Computing extended maximum likelihood estimates for Cox proportional-hazards models

Douglas B. Clarkson, MathSoft Inc.
Robert I. Jennrich, University of California, Los Angeles

Abstract

Infinite parameter estimates can occur in Cox proportional–hazards models when a linear combination of the covariates monotonically increases or decreases with the failure times. Considered here are methods for computing *extended maximum likelihood* estimates for Cox models. An extended estimate is a pair of vectors: a direction vector for the infinite component, and a vector that optimizes the "stratified" log-likelihood of Bryson and Johnson (1981). A method of identifying problems that have extended estimates and an algorithm for finding them is given along with an example illustrating their use.

Key Words: Computer algorithms; Cox regression; Linear programming; Solutions at infinity.

1 Introduction

When the maximizer of a likelihood is at infinity, many standard computational algorithms fail because they are not designed to deal with such solutions. Haberman (1974, appendix B) defines and gives an example of such estimates for frequency data. He calls the estimates obtained *extended maximum likelihood* estimates. Because these estimates contain infinite values, they do not exist in the usual sense, and many authors (Silvapulle and Burridge, 1986, Albert and Anderson, 1984, Hamada and Tse, 1988) have felt that detecting the presence of infinite estimates is sufficient. Bryson and Johnson (1981) note that infinite estimates can be common in the Cox (1972) proportional-hazards model, at least in a simulation study of a model with sample size 20. They also note that infinite estimates occur when the failure times are monotone with a linear function of the covariates, and they give an algorithm in which parameter estimates are computed based upon a "stratified" likelihood, where the stratification is determined by inspection of the covariates. Baker, Clarke, and Lane (1985) discuss an algorithm for computing extended maximum likelihood estimates in sparse contingency tables, and Clarkson and Jennrich (1991) give algorithms for computing extended maximum likelihood estimates in linear-parameter models. While the algorithms in these two papers are not applicable to the Cox proportional-hazards likelihood, they are similar to the algorithms developed here.

An intuitive explanation of why standard algorithms fail on problems with infinite maximizers is that for such problems in a neighborhood a solution some large changes in the parameter vector give very small changes in the log-likelihood assuming the latter is bounded. This suggests that the Hessian of the log-likelihood and the Fisher information matrix are nearly singular. Since standard algorithms typically invert one or the other of these matrices during their iteration process, the process breaks down when these matrices become to close to singular.

Our interest in extended maximum likelihood is that in some cases, including those mentioned above and the Cox proportional-hazards model, the log-likelihood has a finite supremum, but may not have a maximum. Extended maximum likelihood programs are designed to produce suprema and extended estimates even when the maximum does not exist. While the extended estimates may be infinite, useful functions of them can be well defined and finite, for example means in the context of log-linear models and relative hazard rates in the context of the Cox proportional-hazards model. Extended estimates may also be used to form likelihood ratio statistics and likelihood ratio confidence intervals. All of this can be much more valuable than simply failing to find a maximum.

In the next section theorems on extended maximum likelihood estimates in Cox proportional-hazards models are given. These are used as the basis of a computing algorithm given in Section 4. Section 3 gives a linear programming algorithm for solving a system of linear inequalities required by the algorithm in Section 4. Tied observation times are discussed in Section 5. Section 6 discusses an example and a simulation study of Johnson *et al.* (1982). Concluding remarks are given in Section 7.

2 Theory of Extended Maximum Likelihood Estimation in the Cox Proportional Hazards Model

Let $t_1 < t_2 < \cdots < t_n$ denote the observation times in a proportional hazards model and for each t_i let x_i denote a p-vector of covariate values. Some of the t_i are failure times and the remainder are right censoring times. Let $\mathcal{F}$ be the set of index values i for which t_i is a failure time. The log-likelihood for the Cox proportional–hazards model is (see, e.g., Kalbfleish and Prentice, 1980; Elandt-Johnson and Johnson, 1980; Lawless, 1982; Lee, 1980; Cox and

Oakes, 1984):

$$\ell(\beta) = \sum_{i \in \mathcal{F}} \ln \left(\frac{\exp(x_i \beta)}{\sum_{j \geq i} \exp(x_j \beta)} \right) \tag{1}$$

where β is a vector of parameters to be estimated. Because right-censored observations which occur prior to the first failure do not enter the log-likelihood, we assume, without loss of generality, that t_1 is a failure time. For now we have assumed there are no tied observation times t_i. These are discussed later.

