Maximum Likelihood Estimation for Type I Censored Weibull Data Including Covariates

Fritz Scholz*
Mathematics & Computing Technology
Boeing Phantom Works

August 22, 1996

Revised January 5, 2001[1]

*The author wishes to thank James King for several inspiring conversation concerning the uniqueness theorem presented in Section 2, Siavash Shahshahani for showing him more than 30 years ago that it follows from Morse Theory, John Betts for overcoming programming difficulties in using HDNLPR, and Bill Meeker for helpful comments on computational aspects.

[1]The revision mainly corrects misprints, some notational inconsistencies, and indicates the new calling sequence for compiling the Fortran program

Abstract

This report deals with the specific problem of deriving maximum likelihood estimates of the regression model parameters when the residual errors are governed by a Gumbel distribution. As an additional complication the observed responses are permitted to be type I or multiply censored. Since the log-transform of a 2-parameter Weibull random variable has a Gumbel distribution, the results extend to Weibull regression models, where the log of the Weibull scale parameter is modeled linearly as a function of covariates. In the Weibull regression model the covariates thus act as multiplicative modifiers of the underlying scale parameter.

A general theorem for establishing a unique global maximum of a smooth function is presented. The theorem was previously published by Mäkeläinen et al. (1981) with a sketch of a proof. The proof presented here is much shorter than their unpublished proof.

Next, the Gumbel/Weibull regression model is introduced together with its censoring mechanism. Using the above theorem the existence and uniqueness of maximum likelihood estimates for the posed specific Weibull/Gumbel regression problem for type I censored responses is characterized in terms of sufficient and easily verifiable conditions, which are conjectured to be also necessary.

As part of an efficient optimization algorithm for finding these maximum likelihood estimates it is useful to have good starting values. These are found by adapting the iterative least squares algorithm of Schmee and Hahn (1979) to the Gumbel/Weibull case. FORTRAN code for computing the maximum likelihood estimates was developed using the optimization routine HDNLPR. Some limited experience of this algorithm with simulated data is presented as well as the results to a specific example from Gertsbakh (1989).

1 Introduction

In the theory of maximum likelihood estimation it is shown, subject to regularity conditions, that the likelihood equations have a consistent root. The problems that arise in identifying the consistent root among possibly several roots were discussed by Lehmann (1980). It is therefore of interest to establish, whenever possible, that the likelihood equations have a unique root. For example, for exponential family distributions it is easily shown, subject to mild regularity conditions, that the log-likelihood function is strictly concave which in turn entails that the log-likelihood equations have at most one root. However, such global concavity cannot always be established. Thus one may ask to what extent the weaker property of local concavity of the log-likelihood function at all roots of the likelihood equations implies that there can be at most one root. Uniqueness arguments along these lines, although incomplete, may be found in Kendall and Stuart (1973, p. 56), Turnbull (1974), and Copas (1975), for example.

However, it also was pointed out by Tarone and Gruenhage (1975) that a function of two variables may have an infinity of strict local maxima and no other critical points, i.e. no saddle points or minima. To resolve this issue, a theorem is presented which is well known to mathematicians as a special application of Morse Theory, cf. Milnor (1963) and also Arnold (1978) p. 262. Namely, on an island the number of minima minus the number of saddle points plus the number of maxima is always one. The specialization of the theorem establishing conditions for a unique global maximum was first presented to the statistical community by Mäkeläinen et al. (1981). Since Morse Theory is rather deep and since Mäkeläinen et al. only give an outline of a proof, leaving the lengthy details to a technical report, a short (one page) and more accessible proof is given here. It is based on the elementary theory of ordinary differential equations.

It is noted here that although Mäkeläinen et al. have priority in publishing the theorem presented here, a previous version of this paper had been submitted for publication, but was withdrawn and issued as a technical report (Scholz, 1981), when the impending publication of Mäkeläinen et al. became known. Aside from these two independent efforts there was a third by Barndorff-Nielsen and Blæsild (1980), similarly preempted, which remained as a technical report. Their proof of the result appears to depend on Morse Theory. Similar results under weaker assumptions may be found in Gabrielsen (1982, 1986). Other approaches, via a multivariate version of Rolle's theorem were examined in Rai and van Ryzin (1982).

2 The Uniqueness Theorem

In addition to the essential strict concavity at all critical points the uniqueness theorem invokes a compactness condition which avoids the problems pointed out by Tarone and Gruenhage (1975) and which are illustrated in Figure 1. The theorem can be stated as follows:

Theorem 1 *Let G be an open, connected subset of R^n and let $f : G \longrightarrow R$ be twice continuously differentiable on G with the following two properties:*

i) For any $x \in G$ with grad $f(x) = 0$ the Hessian $D^2 f(x)$ is negative definite, i.e. all critical points are strict local maxima.

ii) For any $x \in G$ the set $\{y \in G : \ f(y) \geq f(x)\}$ is compact.

Then f has exactly one critical point, hence one global maximum and no other local maxima on G.

Proof: By i) all critical points are isolated, i.e. for each critical point $x \in G$ of f there exists and $\epsilon(x) > 0$ such that

$$B_{\epsilon(x)}(x) = \{y \in R^n : \ |y - x| < \epsilon(x)\} \subset G$$

contains no other critical point besides x, and such that

$$g(x) \stackrel{\text{def}}{=} \sup \Big\{ f(y) : \ y \in \partial B_{\epsilon(x)}(x) \Big\} < f(x) \ .$$

Let

$$U_{d(x)}(x) = \Big\{ y \in B_{\epsilon(x)}(x) : \ f(y) > f(x) - d(x) \Big\}$$

with $0 < d(x) < f(x) - g(x)$, then $\partial U_{d(x)}(x) \subset B_{\epsilon(x)}(x)$ (note that $f(y) = f(x) - d(x)$ for $y \in \partial U_{d(x)}(x)$). Consider now the following vector function

$$h(z) = \text{grad } f(z) \ \cdot \ | \text{ grad } f(z) \ |^{-2}$$

which is well defined and continuously differentiable on $G - C$, where C is the set of critical points of f in G. Hence the differential equation $\dot{z}(t) = h(z(t))$ with initial

condition $z(0) = z_0 \in G - C$ has a unique right maximal solution $z(t; 0, z_0)$ on some interval $[0, t_0), t_0 > 0$, see Hartman (1964), pp. 8-13. Note that $f(z(t; 0, z_0)) = f(z_0) + t$ for $t \in [0, t_0)$. It follows from ii) that t_0 must be finite. Consider now the following compact set:

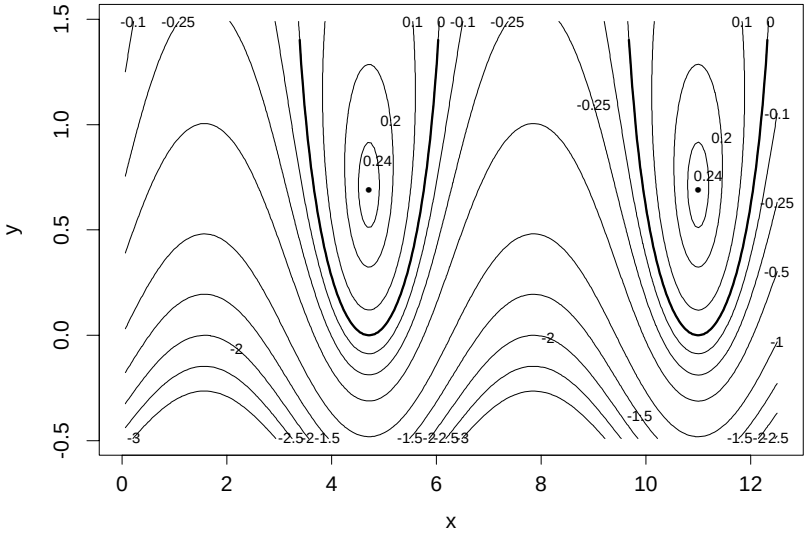
$$K = \{y \in G : \ f(y) \geq f(z_0)\} - \bigcup_{x \in C} U_{d(x)}(x) \ .$$

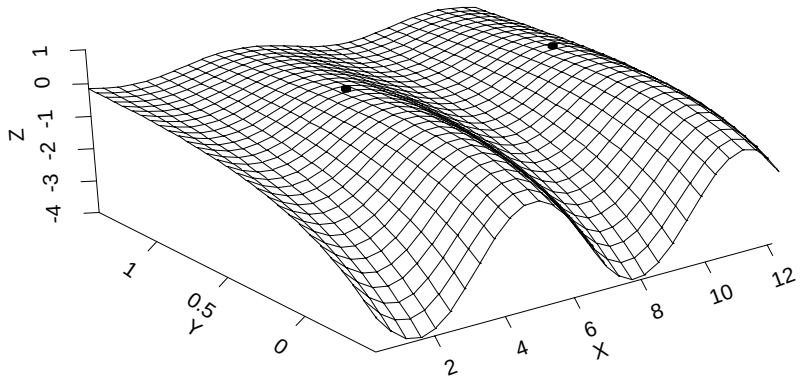
Then $z(t; 0, z_0) \notin K$ for t near t_0, see Hartman pp. 12-13. Hence for t near t_0 $z(t; 0, z_0) \in U_{d(x)}(x)$ for some $x \in C$. From the construction of $U_{d(x)}(x)$ it is clear that once such a solution enters $U_{d(x)}(x)$ it will never leave it. For $x \in C$ let $P(x)$ be the set set containing x and all those points $z_0 \in G - C$ whose solutions $z(t; 0, z_0)$ will wind up in $U_{d(x)}(x)$. It has been shown that $\{P(x) : c \in C\}$ forms a partition of G. Since $z(t; 0, z_0)$ is a continuous function of $z_0 \in G - C$, see Hartman p. 94, it follows that each $P(x)$, $x \in C$ is open. Since G is assumed to be connected, i.e., G cannot be the disjoint union of nonempty open sets, one concludes that all but one of the $P(x)$, $x \in C$, must be empty. Q.E.D.

