Computational Methods for Semiparametric Linear Regression With Censored Data

D. Y. LIN* AND C. J. GEYER†

Semiparametric linear regression with censored data assumes a linear relationship between failure time and covariates without specifying the distributional form of the error term. This approach has attracted considerable attention recently. Most notably, rank regression methods have been derived for parameter estimation, hypothesis testing, and goodness-of-fit analysis. The implementation of these methods requires minimizing discrete objective functions with multiple local minima. Conventional optimization algorithms cannot be used to solve such minimization problems. We develop computational methods to implement rank regression procedures using simulated annealing. Two real data sets are used for illustration. Applications of the new algorithms to the modified least squares estimator of Buckley and James and several other related problems are also described.

Key Words: Accelerated failure time model; Buckley–James estimator; Gibbs distribution; Rank regression; Simulated annealing; Survival analysis.

1. INTRODUCTION

The semiparametric linear regression model for right-censored data specifies that the failure time Y is related to a $p \times 1$ vector of covariates X in the following way:

$$Y_i = \beta_0' X_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where β_0 is a $p \times 1$ vector of unknown regression parameters and the ε_i's are independent and have an unspecified common survival function F. When Y is expressed on a logarithmic scale, this model is known as the accelerated failure time model (Cox and Oakes [1984] and Kalbfleisch and Prentice [1980]). Model (1.1) provides a useful ad-

*D. Y. Lin is Assistant Professor, Department of Biostatistics, SC–32, University of Washington, Seattle, WA 98195

†C. J. Geyer is Assistant Professor, School of Statistics, University of Minnesota, 270A Vincent Hall, 206 Church St. S.E., Minneapolis, MN 55455

Received July 1991; Revised January 1992

Volume 1, Number 1, pp. 77–90

junct to the popular Cox (1972) proportional hazards model and is especially appealing to practitioners due to its straightforward interpretation.

Several methods were proposed over the years to analyze model (1.1). Most notably, Prentice (1978) suggested inference procedures based on linear rank statistics while Buckley and James (1979) provided a natural extension of the least squares estimator. These methods have not been widely used because neither their theoretical nor their numerical aspects were sufficiently addressed.

Recently there have been significant developments in the theory and methods for the semiparametric linear regression model with censored data. In particular, Tsiatis (1990), Wei, Ying, and Lin (1990), Lai and Ying (1991b), and Ying (in press) formalized and extended the work of Prentice (1978). Moreover, Lai and Ying (1991a) provided the usual large-sample theory for the Buckley–James estimator. These new theoretical and methodological advances have added increasing urgency to the development of efficient computing algorithms.

In this article we provide a unified approach to implementing semiparametric regression methods. We focus on the rank regression procedures derived by Wei et al. (1990). As shown in Section 2, the implementation of these procedures requires minimizing discrete objective functions that generally have multiple local minima. Such complicated features of the objective functions preclude the use of conventional optimization algorithms designed for smooth functions. By contrast, the method of simulated annealing described in Section 3 is suitable for optimization of discrete functions and has a better chance of finding global minima than Newton-type algorithms. In Section 4 we specialize this algorithm in the rank regression setting by using the problem-specific information. Section 5 uses two real data sets to illustrate and evaluate the proposed computational methods. Applications of the current approach to the Buckley–James estimator and several other related problems are described in Section 6.

2. RANK REGRESSION METHODS

Let $Z_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i \leq C_i)$, where C_i is the censoring time for the ith subject and $I(\cdot)$ is the indicator function. Also, let $e_i(\beta) = Z_i - \beta' X_i$ $(i = 1, \ldots, n)$. The linear rank statistic based on $\{e_i(\beta), \delta_i, X_i\}$ $(i = 1, \ldots, n)$ is defined by

$$S(\beta) = \sum_{i=1}^{n} \delta_i \phi(\widehat{F}_\beta(e_i(\beta))) \left\{ X_i - \frac{\sum_{j=1}^{n} I(e_j(\beta) \geq e_i(\beta)) X_j}{\sum_{j=1}^{n} I(e_j(\beta) \geq e_i(\beta))} \right\}, \qquad (2.1)$$

where $\widehat{F}_\beta(\cdot)$ is the left-continuous version of the Kaplan–Meier estimator for $F(\cdot)$ based on $\{e_i(\beta), \delta_i\}$ $(i = 1, \ldots, n)$, and ϕ is a twice continuously differentiable function on $[0, 1]$. The statistic (2.1) corresponds to the log-rank statistic if $\phi(\cdot) = 1$, and to the Peto–Prentice generalization of the Wilcoxon statistic if $\phi(u) = u$.

