulatory agencies.

- Misapplication of the ML estimator, however, can have disastrous consequences. When the Weibull distribution (the default value in PROC LIFEREG) was used in the present sample, no significant differences between groups were found (although measures of bias were only an order of magnitude greater). A large-scale simulation might address whether the EM-RT algorithm performs on average better under conditions of misspecification.

Although limited simulations often raise more questions than they answer, it seems clear that survival models cannot be automatically estimated in practice without careful attention to distributional assumptions.

References


SAS® is a registered trademark of the SAS Institute, Inc., Cary, NC, USA.

Author

George W. Pasdirtz, PhD
Hazleton Laboratories
3301 Kinsman Boulevard
Madison, WI 53704
Appendix 1. SAS Code for the EM algorithm.

```sas
%macro mvem (tol=, datin=, ivar=1 dvar=, pvar=, censor=, itmax=);
* Multivariate E-M algorithm

tol      tolerance for convergence
datin    input dataset
ivar     independent variables
dvar     dependent variables
pvar     predicted variable list (similar to dvar)
censor   censored dvar indicator variable
itmax    max iterations

NOTE: The datin data set is modified to contain predicted values for censored observations;

%local it quit std;
%let it = 0; %let quit = 0;
%do %until (&quit = 1 or &it > &itmax); %let it = %eval(&it+1);
   proc reg data = &datin outest = dcoeff noprint;
model &dvar = &ivar;
output out = p p = &pvar;
%if &it = 1 %then %do; * convergence check;
data _null_; set dcoeff;
call symput ('nnse',_nnse_);
stop;
%end;
%else %do;
data _null_; set dcoeff;
rrnse=symget ('rrnse ');
delta = rrnse - _rrnse_; put 'NOTE: ' delta= rrnse= _rrnse;
if delta > &tol then quit = 0;
else quit = 1;
call symput ('rrnse',_rrnse_);
call symput ('quit',quit);
stop;
%end;
data &datin; * estimate censored values;
set p; array d &dvar; array p &pvar; array c &censor;
do over d;
   if c = 1 then d= p; end;
%end;
%if &it = &itmax %then %put WARNING: FAILED TO CONVERGE.;
%else %put NOTE: CONVERGENCE AT &it ITERATIONS;
%mend;
```

Appendix 2. ML Estimator.

LIFEREG PROCEDURE

Data Set = WORK.A
Dependent Variable = DEATH
Censoring Variable = CENSOR
Censoring Value(s) = 1
Noncensored Values = 74  Right Censored Values = 175
Left Censored Values = 0  Interval Censored Values = 0

Log Likelihood for NORMAL -423.4366725

1280
<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Std Err</th>
<th>ChiSquare</th>
<th>Pr&gt;Chi</th>
<th>Label/Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCPT</td>
<td>1</td>
<td>127.076888</td>
<td>5.301131</td>
<td>574.6403</td>
<td>0.0001</td>
<td>Intercept</td>
</tr>
<tr>
<td>D1</td>
<td>1</td>
<td>-34.514835</td>
<td>15.98841</td>
<td>4.660161</td>
<td>0.0309</td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td>1</td>
<td>-43.073153</td>
<td>20.794</td>
<td>4.290789</td>
<td>0.0383</td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td>1</td>
<td>57.805896</td>
<td>26.18906</td>
<td>4.87197</td>
<td>0.0273</td>
<td></td>
</tr>
<tr>
<td>D4</td>
<td>1</td>
<td>97.251856</td>
<td>46.91569</td>
<td>4.296944</td>
<td>0.0382</td>
<td></td>
</tr>
<tr>
<td>SCALE</td>
<td>1</td>
<td>23.936832</td>
<td>2.283036</td>
<td></td>
<td></td>
<td>Normal scale parameter</td>
</tr>
</tbody>
</table>

**Appendix 3. EM Estimator.**

Model: MODEL1
Dependent Variable: LDEATH

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>0.25513</td>
<td>0.06378</td>
<td>6.218</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>244</td>
<td>2.50306</td>
<td>0.01026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>248</td>
<td>2.75819</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.10128
Dep Mean 4.73705
C.V. 2.21480

### Parameter Estimates

| Variable  | DF | Parameter | Standard Error | T for H0: Parameter=0 | Prob > |T| |
|-----------|----|-----------|----------------|------------------------|--------|---|
| INTERCPT  | 1  | 4.625960  | 0.01446915     | 319.712                | 0.0001 |
| D1        | 1  | -0.230845 | 0.04999714     | -4.617                 | 0.0001 |
| D2        | 1  | -0.263490 | 0.06457845     | -4.080                 | 0.0001 |
| D3        | 1  | 0.351816  | 0.08167062     | 4.308                  | 0.0001 |
| D4        | 1  | 0.595015  | 0.14583007     | 4.080                  | 0.0001 |