Remark: It is clear that a disconnected set G allows for easy counterexamples of the theorem. Assumption ii) is violated in the example presented by Tarone and Gruenhage: $f(x, y) = -\exp(-2y) - \exp(-y)\sin(x)$. Figure 1 shows the contour lines of $f(x, y)$ in the upper plot and the corresponding perspective in the lower plot. In thicker line width is indicated the contour $f(x, y) = 0$, given by $y = -\log(-\sin(x))$ over the intervals where $\sin(x) < 0$. This latter contour is unbounded since $y \to \infty$ as $\sin(x) \to 0$ at those interval endpoints. Thus the level set $\{(x, y) : f(x, y) \geq 0\}$ is unbounded. What is happening in this example is that there are saddle points at infinity which act as the connecting agent between the local maxima.

Assumption ii) may possibly be replaced by weaker assumptions; however, it appears difficult to formulate such assumptions without impinging on the simplicity of theorem and proof. The following section will illustrate the utility of the theorem in the context of censored Weibull data with covariates. However, it should be noted that many other examples exist.

Figure 1: Contour and Perspective Plots
of $f(x, y) = -\exp(-2y) - \exp(-y)\sin(x)$





3 Weibull Regression Model Involving Censored Data

Consider the following linear model:

$$y_i = \sum_{j=1}^{p} u_{ij}\beta_j + \sigma\epsilon_i = \boldsymbol{u}_i'\boldsymbol{\beta} + \sigma\epsilon_i \qquad i = 1, \ldots, n$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent random errors, identically distributed according to the extreme value or Gumbel distribution with density $f(x) = \exp[x - \exp(x)]$ and cumulative distribution function $F(x) = 1 - \exp[-\exp(x)]$. The $n \times p$ matrix $\boldsymbol{U} = (u_{ij})$ of constant regression covariates is assumed to be of full rank p, with $n > p$. The unknown parameters $\sigma, \beta_1, \ldots, \beta_p$ will be estimated by the method of maximum likelihood, which here is taken to be the solution to the likelihood equations.

The above model can also arise from the following Weibull regression model:

$$P(T_i \leq t) = 1 - \exp\left(-\left[\frac{t}{\alpha(\boldsymbol{u}_i)}\right]^{\gamma}\right)$$

which, after using the following log transformation $Y_i = \log(T_i)$, results in

$$P(Y_i \leq y) = 1 - \exp\left[-\exp\left(\frac{y - \log[\alpha(\boldsymbol{u}_i)]}{(1/\gamma)}\right)\right] = 1 - \exp\left[-\exp\left(\frac{y - \mu(\boldsymbol{u}_i)}{\sigma}\right)\right] \ .$$

Using the identifications $\sigma = 1/\gamma$ and $\mu(\boldsymbol{u}_i) = \log[\alpha(\boldsymbol{u}_i)] = \boldsymbol{u}_i'\boldsymbol{\beta}$ this reduces to the previous linear model with the density f.

Rather than observing the responses y_i completely, the data are allowed to be censored, i.e., for each observation y_i one either observes it or some censoring time c_i. The response y_i is observed whenever $c_i \geq y_i$ and otherwise one observes c_i, and one knows whether the observation is a y_i or a c_i. One will also always know the corresponding covariates $u_{ij}, j = 1, \ldots, p$ for $i = 1, \ldots, n$. Such censoring is called type I censoring or, since the censoring time points c_i can take on multiple values, one also speaks of multiply censored data. Thus the data consist of

$$\boldsymbol{S} = \{(x_1, \delta_1, \boldsymbol{u}_1), \ldots, (x_n, \delta_n, \boldsymbol{u}_n)\} \ ,$$

where $x_i = y_i$ and $\delta_i = 1$ when $y_i \leq c_i$, and $x_i = c_i$ and $\delta_i = 0$ when $y_i > c_i$. The number of uncensored observations is denoted by $r = \sum_{i=1}^{n} \delta_i$ and the index sets of uncensored and censored observations by $\mathcal{D}$ and $\mathcal{C}$, respectively, i.e.,

$$\mathcal{D} = \{i : \delta_i = 1, i = 1, \ldots, n\} = \{i_1, \ldots, i_r\} \quad \text{and} \quad \mathcal{C} = \{i : \delta_i = 0, i = 1, \ldots, n\} \; .$$

Furthermore, denote the uncensored observations and corresponding covariates by

$$\boldsymbol{y}_{\mathcal{D}} = \begin{pmatrix} y_{i_1} \\ \vdots \\ y_{i_r} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{U}_{\mathcal{D}} = \begin{pmatrix} \boldsymbol{u}'_{i_1} \\ \vdots \\ \boldsymbol{u}'_{i_r} \end{pmatrix} \; .$$

The likelihood function of the data $\boldsymbol{S}$ is

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i \in \mathcal{D}} \frac{1}{\sigma} \exp\left[\frac{x_i - \boldsymbol{u}'_i\boldsymbol{\beta}}{\sigma} - \exp\left(\frac{x_i - \boldsymbol{u}'_i\boldsymbol{\beta}}{\sigma}\right)\right] \prod_{i \in \mathcal{C}} \exp\left[-\exp\left(\frac{x_i - \boldsymbol{u}'_i\boldsymbol{\beta}}{\sigma}\right)\right]$$

and the corresponding log-likelihood is

$$\begin{aligned} \ell(\boldsymbol{\beta}, \sigma) &= \log[L(\boldsymbol{\beta}, \sigma)] \\ &= \sum_{i \in \mathcal{D}} \left[\frac{x_i - \boldsymbol{u}'_i\boldsymbol{\beta}}{\sigma} - \exp\left(\frac{x_i - \boldsymbol{u}'_i\boldsymbol{\beta}}{\sigma}\right)\right] - \sum_{i \in \mathcal{D}} \log \sigma - \sum_{i \in \mathcal{C}} \exp\left(\frac{x_i - \boldsymbol{u}'_i\boldsymbol{\beta}}{\sigma}\right) \; . \end{aligned}$$

3.1 Conditions for Unique Maximum Likelihood Estimates

Here conditions will be stated under which the maximum likelihood estimates of $\boldsymbol{\beta}$ and σ exist and are unique. It seems that this issue has not yet been addressed in the literature although software for finding the maximum likelihood estimates exists and is routinely used. Some problems with such software have been encountered and situations have been discovered in which the maximum likelihood estimates, understood as roots of the likelihood equations

$$\frac{\partial \ell(\boldsymbol{\beta}, \sigma)}{\partial \sigma} = 0 \qquad \text{and} \qquad \frac{\partial \ell(\boldsymbol{\beta}, \sigma)}{\partial \beta_j} = 0 \; , \qquad \text{for} \quad j = 1, \ldots, p \; , \tag{1}$$

do not exist. Thus it seems worthwhile to explicitly present the conditions which guarantee unique solutions to the likelihood equations. These conditions appear to be

reasonable and not unduely restrictive. In fact, it is conjectured that these conditions are also necessary, but this has not been pursued.

Theorem 2 *Let $r \geq 1$ and the columns of $\boldsymbol{U}_{\mathcal{D}}$ be linearly independent. Then for $\boldsymbol{y}_{\mathcal{D}}$ not in the column space of $\boldsymbol{U}_{\mathcal{D}}$ or for $\boldsymbol{y}_{\mathcal{D}} = \boldsymbol{U}_{\mathcal{D}}\widehat{\boldsymbol{\beta}}$ for some $\widehat{\boldsymbol{\beta}}$ with $x_i > \boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}$ for some $i \in \mathcal{C}$ the likelihood equations* (1) *have a unique solution which represents the location of the global maximum of $\ell(\boldsymbol{\beta}, \sigma)$ over $R^p \times (0, \infty)$.*

Comments: The above assumption concerning $\boldsymbol{U}_{\mathcal{D}}$ is stronger than assuming that the columns of $\boldsymbol{U}$ be linearly independent. Also, the event that $\boldsymbol{y}_{\mathcal{D}}$ is in the column space of $\boldsymbol{U}_{\mathcal{D}}$ technically has probability zero if $r > p$, but may occur due to rounding or data granularity problems.

When $r \geq p$ and $\boldsymbol{y}_{\mathcal{D}} = \boldsymbol{U}_{\mathcal{D}}\widehat{\boldsymbol{\beta}}$ with $x_i \leq \boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}$ for all $i \in \{\mathcal{C}\}$, it is easily seen that $\ell(\widehat{\boldsymbol{\beta}}, \sigma) \to \infty$ as $\sigma \to 0$. From the point of view of likelihood maximization this would point to $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}) = (\widehat{\boldsymbol{\beta}}, 0)$ as the maximum likelihood estimates, provided one extends the permissible range of σ from $(0, \infty)$ to $[0, \infty)$. However, the conventional large sample normality theory does not apply here, since it is concerned with the roots of the likelihood equations.

The additional requirement $x_i > \boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}$ for some $i \in \mathcal{C}$ gives the extra information that is needed to get out of the denenerate case, namely the linear pattern $\boldsymbol{y}_{\mathcal{D}} = \boldsymbol{U}_{\mathcal{D}}\widehat{\boldsymbol{\beta}}$, because the actual observation y_i implied by the censored case $x_i > \boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}$ will also satisfy that inequality since $y_i > x_i$ and thus break the linear pattern and yield a $\widehat{\sigma} > 0$. This appears to have been overlooked by Nelson (1982) when on page 392 he suggests that when estimating k parameters one should have at least k distinct failure times, otherwise the estimates do not exist. Although his recommendation was made in a more general context it seems that the conditions of Theorem 2 may have some bearing on other situations as well.