It can be shown that $n^{-1/2} S(\beta_0)$ converges weakly to a zero-mean p-variate Gaussian variable with a covariance matrix that is asymptotically equivalent to $n^{-1} \Lambda(\beta_0)$, where

$$\Lambda(\beta) = \sum_{i=1}^{n} \delta_i \phi^2(\widehat{F}_\beta(e_i(\beta))) \left[\frac{\sum_{j=1}^n I(e_j(\beta) \geq e_i(\beta)) X_j^{\otimes 2}}{\sum_{j=1}^n I(e_j(\beta) \geq e_i(\beta))} - \left\{ \frac{\sum_{j=1}^n I(e_j(\beta) \geq e_i(\beta)) X_j}{\sum_{j=1}^n I(e_j(\beta) \geq e_i(\beta))} \right\}^{\otimes 2} \right], \tag{2.2}$$

and $a^{\otimes 2} = aa'$ (Andersen, Borgan, Gill, and Keiding 1982). These results suggest that we use $S(\beta)$ as an estimation function and estimate β_0 by solving the system of equations $\{S(\beta) = 0\}$. The statistic $S(\beta)$ depends on β only through the ranks of the residuals $e_i(\beta)$ $(i = 1, \ldots, n)$. Thus $S(\beta)$ changes its value only for perturbations of β that alter the ranks of the residuals. In consequence, $S(\beta)$ is a step function of β, and there are usually no exact solutions to $\{S(\beta) = 0\}$. Furthermore, $S(\beta)$ is, in general, nonmonotone. These complicated features of $S(\beta)$ deprive us of the use of traditional Newton-type algorithms for root finding. As in Wei et al. (1990), we define the rank estimator $\widehat{\beta}$ as the value of β that minimizes $\| S(\beta) \|$, where $\| \cdot \|$ denotes the L_1 norm. Other norms such as $S(\beta)'\Lambda(\beta)^{-1}S(\beta)$ could also be used. All estimators that minimize some norm of $S(\beta)$ are asymptotically equivalent due to asymptotic linearity of $S(\beta)$ in the neighborhood of β_0.

The asymptotic properties of the rank estimator $\widehat{\beta}$ have been studied by Lai and Ying (1991b), Tsiatis (1990), Wei et al. (1990), and Ying (in press). Under suitable regularity conditions, $n^{1/2}(\widehat{\beta} - \beta_0)$ converges weakly to a zero-mean p-variate Gaussian variable. The corresponding covariance matrix, however, is very complicated and cannot be realistically estimated.

Wei et al. (1990) derived practical methods for making inference about β_0 that bypass the estimation of the covariance matrix for $\widehat{\beta}$. Suppose that we are interested in ψ, the first $q(1 \leq q \leq p)$ components of $\beta = (\psi', \eta')'$. Let

$$Q(\eta; \psi_0) = S(\widetilde{\beta})'\Lambda(\widehat{\beta})^{-1}S(\widetilde{\beta}), \tag{2.3}$$

where $\widetilde{\beta} = (\psi_0', \eta')'$. Then, under $H_0 : \psi = \psi_0$, the statistic $G(\psi_0) = \min_\eta Q(\eta; \psi_0)$ is asymptotically distributed as χ_q^2, where χ_q^2 denotes a χ^2 variable with q degrees of freedom.

Wei et al. (1990) also proposed a simple way of checking the adequacy of the linear model. If the assumed model (1.1) is valid, two different rank estimators, say $\widehat{\beta}_{\phi_1}$ and $\widehat{\beta}_{\phi_2}$ with distinct weight functions ϕ_1 and ϕ_2, should be close to each other, which implies that there exists a common value of β such that the two score statistics corresponding to ϕ_1 and ϕ_2, denoted by $S_{\phi_1}(\beta)$ and $S_{\phi_2}(\beta)$, are both close to zero. Wei et al. (1990) showed that the statistic $H(\phi_1, \phi_2) = \min_\beta R(\beta; \phi_1, \phi_2)$ is asymptotically distributed as χ_p^2 under model (1.1), where

$$R(\beta; \phi_1, \phi_2) = \begin{bmatrix} S_{\phi_1}(\beta) \\ S_{\phi_2}(\beta) \end{bmatrix}' \begin{bmatrix} \Lambda_{\phi_1\phi_1}(\widehat{\beta}_{\phi_1}) & \Lambda_{\phi_1\phi_2}(\widehat{\beta}_{\phi_1}) \\ \Lambda_{\phi_2\phi_1}(\widehat{\beta}_{\phi_1}) & \Lambda_{\phi_2\phi_2}(\widehat{\beta}_{\phi_1}) \end{bmatrix}^{-1} \begin{bmatrix} S_{\phi_1}(\beta) \\ S_{\phi_2}(\beta) \end{bmatrix}, \tag{2.4}$$

and $\Lambda_{\phi_k\phi_l}(\beta)(k,l=1,2)$ are obtained from $\Lambda(\beta)$ defined in (2.2) with ϕ^2 replaced by $\phi_k\phi_l$. An unusually large value of $H(\phi_1,\phi_2)$ indicates model misspecification.

The implementation of the aforementioned procedures (i.e., parameter estimation, hypothesis testing, and goodness-of-fit analysis) amounts to minimizing discontinuous objective functions that generally have multiple local minima. Because the functions to be minimized (i.e., $\| S(\beta) \|$, $Q(\eta;\psi_0)$ and $R(\beta;\phi_1,\phi_2)$) depend on their arguments only through the score $S(\beta)$ (or $S(\widetilde{\beta})$), their values change only when the score varies, and this in turn happens only when the orders of the residuals $e_i(\beta)$ $(i=1,\ldots,n)$ are altered. Thus the regions of constancy for the objective function are bounded by hyperplanes that are solutions to the linear equations $e_i(\beta)=e_j(\beta)$ for $i,j=1,\ldots,n$ such that $\delta_i=1$ or $\delta_j=1$, since the score contains no terms depending on both $e_i(\beta)$ and $e_j(\beta)$ if $\delta_i=\delta_j=0$. These hyperplanes cut up the parameter space into a huge number of regions and it is impracticable to examine them all. Efficient search algorithms are needed.