In applications of the Cox proportional-hazards model it occasionally happens that there is no maximum likelihood estimate, that is, there is no $\hat{\beta}$ in $\Re^p$ that maximizes ℓ. We call $(\hat{\beta}, d)$ an extended maximum likelihood estimate if

$$\lim_{\rho \to \infty} \ell(\hat{\beta} + \rho d) = \sup \ell \tag{2}$$

One may view $(\hat{\beta}, d)$ as a directed line. Assuming $(\hat{\beta}, d)$ is an extended estimate, as one moves along the line in the direction d the value of ℓ approaches its supremum. One may also view extended estimation as ordinary maximum likelihood estimation after extending the domain of ℓ to the set of directed lines. An ordinary maximum likelihood estimate is an extended estimate with $b = 0$. The Cox proportional hazards model can have extended maximum likelihood estimates that are not ordinary estimates.

Assuming it exists let

$$\ell_d(\beta) = \lim_{\rho \to \infty} \ell(\beta + \rho d) \tag{3}$$

Our approach to finding extended estimates will be to find an appropriate d and maximize ℓ_d to produce $\hat{\beta}$.

For $j = 2, \ldots, n$, let $\gamma(j)$ be the index of the largest failure time before t_j. Let

$$a_j = (x_j - x_{\gamma(j)}) \tag{4}$$

and let A be the matrix of such differences. Let d be any solution to the inequality problem

$$Ad \leq 0, \tag{5}$$

such that Ad has the largest possible number of negative components. The inequality in (5) means each component of Ad is less than or equal to zero.

In the next section we show how to use linear programming to find a d that satisfies (5). The vector d need not be unique even up to a scalar multiple. However, because the sum of two solutions to (5) is also a solution, the components of Ad that are negative is unique. Our theory will show that ℓ_d exists, has a finite maximizer $\hat{\beta}$, and that $(\hat{\beta}, d)$ is an extended maximum likelihood estimator. Moreover, ℓ has a finite maximizer if and only if $Ad = 0$ and when this is true $\hat{\beta}$ maximizes ℓ.

Lemma 1: If $i \in \mathcal{F}$ and $j \geq i$,

$$(x_j - x_i)d \leq 0$$

Proof: The result clearly holds if $j = i$. If $j > i$, $i = \gamma^p(j)$ for some integer p where γ^p is the p-fold composition of γ. Note that

$$x_j - x_i = (x_j - x_{\gamma(j)}) + \cdots + (x_{\gamma^{p-1}(j)} - x_{\gamma^p(j)})$$

Multiplying by d gives

$$(x_j - x_i)d \leq 0$$

because all of the terms on the right hand side are non-positive. •

Corollary 1: The sequence $x_i d$, for $i \in \mathcal{F}$ is non-increasing.
Proof: If $i, j \in \mathcal{F}$ and $i < j$, it follows from Lemma 1 that $(x_j - x_i)d \leq 0$. Hence $x_i d \geq x_j d$. •

Definition 1: $R_i = \{j : j \geq i \text{ and } (x_j - x_i)d = 0\}$

Theorem 1: $\ell_d(\beta)$ exists for all $\beta \in \Re^p$ and is given by

$$\ell_d(\beta) = \sum_{i \in \mathcal{F}} \ln\left(\frac{\exp(x_i\beta)}{\sum_{j \in R_i} \exp(x_j\beta)}\right). \tag{6}$$

Proof: Consider $\ell(\beta+\rho d)$. Its expansion (1) has terms that are the logarithms of ratios of the form

$$\frac{\exp(x_i(\beta+\rho d))}{\sum_{j\geq i}\exp(x_j(\beta+\rho d)} = \frac{\exp(x_i\beta)}{\sum_{j\geq i}\exp(x_j\beta+\rho(x_j-x_i)d)} \to \frac{\exp(x_i\beta)}{\sum_{j\in R_i}\exp(x_j\beta)}$$

as $\rho \to \infty$. Thus $\ell_d(\beta)$ exists for all β and is given by (6). •