**Appendix 4. LIFEREG Bootstrap.**

<table>
<thead>
<tr>
<th>N Obs</th>
<th>Variable</th>
<th>Label</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td><em>LNLIKE</em></td>
<td>Natural Log of Likelihood</td>
<td>-422.7927202</td>
<td>32.9428372</td>
</tr>
<tr>
<td></td>
<td>INTERCPT</td>
<td>Intercept</td>
<td>126.7368065</td>
<td>5.1324464</td>
</tr>
<tr>
<td></td>
<td>D1</td>
<td></td>
<td>-34.3314552</td>
<td>16.1822543</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td></td>
<td>-43.4429256</td>
<td>21.0557882</td>
</tr>
<tr>
<td></td>
<td>D3</td>
<td></td>
<td>58.1393695</td>
<td>26.5790082</td>
</tr>
<tr>
<td></td>
<td>D4</td>
<td></td>
<td>97.5015091</td>
<td>47.3183737</td>
</tr>
<tr>
<td></td>
<td><em>SCALE</em></td>
<td>Scale Parameter for Distribution</td>
<td>23.2253174</td>
<td>2.3833473</td>
</tr>
</tbody>
</table>

**Appendix 5. EM Bootstrap.**

<table>
<thead>
<tr>
<th>N Obs</th>
<th>Variable</th>
<th>Label</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td><em>RMSE</em></td>
<td>Root mean squared error</td>
<td>0.0979440</td>
<td>0.0156099</td>
</tr>
<tr>
<td></td>
<td>INTERCPT</td>
<td>Intercept</td>
<td>4.6263356</td>
<td>0.0112676</td>
</tr>
<tr>
<td></td>
<td>D1</td>
<td></td>
<td>-0.2315634</td>
<td>0.0737352</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td></td>
<td>-0.2654279</td>
<td>0.0818534</td>
</tr>
<tr>
<td></td>
<td>D3</td>
<td></td>
<td>0.3546692</td>
<td>0.1053485</td>
</tr>
<tr>
<td></td>
<td>D4</td>
<td></td>
<td>0.5982501</td>
<td>0.1868194</td>
</tr>
<tr>
<td></td>
<td>LDEATH</td>
<td></td>
<td>-1.0000000</td>
<td>0</td>
</tr>
</tbody>
</table>

1281
Appendix 6. EM-RT Estimator.

Dependent Variable: LDEATH VALUE OF LDEATH REPLACED BY RANK

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>549710.84755</td>
<td>137427.71189</td>
<td>46.667</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>244</td>
<td>718552.65245</td>
<td>2944.88792</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>248</td>
<td>1268263.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 54.26682  R-square 0.4334  Adj R-sq 0.4241

C.V. 43.41346

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > |T| |
|----------|----|--------------------|----------------|-----------------------|--------|---|
| INTERCEP | 1  | 200.448980         | 7.75240308     | 25.856                | 0.0001 |
| D1       | 1  | -328.786939        | 26.7879011     | -12.274               | 0.0001 |
| D2       | 1  | -375.735918        | 34.60039146    | -10.859               | 0.0001 |
| D3       | 1  | 509.954898         | 43.75817804    | 11.654                | 0.0001 |
| D4       | 1  | 838.860816         | 78.13407040    | 10.736                | 0.0001 |

Appendix 7. EM-RT Bootstrap.

<table>
<thead>
<tr>
<th>N Obs</th>
<th>Variable</th>
<th>Label</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>RMSE</td>
<td>Root mean squared error</td>
<td>51.2443963</td>
<td>3.8063328</td>
</tr>
<tr>
<td>INTERCEP</td>
<td>Intercept</td>
<td>198.5118489</td>
<td>14.9903475</td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>-313.2916511</td>
<td>50.6088981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td>-364.5831534</td>
<td>71.5662995</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td>478.5342530</td>
<td>89.6551916</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D4</td>
<td>833.1294047</td>
<td>160.0541413</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Appendix 8. Bias(t).

<table>
<thead>
<tr>
<th>TERM</th>
<th>ML</th>
<th>EM</th>
<th>EM-RT</th>
<th>EM/ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEP</td>
<td>0.66259E-01</td>
<td>-0.33305E-01</td>
<td>.12923</td>
<td>-.50265</td>
</tr>
<tr>
<td>D1</td>
<td>-0.11332E-01</td>
<td>0.97428E-02</td>
<td>-.30618</td>
<td>-.85973</td>
</tr>
<tr>
<td>D2</td>
<td>0.17562E-01</td>
<td>0.23675E-01</td>
<td>-.15584</td>
<td>1.3481</td>
</tr>
<tr>
<td>D3</td>
<td>-0.12546E-01</td>
<td>-0.27083E-01</td>
<td>.35046</td>
<td>2.1587</td>
</tr>
<tr>
<td>D4</td>
<td>-0.52759E-02</td>
<td>-0.17316E-01</td>
<td>0.35809E-01</td>
<td>3.2821</td>
</tr>
</tbody>
</table>

1282