3. SIMULATED ANNEALING

Given an arbitrary real function $U(x)$, there exists a one-parameter exponential family of distributions having densities proportional to $\exp\{-U(x)\theta\}$. These are often referred to as Gibbs distributions, for they are the distributions of a physical system with energy function $U(x)$ in thermal equilibrium at a temperature inversely proportional to θ.

Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) introduced an algorithm for simulating realizations from Gibbs distributions. In each step of the algorithm, the current state x and some other "candidate" state x', the latter being a random displacement of x, are compared. The change in "energy", $\Delta E=U(x')-U(x)$, is calculated. If $\Delta E\le 0$, the candidate is accepted and x' becomes the new state of the system; otherwise, the candidate is accepted with probability $\exp(-\Delta E\cdot\theta)$. If the candidate is rejected, the system remains in state x. The equilibrium distribution of the sequence of iterates has the Gibbs distribution.

Kirkpatrick, Gelatt, and Vecchi (1983) proposed to solve combinatorial optimization problems by performing a sequence of Metropolis algorithms in which the "temperature" is slowly decreased to zero, so that the Gibbs distribution becomes slowly concentrated in the neighborhood of the global minimum of the objective function, which corresponds to the lowest energy of the system. They called this process *simulated annealing*. This optimization technique has been used successfully in many scientific fields. Interesting statistical applications of simulated annealing can be found in Bohachevsky, Johnson, and Stein (1986), Haines (1987), and Lundy (1985).

Results from the theory of Markov chains have been used to derive conditions under which simulated annealing converges with probability 1 to a global minimum. In the special case where the control parameter (i.e., temperature) at the kth step, denoted by c_k, equals $c_0/\log(1+k)$, the condition for convergence is that c_0 be no less than the depth, suitably defined, of the deepest local minimum (Hajek 1986). In any implementation of the algorithm, however, asymptotic convergence can only be approximated. For instance,

one must let c_k go to zero faster than $\{\log(1+k)\}^{-1}$ to obtain a solution in reasonable time. Due to such approximations, the algorithm is no longer guaranteed to find a global minimum with probability 1.

Since simulated annealing is a very general algorithm, its application to a given problem requires several specific choices to be made. These choices require insights into the problem at hand and trial-and-error experiments. In Section 4 we specialize this algorithm in the rank regression setting by examining the general characteristics of the problems both analytically and empirically.

4. SIMULATED ANNEALING FOR RANK REGRESSION

The following algorithm is offered for estimating the parameter vector β_0. In this application of simulated annealing, the random step is generated by a multivariate normal random vector with a diagonal covariance matrix $\{\sigma_j^2; j = 1, \ldots, p\}$. Other random steps could be used instead.

Algorithm 4.1.

1. Set c, β, and $\sigma_j^2 (j = 1, \ldots, p)$ to their initial values.
2. Generate a $p \times 1$ vector of independent variates $\beta^\dagger$, its jth component being normally distributed with mean β_j and variance σ_j^2.
3. Let $D = \| S(\beta^\dagger) \| - \| S(\beta) \|$.
4. If $D \leq 0$, set $\beta = \beta^\dagger$; otherwise, set $\beta = \beta^\dagger$ with probability $\exp(-D/c)$.
5. Decrease c and σ_j^2 $(j = 1, \ldots, p)$ and then go to step 2.

Because there is no way of testing for a global minimum, it is necessary to run the annealing for a fixed number of steps, say m, and stop, declaring the lowest point reached during the run to be the answer. Our task is to find annealing schedules and search variances σ_j^2 $(j = 1, \ldots, p)$ that produce reasonable results for such runs.

The value of the "initial temperature" (c_0) does not seem to make much difference to the performance of the algorithm. It should be of the order of magnitude of the changes in the objective function at the early phase of the search. Unless there is a reason to suspect deep local minima in which the algorithm may be trapped, it is unnecessary to have a very high starting temperature.

Experience has shown that annealing works well when exponential rather than logarithmic cooling rates are used, that is, when c is adjusted in step 5 of the algorithm by $c = \rho c$ for some $\rho < 1$. The value of ρ should be chosen so that $\rho^{m/2}$ is fairly small (say in the order of .0001). Then, for the second half of the run, the annealing algorithm acts essentially as a random search that goes strictly downhill.

There is nothing in the theory of annealing that suggests "cooling" the search variances σ_j^2, but experience has shown that exponential cooling rates are advisable. The initial σ_j should be of the order of magnitude of the distance from the starting point to the minimum. The final σ_j should be of the order of magnitude of the size of the region of constancy that is the minimum of the objective function. If the σ_j^2 are too small, progress toward the minimum will be slow. If the σ_j^2 are too large, the random search

will rarely look at points in the vicinity of the minimum.

Recall that $\| S(\beta) \|$ is the sum of the absolute values of the p components of $S(\beta)$. To avoid extremely uneven contributions from individual components to the value of the objective function, we divide all covariates by their respective sample standard deviations. This standardization can drastically speed up the search especially when covariates are measured on very different scales.