Proof: First it is shown that any any critical point $(\boldsymbol{\beta}, \sigma)$ of ℓ is a strict local maximum. In the process the equations resulting from grad $\ell(\boldsymbol{\beta}, \sigma) = \mathbf{0}$ are used to simplify the Hessian or matrix of second derivatives of ℓ at those critical points. This simplified Hessian is then shown to be negative definite. The condition grad $\ell(\boldsymbol{\beta}, \sigma) = \mathbf{0}$ results in the following equations:

$$\frac{\partial \ell}{\partial \sigma} = -\frac{r}{\sigma} - \sum_{i \in \mathcal{D}} \frac{x_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma^2} + \sum_{i=1}^{n} \frac{x_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma^2} \exp\left(\frac{x_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)$$

$$= \quad -\frac{r}{\sigma} - \frac{1}{\sigma}\left[\sum_{i\in\mathcal{D}} z_i - \sum_{i=1}^{n} z_i \exp(z_i)\right] = 0 \tag{2}$$

with $z_i = (x_i - \boldsymbol{u}_i'\boldsymbol{\beta})/\sigma$ and

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= -\frac{1}{\sigma}\left[\sum_{i\in\mathcal{D}} u_{ij} - \sum_{i=1}^{n} u_{ij}\exp\left(\frac{x_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right)\right] \\ &= -\frac{1}{\sigma}\left[\sum_{i\in\mathcal{D}} u_{ij} - \sum_{i=1}^{n} u_{ij}\exp\left(z_i\right)\right] = 0 \qquad \text{for } j = 1,\ldots,p\,. \end{aligned} \tag{3}$$

The Hessian or matrix $\boldsymbol{H}$ of second partial derivatives of ℓ with respect to $(\boldsymbol{\beta}, \sigma)$ is made up of the following terms for $1 \le j, k \le p$:

$$\frac{\partial^2 \ell(\boldsymbol{\beta},\sigma)}{\partial\beta_j \partial\beta_k} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} u_{ij}u_{ik}\exp(z_i) \tag{4}$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta},\sigma)}{\partial\beta_j \partial\sigma} = \frac{1}{\sigma^2}\left[\sum_{i\in\mathcal{D}} u_{ij} - \sum_{i=1}^{n} u_{ij}\exp(z_i) - \sum_{i=1}^{n} u_{ij} z_i \exp(z_i)\right] \tag{5}$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta},\sigma)}{\partial\sigma^2} = \frac{1}{\sigma^2}\left[r + 2\sum_{i\in\mathcal{D}} z_i - 2\sum_{i=1}^{n} z_i\exp(z_i) - \sum_{i=1}^{n} z_i^2\exp(z_i)\right] \tag{6}$$

From (2) one gets

$$\sum_{i=1}^{n} z_i \exp(z_i) - \sum_{i\in\mathcal{D}} z_i = r$$

and one can simplify (6) to

$$\frac{\partial^2 \ell(\boldsymbol{\beta},\sigma)}{\partial\sigma^2} = \frac{r}{\sigma^2} - \frac{2r}{\sigma^2} - \frac{1}{\sigma^2}\sum_{i=1}^{n} z_i^2\exp(z_i) = -\frac{1}{\sigma^2}\left[r + \sum_{i=1}^{n} z_i^2\exp(z_i)\right]$$

Using (3) one can simplify (5) to

$$\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma)}{\partial \beta_j \partial \sigma} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} z_i u_{ij} \exp(z_i) \; .$$

Thus the matrix $\boldsymbol{H}$ of second partial derivatives of ℓ at any critical point is

$$\boldsymbol{H} = -\frac{1}{\sigma^2} \begin{pmatrix} \sum_{i=1}^{n} \exp(z_i) \boldsymbol{u}_i \boldsymbol{u}_i' & \sum_{i=1}^{n} z_i \exp(z_i) \boldsymbol{u}_i \\ \sum_{i=1}^{n} z_i \exp(z_i) \boldsymbol{u}_i' & r + \sum_{i=1}^{n} z_i^2 \exp(z_i) \end{pmatrix} = -\frac{1}{\sigma^2} \, \boldsymbol{B}$$

Letting $w_i = \exp(z_i) / \sum_{j=1}^{n} \exp(z_j)$ and $W = \sum_{j=1}^{n} \exp(z_j)$ one can write

$$\boldsymbol{B} = W \begin{pmatrix} \sum_{i=1}^{n} w_i \boldsymbol{u}_i \boldsymbol{u}_i' & \sum_{i=1}^{n} w_i z_i \boldsymbol{u}_i \\ \sum_{i=1}^{n} w_i z_i \boldsymbol{u}_i' & r/W + \sum_{i=1}^{n} w_i z_i^2 \end{pmatrix} \; .$$

In this matrix the upper $p \times p$ left diagonal submatrix $\sum_{i=1}^{n} w_i \boldsymbol{u}_i \boldsymbol{u}_i'$ is positive definite. This follows from

$$\boldsymbol{a}' \sum_{i=1}^{n} w_i \boldsymbol{u}_i \boldsymbol{u}_i' \boldsymbol{a} = \sum_{i=1}^{n} w_i |\boldsymbol{a}' \boldsymbol{u}_i|^2 > 0$$

for every $\boldsymbol{a} \in R^p - \{\boldsymbol{0}\}$, provided the columns of $\boldsymbol{U}$ are linearly independent, which follows from our assumption about $\boldsymbol{U}_{\mathcal{D}}$. The lower right diagonal element $r + W \sum_{i=1}^{n} w_i z_i^2$ of $\boldsymbol{B}$ is positive since $r \geq 1$.

The last step in showing $\boldsymbol{B}$ to be positive definite is to verify that $\det(\boldsymbol{B}) > 0$. To this end let

$$\boldsymbol{V} = \left(\sum_{i=1}^{n} w_i \boldsymbol{u}_i \boldsymbol{u}_i' \right)^{-1}$$

and note that for $r > 0$ one has

$$\begin{aligned} \det(\boldsymbol{B}) \;\; &= \;\; W \, \det \left(\sum_{i=1}^{n} w_i \boldsymbol{u}_i \boldsymbol{u}_i' \right) \\ &\qquad \times \, \det \left[r/W + \sum_{i=1}^{n} w_i z_i^2 - \sum_{i=1}^{n} w_i z_i \boldsymbol{u}_i' \boldsymbol{V} \sum_{i=1}^{n} w_i z_i \boldsymbol{u}_i \right] > 0 \end{aligned}$$

since

$$
\begin{aligned}
0 \;&\leq\; \sum_{i=1}^{n} w_i \left[z_i - \boldsymbol{u}_i' \boldsymbol{V} \sum_{j=1}^{n} w_j z_j \boldsymbol{u}_j \right]^2 \\
&=\; \sum_{i=1}^{n} w_i z_i^2 - 2 \sum_{i=1}^{n} w_i z_i \boldsymbol{u}_i' \boldsymbol{V} \sum_{j=1}^{n} w_j z_j \boldsymbol{u}_j + \sum_{i=1}^{n} w_i \boldsymbol{u}_i' \boldsymbol{V} \sum_{j=1}^{n} w_j z_j \boldsymbol{u}_j \; \boldsymbol{u}_i' \boldsymbol{V} \sum_{j=1}^{n} w_j z_j \boldsymbol{u}_j \\
&=\; \sum_{i=1}^{n} w_i z_i^2 - 2 \sum_{i=1}^{n} w_i z_i \boldsymbol{u}_i' \boldsymbol{V} \sum_{j=1}^{n} w_j z_j \boldsymbol{u}_j + \sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j z_j \boldsymbol{u}_j' \boldsymbol{V} \boldsymbol{u}_i \boldsymbol{u}_i' \boldsymbol{V} \sum_{j=1}^{n} w_j z_j \boldsymbol{u}_j \\
&=\; \sum_{i=1}^{n} w_i z_i^2 - 2 \sum_{i=1}^{n} w_i z_i \boldsymbol{u}_i' \boldsymbol{V} \sum_{j=1}^{n} w_j z_j \boldsymbol{u}_j + \sum_{j=1}^{n} w_j z_j \boldsymbol{u}_j' \boldsymbol{V} \sum_{i=1}^{n} w_i \boldsymbol{u}_i \boldsymbol{u}_i' \boldsymbol{V} \sum_{j=1}^{n} w_j z_j \boldsymbol{u}_j \\
&=\; \sum_{i=1}^{n} w_i z_i^2 - \sum_{i=1}^{n} w_i z_i \boldsymbol{u}_i' \boldsymbol{V} \sum_{i=1}^{n} w_i z_i \boldsymbol{u}_i \; .
\end{aligned}
$$

To claim the existence of unique maximum likelihood estimates it remains to demonstrate the compactness condition ii) of Theorem 1. It will be shown that

a) $\ell(\boldsymbol{\beta}, \sigma) \longrightarrow -\infty$ uniformly in $\boldsymbol{\beta} \in R^p$ as $\sigma \to 0$ or $\sigma \to \infty$ and

b) for any $\epsilon > 0$ and $\epsilon \leq \sigma \leq 1/\epsilon$ one has
$\sup\{\ell(\boldsymbol{\beta}, \sigma) : |\boldsymbol{\beta}| \geq \rho\} \longrightarrow -\infty$ as $\rho \to \infty$.

Compact sets in R^{p+1} are characterized by being bounded and closed. Using the continuous mapping $\psi(\boldsymbol{\beta}, \sigma) = (\boldsymbol{\beta}, \log(\sigma))$ map the half space $K^+ = R^p \times (0, \infty)$ onto R^{p+1}. According to Theorem 4.14 of Rudin (1976) ψ maps compact subsets of K^+ into compact subsets of R^{p+1}, the latter being characterized as closed and bounded. This allows the characterization of compact subsets in K^+ as those that are closed and for which $|\boldsymbol{\beta}|$ and σ are bounded above and for which σ is bounded away from zero.

Because of the continuity of $\ell(\boldsymbol{\beta}, \sigma)$ the set

$$Q_0 = \{(\boldsymbol{\beta}, \sigma) : \ell(\boldsymbol{\beta}, \sigma) \geq \ell(\boldsymbol{\beta}_0, \sigma_0)\}$$

is closed and bounded and bounded away from the hyperplane $\sigma = 0$. These boundedness properties of Q_0 are seen by contradiction. If Q_0 did not have these properties, then there would be a sequence $(\boldsymbol{\beta}_n, \sigma_n)$ with either $\sigma_n \to 0$ or $\sigma_n \to \infty$ or with $0 < \epsilon < \sigma_n < 1/\epsilon$ and $|\boldsymbol{\beta}_n| \to \infty$. For either of these two cases the above claims a) and b) state that $\ell(\boldsymbol{\beta}_n, \sigma_n) \to -\infty$ which violates $\ell(\boldsymbol{\beta}_n, \sigma_n) \geq \ell(\boldsymbol{\beta}_0, \sigma_0)$. This completes the main argument of the proof of Theorem 2, subject to demonstrating the claims a) and b).