Lemma 2: Let $i \in \mathcal{F}$ and $j \in R_i$. If $k \in \mathcal{F}$ and $i \leq k \leq j$, then $k \in R_i$.
Proof: Note that

$$x_j - x_i = x_j - x_k + x_k - x_i$$

Multiplying by d gives

$$0 = (x_j - x_k)d + (x_k - x_i)d$$

By Lemma 1 both terms on the right are non-positive and hence $(x_k - x_i)d = 0$. Thus $k \in R_i$. •

Definition 2: $\mathcal{Z} = \{j : a_j d = 0\}$

Lemma 3: If $i \in \mathcal{F}$, $j \in R_i$, and $j > i$, then $j \in \mathcal{Z}$.
Proof: Note that

$$x_j - x_i = x_j - x_{\gamma(j)} + x_{\gamma(j)} - x_i$$

Multiplying on the right by d gives

$$0 = (x_j - x_{\gamma(j)})d + (x_{\gamma(j)} - x_i)d$$

Since $i \leq \gamma(j)$ and $i, \gamma(j) \in \mathcal{F}$, it follows from Lemma 1 that the terms on the right are non-positive and hence zero. Thus $a_j d = 0$ and $j \in \mathcal{Z}$. •

Theorem 2: The function ℓ_d has a finite maximizer.
Proof: We may write ℓ_d in the form

$$\ell_d(\beta) = -\sum_{i \in \mathcal{F}} \log(\sum_{j \in R_i} \exp((x_j - x_i)d)) \tag{7}$$

Let $\mathcal{M}$ be the space spanned by the differences $x_j - x_i$ for which $i \in \mathcal{F}$ and $j \in R_i$. These are the differences that appear in (7). If ℓ_d has no finite maximizer, then there are β_n in $\mathcal{M}$ such that

$$\ell_d(\beta_n) \to \sup \ell_d$$

and such that $\beta_n = \rho_n v_n$ with $\rho_n \to \infty$ and $v_n \to v \neq 0$. Let $i \in \mathcal{F}$ and $j \in R_i$. Then

$$(x_j - x_i)v \leq 0 \tag{8}$$

because otherwise $\ell_d(\beta_n) \to -\infty$.

If $j \in \mathcal{Z}$, then $j \in R_{\gamma(j)}$. Using (8), $(x_j - x_{\gamma(j)})v \leq 0$. Thus

$$a_j v \leq 0$$

for all $j \in \mathcal{Z}$.

Since $v \neq 0$ and $v \in \mathcal{M}$, there is an $i \in \mathcal{F}$ and $j \in R_i$ such that $(x_j - x_i)v \neq 0$. From (8),

$$(x_j - x_i)v < 0. \tag{9}$$

Let j be the smallest index in R_i for which (9) holds. Note that

$$x_j - x_i = x_j - x_{\gamma(j)} + x_{\gamma(j)} - x_i$$

Multiplying by v and using (9) gives

$$0 > a_j v + (x_{\gamma(j)} - x_i)v$$

By Lemma 2, $\gamma(j) \in R_i$ so the second term is non-positive. If it is negative, j is not the smallest index in R_i for which (9) holds. Thus the second term is zero and $a_j v < 0$. From (9), $j \neq i$. By Lemma 3, $j \in \mathcal{Z}$ and hence

$$a_j v < 0 \tag{10}$$

for some $j \in \mathcal{Z}$.

Since $a_j d = 0$ and $a_j v \leq 0$ for all $j \in \mathcal{Z}$ and $a_j d < 0$ for all $j \notin \mathcal{Z}$, for ρ sufficiently large,

$$A(v + \rho d) \leq 0,$$

but from (10), $A(v+\rho d)$ has at least one more negative component than Ad. This contradicts the definition of d and completes the proof. •

Theorem 3: If $\hat{\beta}$ maximizes ℓ_d, then $(\hat{\beta}, d)$ is an extended maximum likelihood estimator.
Proof: From (1) and (6), $\ell \leq \ell_d$. Thus

$$\sup \ell \leq \sup \ell_d = \ell_d(\hat{\beta}) = \lim_{\rho \to \infty} \ell(\hat{\beta} + \rho d) \leq \sup \ell$$

Hence $\lim_{\rho\to\infty} \ell(\hat{\beta} + \rho d) = \sup \ell$. •

Corollary 2: If $\hat{\beta}$ maximizes ℓ_d and $Ad = 0$, then $\hat{\beta}$ maximizes ℓ.
Proof: If $Ad = 0$, $R_i = \{j : j \geq i\}$ and hence $\ell_d = \ell$. Because $\hat{\beta}$ maximizes ℓ_d it maximizes ℓ.