Since the parameter estimators for standardized covariates are usually comparable in magnitude, we set $\sigma_1 = \cdots = \sigma_p = \sigma$. Let us also standardize the response variable Z. Then the parameter vector estimator $\widehat{\beta}^*$ is typically bounded by ± 1, and it seems reasonable to set the initial value of σ to be about .1. To localize the estimate to very small regions of constancy, it may be necessary to cool σ to .001 or smaller by the end of the annealing run. Because the regions of constancy shrink with the sample size, smaller final value of σ should be used for larger sample sizes.

Because there is no way of testing whether the global minimum has been reached, the question naturally arises as to what one can do to gain more confidence in the result. Three suggestions are: First, restart the annealing and see if it converges to the same point; second, do longer annealing runs with slower cooling; and third, start the annealing at different points.

To calculate $G(\psi_0)$ for testing $\psi = \psi_0$, one replaces $S(\beta)$ and β in Algorithm 4.1 by $Q(\eta; \psi_0)$ and η. The guidelines on the selection of annealing schedules and search variances for parameter estimation are also applicable here. It is natural to set the starting point to be $\widehat{\eta}$. This should be close to the final solution under the null hypothesis. The global minimum point can be far away from $\widehat{\eta}$, however, if the covariate effects being tested are highly significant.

Similarly, to implement the goodness-of-fit test, one substitutes $R(\beta; \phi_1, \phi_2)$ for $S(\beta)$ in Algorithm 4.1. Comparisons between parameter estimators with weight functions ϕ_1 and ϕ_2 can facilitate the choice of annealing schedules and search variances. A small value of $H(\phi_1, \phi_2)$ is anticipated if $\widehat{\beta}_{\phi_1}$ and $\widehat{\beta}_{\phi_2}$ are similar. On the other hand, $H(\phi_1, \phi_2)$ can be substantial when there are considerable differences between $\widehat{\beta}_{\phi_1}$ and $\widehat{\beta}_{\phi_2}$ in the components that are highly significant.

Our extensive experimentation with real and simulated data sets has indicated that the following default choices achieve satisfactory empirical performance.

(a) *The choice of initial solution.* Set the initial value of β (or η) to zero or any other reasonable guesses.

(b) *The choice of steps.* Set m to approximately 1,000 times the dimension of the argument, that is, $1{,}000p$ for parameter estimation and goodness-of-fit test, and $1{,}000(p - q)$ for testing covariate effects.

(c) *The choice of initial temperature.* For parameter estimation, let c_0 be smaller than, but in the same order of magnitude as, $\| S(\beta^{(0)}) \|$. For testing covariate effects, let c_0 be of the order of magnitude of $\chi^2_q(.90)$, the 90% percentile point of the χ^2_q distribution. For goodness-of-fit test and for parameter estimation that minimizes $S(\beta)'\Lambda^{-1}(\beta)S(\beta)$, let c_0 be of the order of magnitude of $\chi^2_p(.90)$.

(d) *The choice of cooling rate for temperature.* Choose a ρ such that $\rho^{m/2} \approx .0005$.

(e) *The choice of initial search variance.* Set the initial σ to be roughly .1.

(f) *The choice of cooling rate for search variance.* Choose a cooling rate such that the final σ is about .0005.

5. EXAMPLES

In this section we illustrate and evaluate the techniques proposed in Section 4 using two real data sets. The response variable and covariates are standardized in all calculations, and the results reported in the text are expressed under these standardizations unless otherwise indicated. Readers will notice that the algorithm parameters used in these two examples are consistent with our recommendations in Section 4.

5.1 STANFORD HEART TRANSPLANT DATA

The Stanford heart transplant data as of February 1980 were described in Miller and Halpern (1982). Following these authors, we regressed the base 10 logarithm of the survival time against *age* and age^2 with the 152 patients who survived at least 10 days after entering the study. Out of these 152 patients, 97 had died by February 1980. In our analysis, the variable age was centered at 41.7 (approximate sample mean of the patients' ages) to improve interpretability and to avoid numerical instabilities. With this centering the sample standard deviations for *age* and age^2 are 10.62557 and 165.52726 and that of Z is .67066.

Figures 1 through 3 (pp. 84–86) show the result of an annealing run for estimating β_0 with the log-rank weight function, which finds the apparent global minimum of the norm of the score. The parameter estimate $\widehat{\beta}^* = (-.59396, -.40477)$, where the value of the objective function $\| S(\beta) \|$ is .02267, and the actual score $S(\beta) = (-.0015, -.0212)$. On the original scales the parameter estimate $\widehat{\beta} = (-.037489, -.001640)$, which confirms the result obtained by Wei et al. (1990) via a grid search.

Figure 1 plots the value of the objective function versus iteration number. The last uphill step was at the 1,000th iteration. Earlier on the algorithm took two large uphill steps, which could have gotten it out of deep local minima. At the end of the annealing run, σ has been cooled so that in the last 200 iterations 21 distinct regions of constancy are inspected—not so few that the algorithm is just looking inside one region and not so many that the algorithm is not "seeing" the discreteness of the objective function.

Figure 2 shows the whole path of the annealing run from its start at the origin to its finish in the cluster of points near $(-.6, -.4)$, along with approximate contours of the objective function. A contour plot of the much larger region in which both parameters run from -10 to 10 (not shown) revealed no other local minima on the coarse scale of the plot.