To see a) first deal with the case in which $\boldsymbol{y}_{\mathcal{D}}$ is not in the column space of $\boldsymbol{U}_{\mathcal{D}}$. This entails that for all $\boldsymbol{\beta} \in R^p$

$$|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{U}_{\mathcal{D}}\boldsymbol{\beta}| \geq |\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{U}_{\mathcal{D}}\widehat{\boldsymbol{\beta}}| = \kappa > 0 \qquad \text{where} \qquad \widehat{\boldsymbol{\beta}} = [\boldsymbol{U}'_{\mathcal{D}}\boldsymbol{U}_{\mathcal{D}}]^{-1}\boldsymbol{U}'_{\mathcal{D}}\boldsymbol{y}_{\mathcal{D}} \, .$$

Thus $\max\{|x_i - \boldsymbol{u}'_i\boldsymbol{\beta}| : i \in \mathcal{D}\} \geq \tilde{\kappa} > 0$ uniformly in $\boldsymbol{\beta} \in R^p$ and, using the inequality $x - \exp(x) \leq -|x|$ for all $x \in R$, one has

$$\begin{aligned} \sum_{i\in\mathcal{D}} \left[\frac{x_i - \boldsymbol{u}'_i\boldsymbol{\beta}}{\sigma} - \exp\left(\frac{x_i - \boldsymbol{u}'_i\boldsymbol{\beta}}{\sigma} \right) \right] &\leq -\sum_{i\in\mathcal{D}} \left| \frac{x_i - \boldsymbol{u}'_i\boldsymbol{\beta}}{\sigma} \right| \qquad (7) \\ &\leq -\frac{\max\{|x_i - \boldsymbol{u}'_i\boldsymbol{\beta}| : i \in \mathcal{D}\}}{\sigma} \leq -\frac{\tilde{\kappa}}{\sigma} \longrightarrow -\infty \, , \end{aligned}$$

as $\sigma \to 0$, and thus uniformly in $\boldsymbol{\beta} \in R^p$

$$\ell(\boldsymbol{\beta}, \sigma) \leq -r\log(\sigma) - \frac{\tilde{\kappa}}{\sigma} \longrightarrow -\infty \qquad \text{as} \qquad \sigma \to 0 \, .$$

To deal with the other case, where $\boldsymbol{U}_{\mathcal{D}}\widehat{\boldsymbol{\beta}} = \boldsymbol{y}_{\mathcal{D}}$ and $x_i > \boldsymbol{u}'_i\widehat{\boldsymbol{\beta}}$ for some $i \in \mathcal{C}$, take a neighborhood of $\widehat{\boldsymbol{\beta}}$

$$B_\rho(\widehat{\boldsymbol{\beta}}) = \left\{ \boldsymbol{\beta} : |\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}| \leq \rho \right\}$$

with $\rho > 0$ chosen sufficiently small so that

$$|\boldsymbol{u}'_i\boldsymbol{\beta} - \boldsymbol{u}'_i\widehat{\boldsymbol{\beta}}| \leq \frac{x_i - \boldsymbol{u}'_i\widehat{\boldsymbol{\beta}}}{2} \qquad \text{for all } \boldsymbol{\beta} \in B_\rho(\widehat{\boldsymbol{\beta}}).$$

This in turn implies

$$\begin{aligned} x_i - \boldsymbol{u}_i'\boldsymbol{\beta} &= x_i - \boldsymbol{u}_i'\widehat{\boldsymbol{\beta}} + \boldsymbol{u}_i'\widehat{\boldsymbol{\beta}} - \boldsymbol{u}_i'\boldsymbol{\beta} \\ &\geq x_i - \boldsymbol{u}_i'\widehat{\boldsymbol{\beta}} - \frac{x_i - \boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}}{2} = \frac{x_i - \boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}}{2} \end{aligned}$$

for all $\boldsymbol{\beta} \in B_\rho(\widehat{\boldsymbol{\beta}})$. For some $\kappa' > 0$ and for all $\boldsymbol{\beta} \notin B_\rho(\widehat{\boldsymbol{\beta}})$ one has $|\boldsymbol{y}_{\mathcal{D}} - \boldsymbol{U}_{\mathcal{D}}\boldsymbol{\beta}| \geq \kappa'$. Bounding the first term of the likelihood $\ell(\boldsymbol{\beta}, \sigma)$ as in (7) for all $\boldsymbol{\beta} \notin B_\rho(\widehat{\boldsymbol{\beta}})$ and bounding the last term of the likelihood by

$$-\exp\left(\frac{x_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma}\right) \leq -\exp\left(\frac{x_i - \boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}}{2\sigma}\right) \qquad \text{for all } \boldsymbol{\beta} \in B_\rho(\widehat{\boldsymbol{\beta}})$$

one finds again that either of these bounding terms will dominate the middle term, $-r\log\sigma$, of $\ell(\boldsymbol{\beta}, \sigma)$ as $\sigma \to 0$. Thus again uniformly in $\boldsymbol{\beta} \in R^p$ one has $\ell(\boldsymbol{\beta}, \sigma) \to -\infty$ as $\sigma \to 0$.

As for $\sigma \to \infty$ note $x - \exp(x) \leq -1$ and one has

$$\ell(\boldsymbol{\beta}, \sigma) \leq -r\log(\sigma) - r \longrightarrow -\infty \qquad \text{as} \qquad \sigma \to \infty$$

uniformly in $\boldsymbol{\beta} \in R^p$. This establishes a).

Now let $\epsilon \leq \sigma \leq 1/\epsilon$. From our assumption that the columns of $\boldsymbol{U}_{\mathcal{D}}$ are linearly independent it follows that

$$\inf\{|\boldsymbol{U}_{\mathcal{D}}\boldsymbol{\beta}| : \ |\boldsymbol{\beta}| = 1\} = m > 0$$

where m^2 is the smallest eigenvalue of $\boldsymbol{U}_{\mathcal{D}}'\boldsymbol{U}_{\mathcal{D}}$. Thus for all $\boldsymbol{\beta} \in R^p$

$$|\boldsymbol{U}_{\mathcal{D}}\boldsymbol{\beta} - \boldsymbol{y}_{\mathcal{D}}| \geq |\boldsymbol{U}_{\mathcal{D}}\boldsymbol{\beta}| - |\boldsymbol{y}_{\mathcal{D}}| \geq m|\boldsymbol{\beta}| - |\boldsymbol{y}_{\mathcal{D}}| \ ,$$

and using the inequality $\sum |x_i| \geq \sqrt{\sum x_i^2}$ one has

$$\begin{aligned} \ell(\boldsymbol{\beta}, \sigma) &\leq -r\log(\sigma) - \sum_{i \in \mathcal{D}} \left| \frac{x_i - \boldsymbol{u}_i'\boldsymbol{\beta}}{\sigma} \right| \leq -r\log(\sigma) - \frac{|\boldsymbol{U}_{\mathcal{D}}\boldsymbol{\beta} - \boldsymbol{y}_{\mathcal{D}}|}{\sigma} \\ &\leq -r\log(\sigma) - \frac{m|\boldsymbol{\beta}| - |\boldsymbol{y}_{\mathcal{D}}|}{\sigma} \longrightarrow -\infty \qquad \text{as} \qquad |\boldsymbol{\beta}| \longrightarrow \infty \ , \end{aligned}$$

again uniformly in $|\boldsymbol{\beta}| \geq K$, with $K \longrightarrow \infty$. This concludes the proof of Theorem 2.

4 Solving the Likelihood Equations

The previous section showed that the solution to the likelihood equations is unique and coincides with the unique global maximum of the likelihood function. This section discusses some computational issues that arise in solving for these maximum likelihood estimates. One can either use a multidimensional root finding algorithm to solve the likelihood equations or one can use an optimization algorithm on the likelihood or log-likelihood function. It appears that in either case one can run into difficulties when trying to evaluate the exponential terms $\exp([x_i - \boldsymbol{u}_i'\boldsymbol{\beta}]/\sigma)$. Depending on the choice of σ and $\boldsymbol{\beta}$ this term could easily overflow and terminate all further calculation. Such overflow leads to a likelihood that is practically zero, indicating that σ and $\boldsymbol{\beta}$ are far away from the optimum. It seems that this problem is what troubles the algorithm `survreg` in S-PLUS. In some strongly censored data situations `survreg` simply crashes with overflow messages. One such data set is given in Table 1 with a dagger indicating the three failure times. The histogram of this data set is given in Figure 2 with the three failure cases indicated by dots below the histogram.

Table 1: Heavily Censored Sample

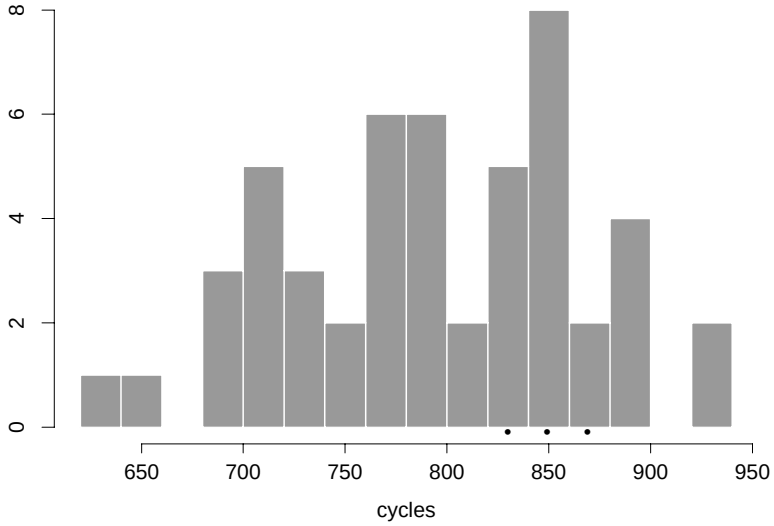
626.1	651.7	684.7	686.3	698.2	707.7	709.8	714.7	718.0	719.6
720.9	721.9	726.7	740.3	752.9	760.3	764.0	764.8	768.3	773.6
774.4	784.1	785.3	788.9	790.3	793.2	794.0	806.1	816.2	825.0
826.5	$829.8^{\dagger}$	832.3	839.4	840.5	843.1	845.2	849.1	$849.2^{\dagger}$	856.2
856.8	859.1	$868.9^{\dagger}$	869.3	881.1	887.8	890.5	898.2	921.5	934.8

In the case of simple Weibull parameter estimation without covariates this overflow problem can be finessed in the likelihood equations by rewriting these equations so that the exponential terms only appear simultaneously in numerator and denominator of some ratio, see equation (4.2.2) in Lawless (1982). One can then use a common scaling factor so that none of the exponential terms overflow.