Theorem 4: The function ℓ has a finite maximizer if and only if $Ad = 0$.
Proof: If $Ad = 0$, $R_i = \{j : j \geq i\}$ and hence $\ell = \ell_d$ which has a finite maximizer by Theorem 2.

If $Ad \neq 0$, $(x_j - x_i)d < 0$ for some $i \in \mathcal{F}$ and $j > i$. From Lemma 1, if $i \in \mathcal{F}$ and $j \geq i$, $(x_j - x_i)d \leq 0$. Write ℓ in the form

$$\ell(\beta) = -\sum_{i\in\mathcal{F}} \log \sum_{j\geq i} \exp((x_j - x_i)\beta)$$

Clearly $\ell(\beta + d) > \ell(\beta)$ for any β and hence ℓ has no finite maximizer. •

3 A linear programming algorithm for solving Ad≤0

We are looking for a vector d such that

$$Ad \leq 0 \tag{11}$$

and Ad has as many negative components as possible. This is equivalent to finding a non-positive vector in the column space of A with as many negative components as possible or equivalently a non-positive vector in the row space of A' with as many negative components as possible.

To find a non-positive vector in the row space of A' with as many negative components as possible let $\tilde{A}$ be any matrix whose rows are a basis for the row space of A' and consider the linear programming problem:

$$\min c\delta \quad \text{given} \quad \tilde{A}\delta = b \ , \ \delta \geq 0$$

where b is a non-negative combination of the columns of $\tilde{A}$, c is a row vector of ones, and the minimization is with respect to δ. Because of the way b is chosen, there is at least one non-negative solution δ to the constraint equations $\tilde{A}\delta = b$. The tabular form of this problem is

$$\begin{array}{c|c} \tilde{A} & \text{b} \\ \hline c & 0 \end{array} \tag{12}$$

Apply the simplex algorithm to the tableau (12) and consider the final tableau when the algorithm stops. If the linear programming problem has a solution, the reduced cost vector $\tilde{c} \leq 0$ and $v = \tilde{c} - c < 0$. Moreover, v is a linear combination of the rows of $\tilde{A}$. Thus v is a non-positive vector in the row space of A' that has as many negative components as possible.

It is possible that this is true even if the linear programming problem (12) does not have a solution. This happens when every positive reduced cost has only zeros above it. These reduced costs must equal the corresponding initial costs and hence v will be zero in every column with a positive reduced cost and negative in all others. Clearly v is a linear combination of the rows of A' that has as many negative components as possible.

Finally, what happens if there are positive reduced costs with negative values above them? All other values above such reduced costs must be zero. Hence there must be columns of the coefficient matrix of the form

$$\begin{array}{ccc} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}$$

Clearly if there is a non-positive linear combination of the rows of the current coefficient matrix that is non-zero, it must be a non-positive linear combination of its last two rows. Thus any non-zero solution v must be a linear combination of these two rows. We may proceed by creating a new tableau of the form (12) with $\tilde{A}$ equal to these two rows and applying the simplex algorithm. If this produces a positive reduced cost with one or more negative coefficients above it, we may again reduce the number of rows in $\tilde{A}$. If we run out of rows, the only non-positive vector in the row space of A' is $v = 0$. This procedure may be summarized as follows:

LP Algorithm: Let $\tilde{A}$ be any matrix whose rows are a basis for the row space of A', let b be any non-negative combination of the columns of $\tilde{A}$, and let c be a vector of ones of the same length as the rows of $\tilde{A}$. Form the tableau (12).