Figure 3 displays the end of the annealing run magnified. The discreteness of the objective function is now noticeable. The apparent global minimum is the small quadrangle labeled 23 near the center of the plot, in which there are two accepted points. The region in the upper left labeled 127 is a local minimum. The density of points inspected

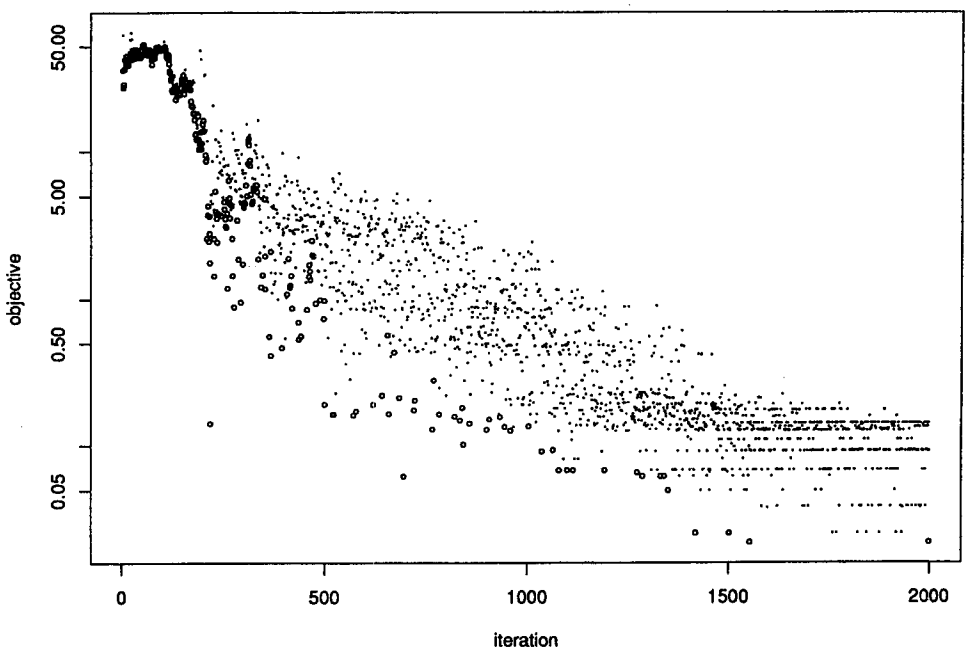

Figure 1. Annealing run for parameter estimation with the log-rank weight function for the Stanford heart transplant data. Plot of objective function value $\| S(\beta) \|$ *versus iteration number. Small dots are rejected candidate points and large hollow dots are accepted moves of the annealing. The algorithm parameters are* $m = 2{,}000$, $\beta^{(0)} = 0$, $c_0 = 10$, *initial* $\sigma = .1$, *and cooling rates .993 for c and .997 for* σ. *Note log scale for the y axis.*

shows that there is a fair chance of a run with these parameters finding this global minimum. The area of global minimum, however, is very tiny; therefore, it is possible for a particular run to miss this area.

As Figure 3 indicates, the rank estimates inside the global minimum region may differ in the fifth decimal point. In addition, the solutions in areas 23 and 26 are identical if rounded to the third decimal point.

To study the behavior of the procedure, we reran the annealing program 200 times using the same algorithm parameters. Sixty of these 200 runs stopped in the global minimum region and 138 runs stopped in the adjacent region labeled 26 in Figure 3 (that is, 99% of the runs stopped at either .02267 or .02615). The remaining two runs stopped at the quite high objective function values of 4.03143 and 18.37084. These results indicate that the answer from the first run was not merely "lucky" and that the algorithm parameters used were good examples for this problem. In ordinary practice one would not make so many runs; we have done this only to illustrate the performance of the algorithm.

For testing $\beta_1 = 0$, it is possible to examine all the intervals of constancy for $Q(\beta_2; 0)$. The global minimum of $Q(\beta_2; 0)$ with the log-rank weight function is 15.02773 for $-.0236 \leq \beta_2 \leq -.0231$. A grid search would require grid size .0001. We ran the annealing algorithm 100 times with $m = 800$, $\eta^{(0)} = -.40477$, $c_0 = 1$, initial $\sigma = .1$,