In the current case with covariates it appears that this same trick will not work. Thus it is proposed to proceed as follows. Find a starting value $(\boldsymbol{\beta}_0, \sigma_0)$ by way of the Schmee-Hahn regression algorithm presented below. It is assumed that the starting value will not suffer from the overflow problems mentioned before.

Next, employ an optimization algorithm that allows for the possibility that the function to be optimized may not be able to return a function value, gradient or Hessian

Figure 2: Histogram for Data Table 1



at a desired location. In that case the optimization algorithm should reduce its step size and try again. The function box which calculates the function value, gradient and Hessian should take care in trapping exponential overflow problems, i.e., state when they cannot be resolved. These problems typically happen only far away from the function optimum where the log-likelihood drops off to $-\infty$.

Another precaution is to switch from σ to $\alpha = \log(\sigma)$ in the optimization process. Furthermore, it was found that occasionally it was useful to rescale σ, x_i and u_{ij} by a common scale factor so that σ is in the vicinity of one. This is easily done using the preliminary Schmee-Hahn estimates.

Optimization algorithms usually check convergence based on the gradient (among other criteria) and the gradient is proportional to the scale of the function to be optimized. Thus it is useful to rescale the log-likelihood to get its minimum value into a proper range, near one. This can be done approximately by evaluating the absolute value of the log-likelihood at the initial estimate and rescale the log-likelihood function by dividing by that absolute value.

4.1 Schmee-Hahn Regression Estimates with Censored Data

This method was proposed by Schmee and Hahn (1979) as a simple estimation method for dealing with type I censored data with covariates. It can be implemented by using a least squares algorithm in iterative fashion.

We assume the following regression model

$$Y_i = \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \sigma e_i \ , \qquad i = 1, \ldots, n$$

or in vector/matrix notation

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & \ldots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \ldots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \ \sigma \ \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

or more compactly

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{e} \ .$$

Here $\boldsymbol{Y}$ is the vector of observations, $\boldsymbol{X}$ is the matrix of covariates corresponding to $\boldsymbol{Y}$, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\sigma\boldsymbol{e}$ is the vector of independent and identially distributed error terms with $E(e_i) = 0$ and $\text{var}(e_i) = 1$. We denote the density of e by $g_0(z)$. Often one has $X_{i1} = 1$ for $i = 1, \ldots, n$. In that case the model has an intercept.

Rather than observing this full data set $(\boldsymbol{Y}, \boldsymbol{X})$ one observes the Y_i in partially censored form, i.e., there are censoring values $\boldsymbol{c}' = (c_1, \ldots, c_n)$ such that Y_i is observed whenever $Y_i \leq c_i$, otherwise the value c_i is observed. Also, it is always known whether the observed value is a Y_i or a c_i. This is indicated by a $\delta_i = 1$ and $\delta_i = 0$, respectively. Thus the observed censored data consist of

$$\mathcal{D} = (\widetilde{\boldsymbol{Y}}, \boldsymbol{X}, \boldsymbol{\delta})$$

where $\boldsymbol{\delta}' = (\delta_1, \ldots, \delta_n)$ and $\widetilde{\boldsymbol{Y}}' = (\widetilde{Y}_1, \ldots, \widetilde{Y}_n)$ with

$$\widetilde{Y}_i = \begin{cases} Y_i & \text{if } \delta_i = 1, \quad \text{i.e. when } Y_i \leq c_i \\ \\ c_i & \text{if } \delta_i = 0, \quad \text{i.e. when } Y_i > c_i \end{cases}$$

Based on this data the basic algorithm consist in treating the observations initially as though they are not censored and apply the least squares method to $(\widetilde{\boldsymbol{Y}}, \boldsymbol{X})$ to find initial estimates $(\widehat{\sigma}_0, \widehat{\boldsymbol{\beta}}_0')$ of $(\sigma, \boldsymbol{\beta}')$.

Next, replace the censored values by their expected values, i.e., replace $\widetilde{Y}_i$ by

$$\widetilde{Y}_{i,1} = E(Y_i | Y_i > c_i \, ; \sigma, \boldsymbol{\beta}') \qquad \text{whenever} \quad \delta_i = 0 \, ,$$

computed by setting $(\sigma, \boldsymbol{\beta}') = (\widehat{\sigma}_0, \widehat{\boldsymbol{\beta}}_0')$. Denote this modified $\widetilde{\boldsymbol{Y}}$ vector by $\widetilde{\boldsymbol{Y}}_1$. Again treat this modified data set as though it is not censored and apply the least squares method to $(\widetilde{\boldsymbol{Y}}_1, \boldsymbol{X})$ to find new estimates $(\widehat{\sigma}_1, \widehat{\boldsymbol{\beta}}_1')$ of $(\sigma, \boldsymbol{\beta}')$. Repeat the above step of replacing censored $\widetilde{Y}_i$ values by estimated expected values

$$\widetilde{Y}_{i,2} = E(Y_i | Y_i > c_i \, ; \sigma, \boldsymbol{\beta}') \qquad \text{whenever} \quad \delta_i = 0 \, ,$$

this time using $(\sigma, \boldsymbol{\beta}') = (\widehat{\sigma}_1, \widehat{\boldsymbol{\beta}}_1')$. This process can be iterated until some stopping criterion is satisfied. Either the iterated regression estimates $(\widehat{\sigma}_j, \widehat{\boldsymbol{\beta}}_j')$ do not change much any more or the residual sum of squares has stabilized.

In order to carry out the above algorithm one needs to have a computational expression for

$$E(Y | Y > c \, ; \sigma, \boldsymbol{\beta}') \, ,$$

where

$$Y = \beta_1 x_1 + \ldots + \beta_p x_p + \sigma e = \mu(\boldsymbol{x}) + \sigma e$$

and the error term e has density $g_0(z)$. Then Y has density

$$g(y) = \frac{1}{\sigma} \, g_0\!\left(\frac{y - \mu(\boldsymbol{x})}{\sigma} \right) \, .$$

The conditional density of Y, given that $Y > c$, is

$$g_c(y) = \begin{cases} g(y)/[1 - G(c)] & \text{for } \; y > c \\ 0 & \text{for } \; y \leq c \, . \end{cases}$$

The formula for $E(Y | Y > c \, ; \sigma, \boldsymbol{\beta}')$ is derived for two special cases, namely for $g_0(z) = \varphi(z)$, the standard normal density with distribution function $\Phi(z)$, and for

$$g_0(z) = \delta \tilde{g}_0(\delta z - \gamma) = \delta \exp\left[\delta z - \gamma - \exp(\delta z - \gamma)\right] \, ,$$

where $\delta = \pi/\sqrt{6} \approx 1.28255$ and $\gamma \approx .57721566$ is Euler's constant. Here $\tilde{g}_0(z) = \exp\left[z - \exp(z)\right]$ is the standard form of the Gumbel density with mean $-\gamma$ and standard deviation δ. Thus $g_0(z)$ is the standardized density with mean zero and variance

one. The distribution function of $g_0(z)$ is denoted by $\mathcal{G}_0(z) = \tilde{\mathcal{G}}_0(\delta z - \gamma)$ and is given by

$$\mathcal{G}_0(z) = 1 - \exp(-\exp[\delta z - \gamma]) \; .$$

The Gumbel distribution is covered for its intimate connection to the Weibull distribution.

When $g_0(z) = \varphi(z)$ and utilizing $\varphi'(z) = -z\varphi(z)$ one finds

$$\begin{aligned}
E(Y|Y > c\,;\sigma,\boldsymbol{\beta}') &= \int_c^\infty y\; g_c(y)\; dy \\
&= \left[1 - \Phi\left(\frac{c-\mu(\boldsymbol{x})}{\sigma}\right)\right]^{-1} \int_c^\infty y\; \frac{1}{\sigma}\; \varphi\left(\frac{y-\mu(\boldsymbol{x})}{\sigma}\right)\; dy \\
&= \left[1 - \Phi\left(\frac{c-\mu(\boldsymbol{x})}{\sigma}\right)\right]^{-1} \int_{[c-\mu(\boldsymbol{x})]/\sigma}^\infty [\mu(\boldsymbol{x}) + \sigma z]\varphi(z)\; dz \\
&= \mu(\boldsymbol{x}) - \sigma\left[1 - \Phi\left(\frac{c-\mu(\boldsymbol{x})}{\sigma}\right)\right]^{-1} \int_{[c-\mu(\boldsymbol{x})]/\sigma}^\infty \varphi'(z)\; dz \\
&= \mu(\boldsymbol{x}) + \sigma\left[1 - \Phi\left(\frac{c-\mu(\boldsymbol{x})}{\sigma}\right)\right]^{-1} \varphi\left(\frac{c-\mu(\boldsymbol{x})}{\sigma}\right) \; ,
\end{aligned}$$

which is simple enough to evaluate for given σ and $\mu(\boldsymbol{x})$.