1. Apply the simplex algorithm to (12).
2. If there are no positive reduced costs with negative values above them, skip this step. Otherwise delete all rows of the current tableau that contain a negative value above a positive reduced cost. If no rows remain above the reduced costs, set $v = 0$ and stop. Otherwise replace the reduced costs by c and go to 1.
3. Compute $v = \tilde{c} - c$ where $\tilde{c}$ is the reduced cost vector. Stop.

Note that:

- The vector b may be zero.
- If A has full column rank, the initial $\tilde{A}$ may be set equal to A'.
- All $\tilde{A}$ after the initial $\tilde{A}$ are in standard form.
- If the initial $\tilde{A}$ is in standard form it is sufficient to choose any $b \geq 0$. If the components of b are a random sample from the unit interval, with probability one, the initial tableau and all later tableau will have no degenerate basic solutions.

The last two observations are important because when $\tilde{A}$ is in standard form it is easy to put tableau (12) in standard form. Moreover, if there are no degenerate basic solutions, one need not use a linear programing algorithm designed to handle them.

Whatever the value of v, the equation $Ad = v'$ has a solution d. Any such d satisfies (11) and Ad has as many negative components as possible.

4 An algorithm for extended maximum likelihood estimation

Before defining an algorithm let $c_1, \cdots, c_m$ be the distinct values of $x_i d$ for $i \in \mathcal{F}$ and for $r = 1, \cdots, m$ let

$$S_r = \{j : x_j d = c_r\} \tag{13}$$

These are distinct sets. Let $\tilde{\ell}_r$ be the Cox likelihood for the observations in S_r. Then

$$\ell_d = \tilde{\ell}_1 + \cdots + \tilde{\ell}_m$$

The expression on the right is called the Cox likelihood for the stratified data $S_1, \cdots, S_m$. There are standard algorithms for finding finite maximizers of such likelihoods. We assume we have such an algorithm. The following algorithm will find finite and extended maximizers of (1).

EMLE algorithm:

1. Form A as defined by (4) and apply the LP algorithm to find d.
2. Find the strata $S_1, \cdots, S_m$ defined by (13).
3. Apply the stratified Cox regression algorithm to the strata $S_1, \cdots, S_m$ to produce $\hat{\beta}$.

The pair $(\hat{\beta}, d)$ produced by the EMLE algorithm is an extended maximum likelihood estimate. If $Ad = 0$, $\hat{\beta}$ is a finite maximum likelihood estimate.

5 Ties

Assume that $t_1 \leq t_2 \leq \ldots \leq t_n$ so that tied observations are possible. Let $\mathcal{F}$ be the set of failure times and as before assume $t_1 \in \mathcal{F}$. For each $t \in \mathcal{F}$, let $\tilde{R}_t$ denote the set of indices for observations which fail or are censored in the interval $[t, \infty)$, let Q_t denote the indices i such that $t_i = t$, and let q_t denote the number of indices in the set Q_t. Then (see, e.g., Cox and Oakes, 1984, page 103) the log-likelihood for tied observations is

$$\ell(\beta) = \sum_{t \in \mathcal{F}} \ln \left(\frac{q_i! \prod_{j \in Q_i} \exp(x_j \beta)}{\sum_{k \in \tilde{R}_t} \exp(x_k \beta)} \right)$$

This log-likelihood is equivalent to (1) when there are no ties.

The difficulty with ties lies in defining A appropriately. We begin by redefining $\gamma(j)$. Given an observation time t_j, let t be the largest failure time such that $t \leq t_j$. If $Q_t \neq \{j\}$, let $\gamma(j) = Q_t - \{j\}$. Otherwise let $\gamma(j) = Q_s$ where s is the largest failure time such that $s < t_j$. Using this definition, let

$$a_j = x_j - \sum_{i \in \gamma(j)} x_j \tag{14}$$

for $j = 2, \cdots, n$ and let A be the matrix with these vectors as rows. Find a vector d such that $Ad \leq 0$ and Ad has as many negative components as possible. For each $t \in \mathcal{F}$ define the reduced risk set

$$R_t = \{j : t \in \gamma(j) \text{ and } a_j d = 0\} \tag{15}$$

Using these modified definitions, appropriately modified versions of the results in Section 2 may be proved. The only changes required for the EMLE algorithm are to replace (4) by (14) and Definition 1 by (15).