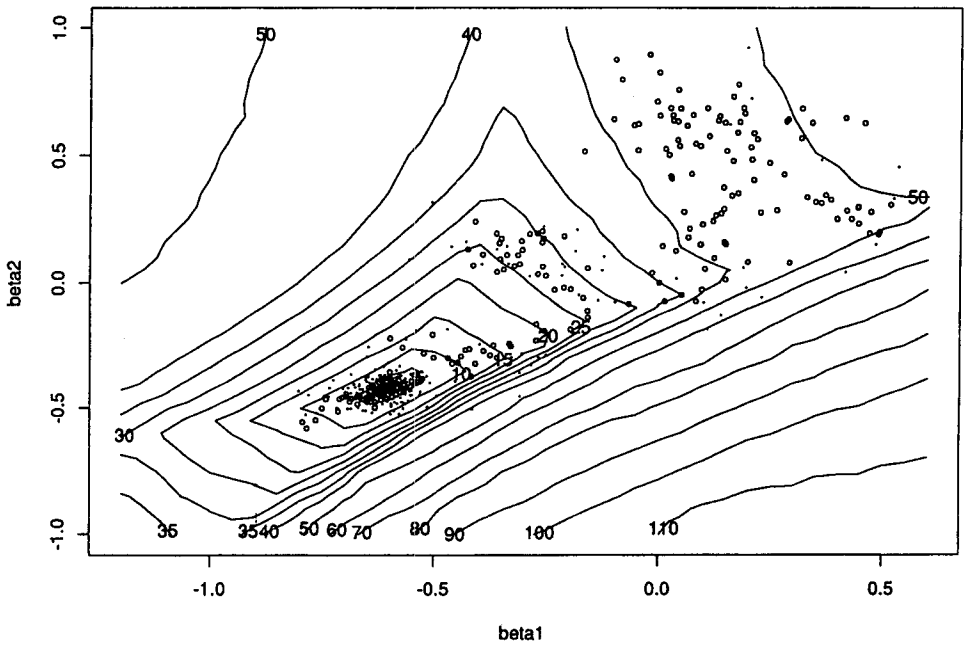

Figure 2. Annealing run for the Stanford heart transplant data (same run as in Figure 1). Plot of points in parameter space along with contours of the objective function. Small dots are rejected points and large hollow dots are accepted points.

and cooling rates .99 for c and .995 for σ. They all stopped at the global minimum region. The same algorithm parameters were also used to test $\beta_2 = 0$, and the global minimum 5.15456 was found in all 100 annealing runs.

For parameter estimation and hypothesis testing with the Peto–Prentice weight function, we used the same sets of algorithm parameters used for the log-rank weight function. The performance of the algorithm with this weight function was similar to that of the log-rank weight function. The rank estimate was $\widehat{\beta}^* = (-.56170, -.40749)$, at which the objective function $\| S(\beta) \|$ equals .01584. The G statistics for testing $\beta_1 = 0$ and $\beta_2 = 0$ are 15.16309 and 6.21079.

To calculate the goodness-of-fit test statistic based on the log-rank versus Peto–Prentice weight functions, we set $\beta^{(0)}$ to be the rank estimate with the log-rank weight function, the initial temperature to be 1, and the initial σ to be .1. The number of steps was 2,000, and the cooling rates were .993 for c and .997 for σ. At the starting point $R(\beta; \phi_1, \phi_2) = 1.01426$, which is rather small because the parameter estimates using these two weight functions are very close. In practice, one could have terminated the program at the starting point because any statistic less than 1.01 will be nonsignificant when compared with the χ^2_2 distribution. The minimum found in the first annealing run was .80536 at $\beta = (-.54282, -.40336)$. This result was reproduced in 75 out of 100 reruns with the given parameters. Thirteen of these reruns found a slightly smaller objective function value .80517. The answers of the remaining 12 runs ranged from .80850 to .87226.

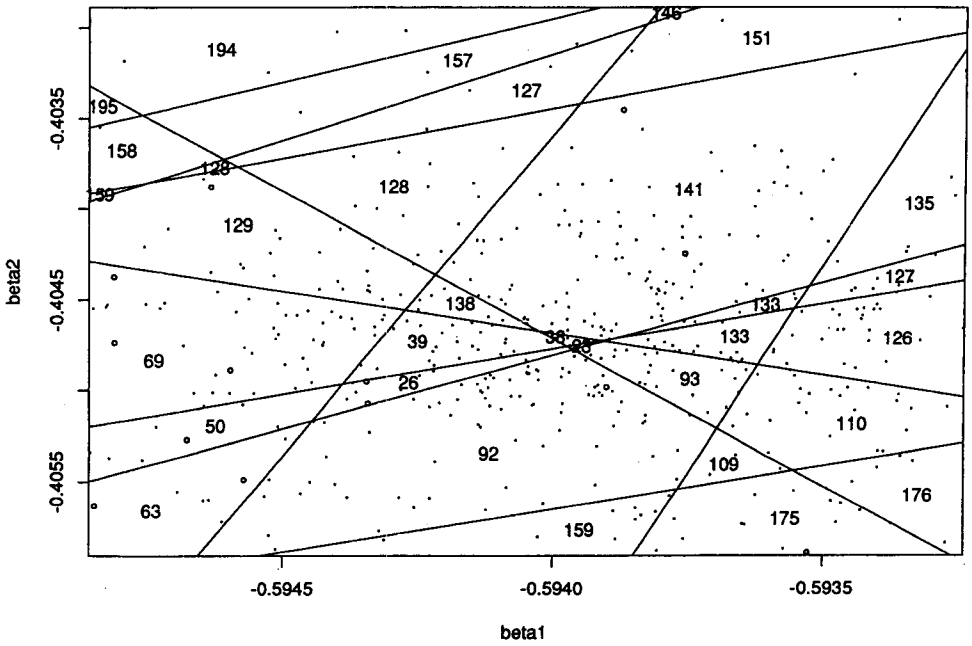

Figure 3. Magnification of central region of Figure 2. Lines are boundaries of regions of constancy. Numbers are values of the objective function $\| S(\beta) \|$ *times 1,000. Small dots are rejected points and large hollow dots are accepted points. The apparent global minimum is the small quadrangle with objective function label 23 (actual value .02267) near the center. An extremely small triangular region at the right end of the global minimum has the value .141 but is too small to be labeled.*

5.2 MAYO PBC LIVER DATA

The Mayo Clinic developed a data base of 418 patients having primary biliary cirrhosis (PBC) who were referred to the clinic between January 1974 and May 1984. PBC is a fatal chronic liver disease of unknown cause. Because PBC is a rare disease, the Mayo data base has been a valuable resource to liver specialists. These data are listed in Appendix D.1 of Fleming and Harrington (1991). As of the date of data listings, 161 patients had died.