For $g_0(z) = \delta \exp\left[\delta z - \gamma - \exp(\delta z - \gamma)\right]$ one obtains in similar fashion

$$\begin{aligned}
E(Y|Y > c\,;\sigma,\boldsymbol{\beta}') &= \int_c^\infty y\; g_c(y)\; dy \\
&= \left[1 - \mathcal{G}_0\left(\frac{c-\mu(\boldsymbol{x})}{\sigma}\right)\right]^{-1} \int_c^\infty y\; \frac{1}{\sigma}\; g_0\left(\frac{y-\mu(\boldsymbol{x})}{\sigma}\right)\; dy \\
&= \left[1 - \mathcal{G}_0\left(\frac{c-\mu(\boldsymbol{x})}{\sigma}\right)\right]^{-1} \int_{[c-\mu(\boldsymbol{x})]/\sigma}^\infty [\mu(\boldsymbol{x}) + \sigma z] g_0(z)\; dz \\
&= \mu(\boldsymbol{x}) + \sigma\left[1 - \mathcal{G}_0\left(\frac{c-\mu(\boldsymbol{x})}{\sigma}\right)\right]^{-1} \int_{[c-\mu(\boldsymbol{x})]/\sigma}^\infty z g_0(z)\; dz \; .
\end{aligned}$$

Here, substituting and integrating by parts, one has

$$\begin{aligned}
\int_a^\infty zg_0(z)\,dz &= \int_a^\infty [\delta z-\gamma+\gamma]\exp\left[\delta z-\gamma-\exp(\delta z-\gamma)\right]\,dz \\
&= \delta^{-1}\int_{\exp(\delta a-\gamma)}^\infty [\log(t)+\gamma]\,\exp(-t)\,dt \\
&= \delta^{-1}\left(\delta a\exp[-\exp(\delta a-\gamma)]+\int_{\exp(\delta a-\gamma)}^\infty \exp(-t)\,t^{-1}\,dt\right) \\
&= a\exp[-\exp(\delta a-\gamma)]+\delta^{-1}E_1[\exp(\delta a-\gamma)]\,.
\end{aligned}$$

Here $E_1(z)$ is the exponential integral function, see Abramowitz and Stegun (1972). There one also finds various approximation formulas for

$$E_1(z)=\int_z^\infty \exp(-t)\,t^{-1}\,dt\,,$$

namely for $0\le z\le 1$ and coefficients a_i given in Table 2 one has

$$E_1(z)=-\log(z)+a_0+a_1z+a_2z^2+a_3z^3+a_4z^4+a_5z^5+\epsilon(z)$$

with $|\epsilon(z)|<2\times 10^{-7}$, and for $1\le z<\infty$ and coefficients a_i and b_i given in Table 3 one has

$$z\exp(z)\,E_1(z)=\frac{z^4+a_1z^3+a_2z^2+a_3z+a_4}{z^4+b_1z^3+b_2z^2+b_3z+b_4}+\epsilon(z)$$

with $|\epsilon(z)|<2\times 10^{-8}$.

Table 2: Coefficient for $E_1(z)$ Approximation ($0\le z\le 1$)

$a_0=-.57721566$	$a_1=.99999193$	$a_2=-.24991055$
$a_3=.05519968$	$a_4=-.00976004$	$a_5=.00107857$

Table 3: Coefficient for $E_1(z)$ Approximation ($1\le z<\infty$)

$a_1=8.5733287401$	$a_2=18.0590169730$	$a_3=8.6347608925$	$a_4=.2677737343$
$b_1=9.5733223454$	$b_2=25.6329561486$	$b_3=21.0996530827$	$b_4=3.9584969228$

Combining the above one obtains the following formula for $E(Y|Y > c; \sigma, \boldsymbol{\beta}')$:

$$E(Y|Y > c; \sigma, \boldsymbol{\beta}') = \mu(\boldsymbol{x}) + \delta^{-1}\sigma \exp\left[\exp\left(\frac{c-\mu(\boldsymbol{x})}{\sigma/\delta} - \gamma\right)\right]$$

$$\times \left(\frac{c-\mu(\boldsymbol{x})}{\sigma/\delta}\ \exp\left[-\exp\left(\frac{c-\mu(\boldsymbol{x})}{\sigma/\delta} - \gamma\right)\right] + E_1\left[\exp\left(\frac{c-\mu(\boldsymbol{x})}{\sigma/\delta} - \gamma\right)\right]\right)$$

$$= c + \delta^{-1}\sigma \exp\left[\exp\left(\frac{c-\mu(\boldsymbol{x})}{\sigma/\delta} - \gamma\right)\right] E_1\left[\exp\left(\frac{c-\mu(\boldsymbol{x})}{\sigma/\delta} - \gamma\right)\right] .$$

Note that for $\epsilon = \exp(\delta[c-\mu(\boldsymbol{x})]/\sigma - \gamma) \approx 0$ one has

$$\begin{aligned} E(Y|Y > c; \sigma, \boldsymbol{\beta}') &= \mu(\boldsymbol{x}) + \delta^{-1}\sigma \exp(\epsilon)\left[(\gamma + \log(\epsilon))\exp(-\epsilon) + E_1(\epsilon)\right] \\ &= \mu(\boldsymbol{x}) + \delta^{-1}\sigma \exp(\epsilon) \\ &\quad \times \left([\gamma + \log(\epsilon)]\left[1 - \epsilon + O(\epsilon^2)\right] - \log(\epsilon) - \gamma + a_1\epsilon + O(\epsilon^2)\right) \\ &= \mu(\boldsymbol{x}) + \delta^{-1}\sigma \exp(\epsilon)\left[(a_1 - \gamma)\epsilon - \epsilon\log(\epsilon)\right] + O(\epsilon^2\log(\epsilon)) \end{aligned}$$

where a_1 is as in Table 2. In particular, in the limiting case as $\epsilon \to 0$, one has

$$E(Y|Y > c; \sigma, \boldsymbol{\beta}') \longrightarrow \mu(\boldsymbol{x}) .$$

This makes intuitive sense since in that case the censored observation is so low as to provide no information about the actual failure time. In that case it reasonable to replace a "completely missing" observation by its mean value.

For $\lambda = \exp(\delta[c-\mu(\boldsymbol{x})]/\sigma - \gamma)$ very large one has

$$E(Y|Y > c; \sigma, \boldsymbol{\beta}') = c + \delta^{-1}\sigma \exp(\lambda)E_1(\lambda) = c + \delta^{-1}\sigma\left(\frac{1}{\lambda} + O(1/\lambda^2)\right) \approx c .$$

4.2 Some Specific Examples and Simulation Experiences

The data set Table 4 is taken from Gertsbakh (1989) for illustrative and comparative purposes. It gives the log-life times for 40 tested motors under different temperature and load conditions. The failure indicator is one when the motor failed and zero when it was still running at the termination of the test. The maximum likelihood estimates for the regression coefficients and scale parameter were given by Gertsbakh as the entries in the first row of Table 5. The corresponding estimates as computed by our algorithm are given to the same number of digits in the second row of that table. The results are reasonably close to each other.

The data in Table 1 can be taken as another example, although here there are no covariates. This however provides an independent way of gauging the accuracy of our algorithm, since in that case we have an independent double precision algorithm based on root solving. The answers by these two methods are given in Table 6 to the relevant number of digits for comparison. The agreement is very good (at least nine digits) in this particular example.

As another check on the algorithm various simulations were performed, either with noncensored samples and or with various degrees of censoring. In all cases only one covariate was used. For the noncensored case 1000 samples each were generated at sample sizes $n = 5, 20, 50, 100$. The data were generated according to the Gumbel model with a linear model $\beta_1 + \beta_2 u_i$, with $\beta_1 = 1$ and $\beta_2 = 2$. The u_i were randomly generated from a uniform distribution over $(0, 1)$. The scale paramater was $\sigma = .5$. Figures 3 and 4 illustrate the results. The dashed vertical line in the histogram for $\widehat{\sigma}$ is located at $\sigma\sqrt{(n-2)/n}$. It appears to be a better indication of the mean of the $\widehat{\sigma}$. Equivalently one should compare $\widehat{\sigma}\sqrt{n/(n-2)}$ against $\sigma = .5$. The $n - 2$ "accounts" for the two degrees of freedom lost in estimating β_1 and β_2. Judging from these limited simulation results it appears that the factor $\sqrt{n/(n-2)}$ corrects for the small sample bias reasonably well.

Figures 5-7 illustrate the statistical properties of the maximum likelihood estimates for medium and heavily censored samples of size $n = 50, 500$ and 1000. The censoring was done as follows. For each lifetime Y_i in the sample a random censoring time $V_i = .5 + 3\gamma W_i$ was generated, with W_i taken from a uniform $(0, 1)$ distribution. The smaller of Y_i and V_i was then taken as the i^{th} observation and the censoring indicator was set appropriately. The parameter γ controls the censoring. A small value of γ means heavy censoring and larger γ means medium to light censoring. In

this simulation $\gamma = .2$ and $\gamma = 1$ were used.

The presentations in Figures 5-7 plot for each sample the estimate versus the corresponding censoring fraction. Originally $N = 1000$ samples were generated under each censoring scenario, but under $n = 50$ and heavy censoring two samples did not permit a solution, since at least 49 lifetimes were censored in those cases. The percentages given in these plots indicate the proportion of estimates above the respective target line. The percentages given in parentheses use the dashed target line, which as in Figures 3-4 is an attempt at bias correction. Note how the increasing sample size entails a reduction in the scatter of the estimates. Also note how the scatter increases with increasing censoring fraction.

Also shown in each plot of Figures 5-7 is the least squares regression line to indicate trends in the estimates against the censoring fraction. It appears that for heavy censoring there is a definite trend for the intercept estimates $\widehat{\beta}_1$. Namely, as the censoring fraction increases so does the intercept estimate. We do not know whether this effect has been discussed in the literature. The usefulness of this relationship is questionable, since one usually does not know whether the regression line is above or below the target line, since the latter is unknown. Note that the median of the estimates $\widehat{\beta}_1$ is close to target.