6 Examples

Consider the data:

t_i	x_1	x_2	z
1	3	-1	2
2	5	1	2
3	3	1	1
4	4	2	1
5	3	1	1

In the table $z = (x_1 - x_2)/2$ is a monotone function of the failure times.

The EMLE algorithm was used to fit a Cox regression model to these data. As expected, it split the observations into the two strata identified by z. It required 5 iterations of the stratified Cox regression algorithm to converge. The EMLE algorithm yielded the extended maximum likelihood estimate $(\hat{\beta}, d)$ where

$$\hat{\beta} = \begin{pmatrix} -.6298 \\ 0 \end{pmatrix} \quad , \quad d = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

The same stratified algorithm applied to the single stratum S containing all of the original data required 7 iterations to produce the estimate

$$\tilde{\beta} = \begin{pmatrix} 15.4695 \\ -16.0993 \end{pmatrix}$$

If $\rho = 16.0993$

$$\tilde{\beta} = \hat{\beta} + \rho d$$

to the precision displayed. Also $\ell(\tilde{\beta}) = \ell_d(\hat{\beta}) = -2.2359$ to the precision displayed. The values of $x_i\tilde{\beta}$ were

$$(30.48,\ 30.76,\ -.1767,\ -.8052,\ -.1767) \tag{16}$$

These values appear to identify the same two strata as those identified by the EMLE algorithm.

Because standard routines for proportional hazards models perform no special computations for extended maximum likelihood estimates, it is of interest to use this example as input to them. SURVREG (Preston and Clarkson, 1983), and the user contributed SAS procedure COXREGR (SAS, 1983) reported that the Hessian of the log-likelihood was singular and stopped the analysis, returning essentially nothing. The SYSTAT module SURVIVAL (Steinberg and Colla, 1988), the user contributed SAS procedure PHGLM (SAS, 1983), and the S-plus procedure COXPH (MathSoft, 1995) did somewhat better. They converged with large (absolute) values for the estimated parameters. The usual output statistics were also printed. For all three programs the reported optimal log-likelihood agreed with that given above to the precision displayed. Note, however, in other ways these programs failed. For example they reported finite maximum likelihood estimates when in fact no such estimate exists. The S-plus procedure procedure did give a warning message stating that the estimates might be infinite.

Motivated by a standard algorithm result like (16), one might use it to divide the data into strata and re-apply the standard algorithm. In our example this leads to the correct strata and the second application of the standard algorithm will produce the correct first component of the extended estimate $(\hat{\beta}, d)$ and a maximum value equal to $\sup \ell$. Clarkson (1989) formalized this into an heuristic algorithm. It is attractive because it does not require a linear programming step and involves only a minor modification to a standard algorithm. Unfortunately, it also may not produce correct results. It frequently does, but we are not ready to discuss this here.

In order to reassure the reader that infinite estimates do occur, note that Johnson, Tolley, Bryson, and Goldman (1982, page 693) report an infinite estimate occurrence rate of 22.39% for a Cox regression model with a sample size of 40 and two covariates. They considered an unbalanced two-way analysis of variance with cell means μ_{ij} and cell sizes n_{ij} given by

$$(\mu_{ij}) = \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} \quad (n_{ij}) = \begin{pmatrix} 16 & 4 \\ 4 & 16 \end{pmatrix}$$

The cell means satisify a two-way additive model with zero row effect. Let X be a 40 by 2 design matrix for such an analysis of variance model and let β be choosen so the cell means in $X\beta$ are the values μ_{ij}. The rows of X were used to define relative risks $e^{x_i\beta}$. These in turn were used together with type II censoring to generate 24 failure and 16 censoring times so the resulting survival data represent a random sample from a Cox regression model with design matrix X. Type II censoring means the failure times precede the censorimg times. A total of 10,000 data sets were generated. Of these 2,239 did not have finite estimates.

7 Final comments

It is useful to consider the meaning of solutions at infinity from the point of view of statistical practice. If ℓ does not have a finite maximizer, the second component of the extended estimate $(\hat{\beta}, d)$ defines the reduced risk sets R_i. Of the subjects at risk at failure time t_i, those not in R_i have zero probability of failure and for the subjects $j \in R_i$, the relative risks are $e^{x_j\hat{\beta}}$. If $Ad < 0$, that is if all components of Ad are negative, each R_i contains only one point and the extended estimator predicts failure perfectly. This is actually the usual case when ℓ does not have a finite maximizer.