The Cox regression method and comprehensive data from the 418 Mayo patients were used by Dickson, Grambsch, Fleming, Fisher, and Langworthy (1989) to derive a natural history model for PBC based on patients' age, total serum bilirubin and serum albumin concentrations, prothrombin time, and the severity of edema. The first two columns of Table 1 provide the variable transformations and regression results for that model.

For comparisons we regressed the natural logarithm of the survival time against the same covariates used by Dickson et al. (1989). The sample standard deviation of Z was .83769, and those of the five covariates in the orders shown in Table 1 were 10.44717, .12816, 1.02380, .25342, and .08812.

The following annealing parameters were used for estimating β_0: $m = 5,000$, $c_0 = 10$, $\rho = .997$, initial $\sigma = .1$, and cooling rate for $\sigma = .9985$. We performed 20 such runs for the log-rank weight function. It is much harder to hit the global minimum

Table 1. Regression Analyses of the Mayo Clinic PBC Data

Variables	Cox model	Linear model $\phi(u) = 1$	Linear model $\phi(u) = u$
Age			
$\hat{\beta}$	.0394	−.0265	−.0271
χ_1^2	26.53	20.23	21.74
$\log_e$ (Albumin)			
$\hat{\beta}$	−2.5328	1.6594	1.5985
χ_1^2	15.27	9.44	9.92
$\log_e$ (Bilirubin)			
$\hat{\beta}$	.8707	−.5831	−.5890
χ_1^2	111.03	68.11	76.62
Edema			
$\hat{\beta}$	.8592	−.6921	−.7974
χ_1^2	10.04	7.36	10.25
$\log_e$ (Prothrombin time)			
$\hat{\beta}$	2.3797	−1.8773	−2.2983
χ_1^2	9.64	6.90	9.39

NOTE: Wald tests are used for the Cox model.

this time than in the Stanford example because larger sample size and larger dimension of covariate vector create many more distinct regions of constancy. Because the areas are very tiny, the estimates from all these runs are nearly identical up to the third decimal point even though the objective function values agree only to one significant digit. The apparent global minimum .27178 was reached in four runs. The rank estimate $\widehat{\beta}^* \approx (-.32993, .25388, -.71271, -.20938, -.19748)$. Similarly, we obtained the rank estimate $\widehat{\beta}^*$ with the Peto–Prentice weight function as (−.33836, .24456, −.71984, −.24122, −.24177), corresponding to the objective function value .01618. These two rank estimates are expressed on the original scales in Table 1. Note that the linear regression estimates have opposite signs from the Cox estimates since larger failure time corresponds to lower hazard rate.

Table 1 also provides the values of the G statistics for testing individual covariate effects with the log-rank and Peto–Prentice weight functions. Each of these numbers was the smallest value of the minima identified in five annealing runs and is to be compared with the χ_1^2 distribution. We used the same algorithm parameters used for parameter estimation except that the initial points were the $\widehat{\eta}$'s. The results from the five runs for a given testing problem were fairly close. For example, the five minima for testing $\beta_1 = 0$ were 20.25695, 20.22655, 20.25028, 20.25880, and 20.28705.

The estimates of β_4 and β_5 are considerably different between the log-rank and Peto–Prentice weight functions. These two parameters, however, are the least significant among the five covariates. Thus we anticipated a moderate value of $H(\phi_1, \phi_2)$ with these two weight functions. Using the algorithm parameters described above, but setting $\beta^{(0)}$ to be the rank estimate with the log-rank weight function, we obtained 6.08591, 6.14921, 6.07037, 6.07107, and 6.11649 as the minima of $R(\beta; \phi_1, \phi_2)$ in five annealing runs. In

comparison with the χ^2_5 distribution, the observed statistic 6.07 is not significant at the 10% level.

6. REMARKS

Results from Section 5 demonstrate that the method of simulated annealing works well for the rank regression procedures. Running our FORTRAN-77 programs with double precision on a SUN4/20 workstation, it took about 1.5 and 15 minutes of processing time to perform 1,000 iterations for the Stanford and the Mayo data sets. An IBM RS/6000 workstation was about 10 times faster. Such computing times should be acceptable to most practitioners now that fast workstations are widely accessible. A portable automatic program with default choices built in but overridable by the users is available from the StatLib archive.