4.3 The Fortran Code GMLE

The file with the Fortran subroutine GMLE, developed out of the above considerations, is called gmle.f and is documented in Appendix A. Although the source code for it could easily be made available, it still requires linking with three BCSLIB subroutine libraries, namely **optlib**, **bcsext**, and **bcslib**. Once one has written an appropriate driver for GMLE (which may be contained in the file gmledrv.f, also available) one needs to compile these as follows on a Sun workstation

```
f77 gmledrv.f gmle.f -loptlib -lbcsext -lbcslib .
```

Table 4: Motor Failure Data, Two Factors (from Gertsbakh, p. 206)

log failure time	rescaled load index	rescaled temper.	failure indicator	log failure time	rescaled load index	rescaled temper.	failure indicator
5.45	1	1	1	5.15	-1	1	1
5.74	1	1	1	6.11	-1	1	1
5.80	1	1	1	6.11	-1	1	1
6.37	1	1	1	6.23	-1	1	1
6.49	1	1	1	6.28	-1	1	1
6.91	1	1	1	6.32	-1	1	1
7.02	1	1	1	6.41	-1	1	1
7.10	1	1	0	6.56	-1	1	1
7.10	1	1	0	6.61	-1	1	1
7.10	1	1	0	6.90	-1	1	0
5.07	1	-1	1	3.53	-1	-1	1
5.19	1	-1	1	4.22	-1	-1	1
5.22	1	-1	1	4.73	-1	-1	1
5.58	1	-1	1	5.22	-1	-1	1
5.83	1	-1	1	5.46	-1	-1	1
6.09	1	-1	1	5.58	-1	-1	1
6.25	1	-1	1	5.61	-1	-1	1
6.30	1	-1	0	5.97	-1	-1	1
6.30	1	-1	0	6.02	-1	-1	1
6.30	1	-1	0	6.10	-1	-1	0

Table 5: Comparison of MLE's for Data in Table 4

Source	intercept	load coefficient	temperature coefficient	scale
Gertsbakh	6.318	0.253	0.391	0.539
our code	6.317	0.253	0.391	0.538

Table 6: Comparison of MLE's for Data in Table 1

Source	scale parameter	shape parameter
root solver	952.3774020	23.90139575
optimization code	952.3774021	23.90139576

Figure 3: 1000 Simulations at $n = 5$ and $n = 20$ (uncensored)

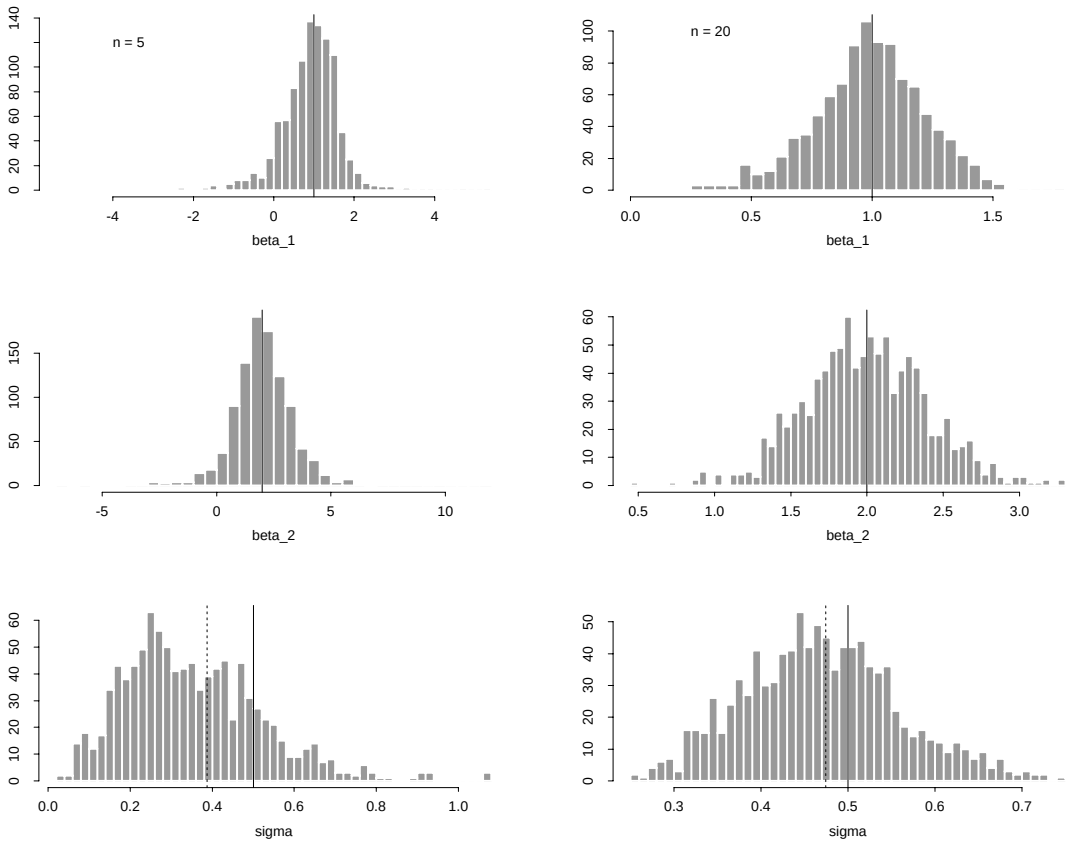


Figure 4: 1000 Simulations at $n = 50$ and $n = 100$ (uncensored)