Assume we have a data set containing males and females with failure represented by uterine cancer. If the males can be seperated from the females by a linear function of the covariates, for example if sex is a covariate, then ℓ does not have a finite maximizer and each R_i contains no males. This means the extended maximum likelihood estimator correctly predicts that males will not fail. One might also discover non-failures of greater interest. For example one might find the R_i also contain no non-smoking females. The extended estimate predicts these cannot fail. This may not be true, but it is consistant with the data set and model used.

To summarise, we have given a number of theoretical results about extended maximum likelihood estimates for Cox proportional-hazards models. These were used along with a linear programming algorithm and a standard Cox regression algorithm for stratified data to produce an algorithm for computing extended maximum likelihood estimates. Because infinite estimates are not common in practice, it may be reasonable to begin with a standard algorithm and if it looks like it is producing an infinite estimate, switch to an extended estimate algorithm. In any event, general programs must deal with the possibility of extended estimates because they do in fact occur.

We would like to thank our reviewer for a number of insights that helped to motivate our work and clearify its presentation.

References

Albert, A., and J. A. Anderson (1984), On the existence of maximum likelihood estimates in logistic regression models, *Biometrika*, **71**, 1–10.

Bryson, Maurice C., and Mark E. Johnson (1981), The incidence of monotone likelihood in the Cox model, *Technometrics*, **23**, 381–384.

Baker, R. J., M. R. B. Clarke, and P. W. Lane (1985), Zero entries in contingency tables, *Computational Statistics and Data Analysis*, **3**, 33–45.

Clarkson, Douglas B. (1989), Computing extended maximum likelihood estimates in monotone likelihood proportional–hazards models, *Proceedings of the 21th Symposium of the Interface*, American Statistical Association, Washington, D. C., in press.

Clarkson, Douglas B., and Robert I. Jennrich (1991), Computing extended maximum likelihood estimates for linear parameter models, *Journal of the Royal Statistical Society*, B, **53**, 417-426.

Cox, D. R. (1972), Regression models and life tables (with discussion), *Journal of the Royal Statistical Society,* Series B*, Methodology*, **34**, 187–220.

Cox, D. R., and D. Oakes (1984), *Analysis of Survival Data*, Chapman and Hall, London.

Elandt-Johnson, Regina C., and Norman L. Johnson (1980), *Survival Models and Data Analysis*, John Wiley & Sons, New York.

Haberman, S. J. (1974), *The Analysis of Frequency Data*, The University of Chicago Press, Chicago.

Hamada, M., and S. K. Tse (1988), A note on the existence of maximum likelihood estimates in linear regression models using interval censored data, *Journal of the Royal Statistics Society, Series B*, **50**, 293–296.

IMSL (1987), STAT/LIBRARY *Users Manual*, Version 1.0, IMSL, Houston.

Johnson, Mark E., H. Dennis Tolley, Maurice C. Bryson, and Aaron S. Goldman (1982), Covariate analysis of survival data: a small-sample study of Cox's model, *Biometrics*, **38**, 685–698.

Kalbfleisch, John D., and Ross L. Prentice (1980), *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, New York.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.

Lee, Elisa T. (1980), *Statistical Methods for Survival Data Analysis*, Lifetime Learning Publications, Belmont, California.

MathSoft (1995), *Guide to Statistical and Mathematical Analysis*, MathSoft Inc., Seattle, Washington.

Preston, Dale L., and Douglas B. Clarkson (1983), SURVREG - an interactive program for the interactive analysis of survival regression models, *The American Statistician*, **37**, 174.

SAS (1983), *SUGI Supplemental Library User's Guide*, SAS Institute Inc., Cary, North Carolina.

Steinberg, Dan and Phillip Colla (1988), *SURVIVAL: A Supplementary module for SYSTAT*, SYSTAT, Inc., Evanston, Illinois.

Silvapulle, M. J., and Burridge J. (1986), Existence of maximum likelihood estimates in regression models for grouped and un grouped data, *Journal of the Royal*
Statistics Society, Series B, **48**, 100–106.