The proposed techniques can also be used to implement semiparametric linear regression methods based on different estimation functions. We now consider the Buckley–James estimation procedure. When there is no censoring, the least squares estimator $\hat{\beta}_L$ for β_0 is the solution to the following system of equations:

$$\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \beta' X_i) = 0, \tag{6.1}$$

where $\overline{X}$ is the vector of sample means for X. In the presence of censored observations, Buckley and James (1979) suggested that in (6.1) Y_i be replaced by

$$\tilde{Y}_i(\beta) = Z_i + (1 - \delta_i)\int_{e_i(\beta)}^{\infty} \hat{F}_\beta(u)du / \hat{F}_\beta(e_i(\beta)),$$

where $e_i(\beta)$ and $\hat{F}_\beta(\cdot)$ are defined as in the beginning of Section 2. Let $L(\beta)$ denote the resulting estimation function $\sum_{i=1}^{n}(X_i - \overline{X})\{\tilde{Y}_i(\beta) - \beta' X_i\}$. Similar to $S(\beta)$, $L(\beta)$ is a discontinuous and nonmonotone function of β, and one may define the estimator $\hat{\beta}_L$ as the value of β that minimizes $\| L(\beta) \|$.

With a slight modification to get around some technical difficulties, Lai and Ying (1991a) proved that $n^{-1/2}L(\beta_0)$ converges in distribution to a p-variate normal with mean 0 and with a covariance matrix that can be consistently estimated. These authors also established the asymptotic normality of $n^{1/2}(\hat{\beta}_L - \beta_0)$. As in the case of rank regression, however, the limiting covariance matrix of $n^{1/2}(\hat{\beta}_L - \beta_0)$ is very complicated and cannot be reliably estimated in applications. Lin and Wei (in press a) proposed a statistic, which is in the same form as $G(\psi_0)$ of Section 2, for making inference about subsets of regression parameters. Thus the techniques developed in Section 4 for the rank regression procedures can also be used for estimating and testing β_0 based on the Buckley–James estimation function.

For survival studies where each subject may experience two or more events or failures, it is natural to assume that each marginal failure time variable satisfies a linear regression model. Lin and Wei (in press b) derived simple procedures for making inference involving parameters of different failure time variables based on marginal linear

rank statistics and Buckley–James estimation functions. Their test statistics are again the minima of certain quadratic forms. Similar statistics were obtained by Lee, Wei, and Ying (submitted manuscript) for highly stratified failure time data. Therefore, the methods of Section 4 can be applied to these situations.

ACKNOWLEDGMENTS

This work was supported by NIH Grant 5 R01 AI 29168 (for Lin) and NSF Grant DMS-9007833 (for Geyer). The authors are grateful to Alun Thomas for advice on simulated annealing. They also thank the editor, an associate editor, and two referees for their helpful comments and suggestions.

REFERENCES

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1982), "Linear Nonparametric Tests for Comparison of Counting Processes, With Applications to Censored Data," *International Statistical Review*, 50, 219–258.

Bohachevsky, I. O., Johnson, M. E., and Stein, M. L. (1986), "Generalized Simulated Annealing for Function Optimization," *Technometrics*, 28, 209–217.

Buckley, J., and James, I. (1979), "Linear Regression With Censored Data," *Biometrika*, 66, 429–436.

Cox, D. R. (1972), "Regression Models and Life-Tables" (With Discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

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

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. (1989), "Prognosis in Primary Biliary Cirrhosis: Model for Decision Making," *Hepatology*, 10, 1–7.

Fleming, T. R., and Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, New York: Wiley.

Haines, L. M. (1987), "The Application of the Annealing Algorithm to the Construction of Exact Optimal Designs for Linear-Regression Models," *Technometrics*, 29, 439–447.

Hajek, B. (1986), "Optimization by Simulated Annealing: A Necessary and Sufficient Condition for Convergence," in *Adaptive Statistical Procedures and Related Topics*, ed. J. V. Ryzin, Hayward, CA: Institute of Mathematical Statistics, pp. 417–427.

Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Kirkpatrick, S., Gelatt, C. D., Jr., and Vecchi, M. P. (1983), "Optimization by Simulated Annealing," *Science*, 220, 671–680.

Lai, T. L., and Ying, Z. (1991a), "Large Sample Theory of a Modified Buckley–James Estimator for Regression Analysis With Censored Data," *The Annals of Statistics*, 19, 1370–1402.

——— (1991b), "Rank Regression Methods for Left Truncated and Right Censored Data," *The Annals of Statistics*, 19, 531–556.

Lee, E. W., Wei, L. J., and Ying, Z. (1992), "Linear Regression Analysis for Highly Stratified Failure Time Data," unpublished manuscript, submitted to *Journal of the American Statistical Association*.

Lin, J. S., and Wei, L. J. (in press a), "Linear Regression Analysis Based on Buckley–James Estimation Equations," *Biometrics*.

——— (in press b), "Linear Regression Analysis for Multivariate Failure Time Observations," *Journal of the American Statistical Association*.

Lundy, M. (1985), "Applications of the Annealing Algorithm to Combinatorial Problems in Statistics," *Biometrika*, 72, 191–198.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1092.

Miller, R., and Halpern, J. (1982), "Regression With Censored Data," *Biometrika*, 69, 521–531.

Prentice, R. L. (1978), "Linear Rank Tests With Right Censored Data," *Biometrika*, 65, 167–179.

Tsiatis, A. A. (1990), "Estimating Regression Parameters Using Linear Rank Tests for Censored Data," *The Annals of Statistics*, 18, 354–372.

Wei, L. J., Ying, Z., and Lin, D. Y. (1990), "Linear Regression Analysis of Censored Survival Data Based on Rank Tests," *Biometrika*, 77, 845–851.

Ying, Z. (in press), "A Large Sample Study of Rank Estimation for Censored Regression Data," *The Annals of Statistics*.