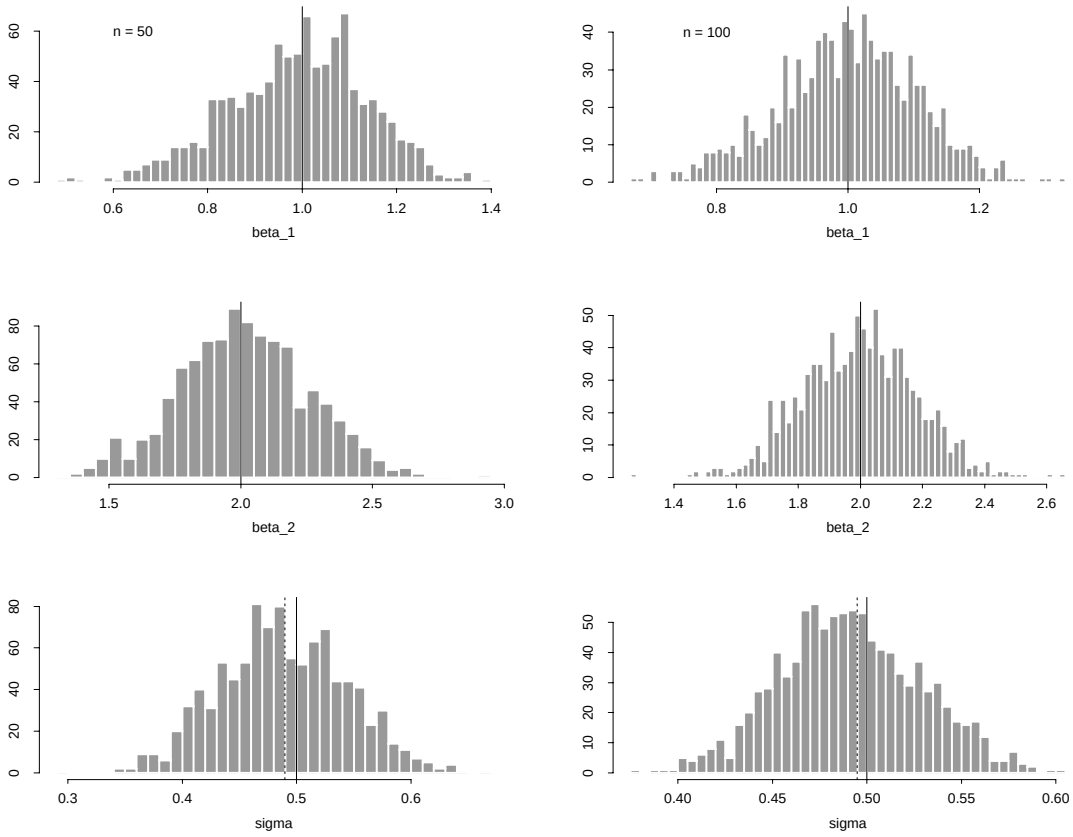


Figure 5: 1000 Simulations at $n = 50$, medium and heavy censoring

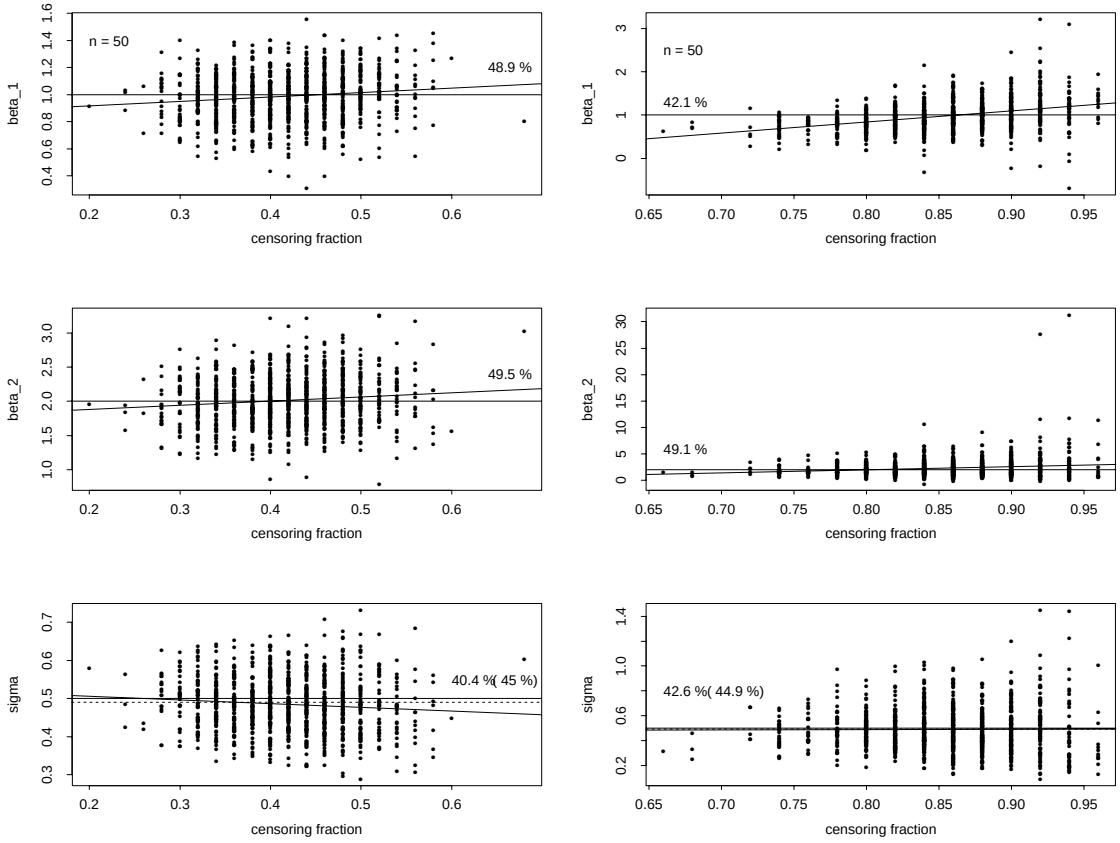


Figure 6: 1000 Simulations at $n = 500$, medium and heavy censoring

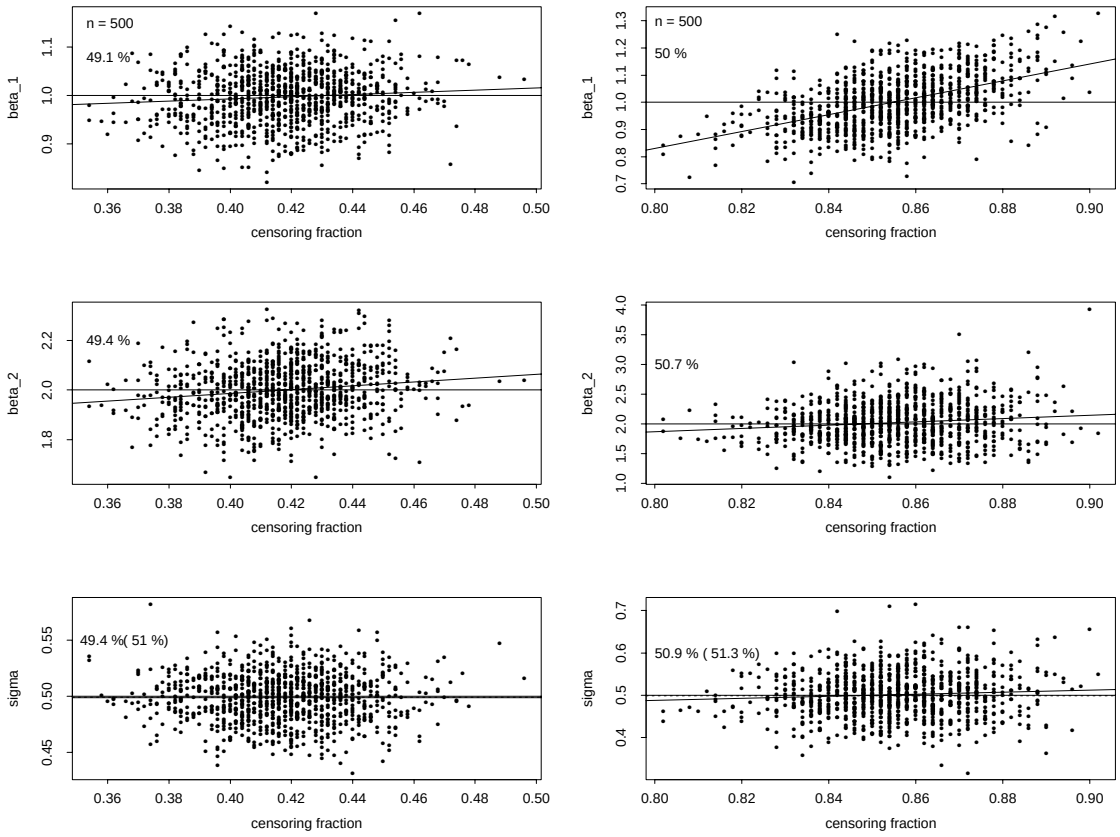
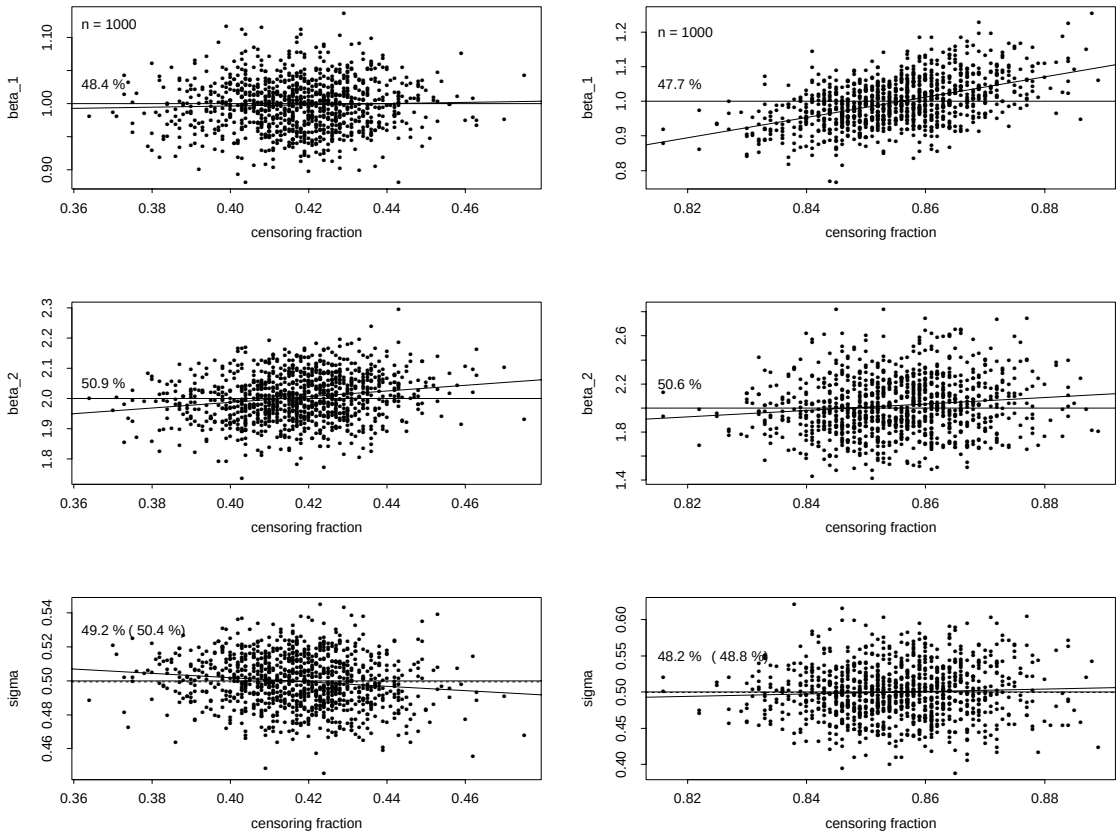


Figure 7: 1000 Simulations at $n = 1000$, medium and heavy censoring



References

Abramowitz, M. and Stegun, I.A. (1972). *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables.* Dover Publication, Inc., New York.

Arnold, V.I. (1978). *Ordinary Differential Equations*, The MIT Press, Cambridge Massachusetts.

Barndorff-Nielsen, O. and Blæsild, P. (1980). "Global maxima and likelihood in linear models." Institute of Mathematics, University of Aarhus. Research Report No. 57.

Copas, J.B. (1975). "On the unimodality of the likelihood for the Cauchy distribution." *Biometrika* **62**, 701-704.

Gabrielsen, G. (1982). "On the unimodality of the likelihood for the Cauchy distribution: Some comments." *Biometrika* **69**, 677-678.

Gabrielsen, G. (1986). "Global maxima of real-valued functions." *Journal of Optimization Theory and Applications* **50**, 257-266.

Gertsbakh, I.B. (1989). *Statistical Reliability Theory*, Marcel Dekker, New York.

Hartman, P. (1964). *Ordinary Differential Equations*, John Wiley & Sons, Inc., New York.

Kendall, M.G. and Stuart, A. (1973). *Advanced Theory of Statistics, Vol. 2, 3rd Edition*, Haffner Publishing Company, New York.

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

Lehmann, E.L. (1980). "Efficient likelihood estimators." *The American Statistician* **34**, 233-235.

Mäkeläinen, T., Schmidt, K., and Styan, G.P.H. (1981). "On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples." *The Annals of Statistics* **9**, 758-767.

Milnor, J. (1963). *Morse Theory*, Princeton University Press.

Nelson, W. (1982). *Applied Life Data Analysis*, John Wiley & Sons, New York.

Rai, K. and van Ryzin, J. (1982). "A note on a multivariate version of Rolle's theorem and uniqueness of maximum likelihood roots." *Commun. Statist.-Theory and Methods* **11**, 1505-1510.

Rudin, W. (1976). *Principles of Mathematical Analysis*, McGraw Hill, New York.

Schmee, J. and Hahn, G.J. (1979). "A simple method for regression analysis with censored data," (with discussion). *Technometrics* **21**, 417-434.

Scholz, F.W. (1981). "On the uniqueness of roots of the likelihood equations." Technical Report No. 14, Department of Statistics, University of Washington, Seattle Washington 98195.

Tarone, R.E. and Gruenhage, G. (1975). "A note on the uniqueness of roots of the likelihood equations for vector-valued parameters." *Journal of the American Statistical Association* **70**, 903-904.

Turnbull, B.W. (1974). "Nonparametric estimation of a survivorship function with censored data." *Journal of the American Statistical Association* **69**, 169-173.

Appendix A

GMLE: Maximum Likelihood Estimates from Censored Gumbel/Weibull Data with Covariates

VERSION

`GMLE` — `Double Precision`

PURPOSE

GMLE computes maximum likelihood estimates of regression and scale parameters for type I or multiply censored data when the data are assumed to come from a Gumbel distribution with location parameter being a linear function of known covariates. By setting the IDIST switch to 2, GMLE will analyze the log-transformed data. This allows to view the original data to come from a Weibull distribution with the log-scale parameter modeled by a linear function of known covariates. The Weibull shape parameter becomes reciprocal of the Gumbel scale parameter.

USAGE

```
INTEGER IDIST, N, NMAX, NP, JFAIL(NMAX), IER
DOUBLE PRECISION RESP(NMAX), COV(NMAX,NP), COEF(NP), SIGMA
CALL GMLE(IDIST,N,NMAX,NP,RESP,COV,JFAIL,COEF,SIGMA,IER)
```

ARGUMENTS

IDIST [INPUT,INTEGER]
Choice of the data distribution model
log(Weibull) = Gumbel or Weibull
IDIST = 1 for the Gumbel model and
IDIST = 2 for the Weibull model.

N [INPUT,INTEGER]
Sample size, NP $<$ N $\leq$ NMAX

NMAX [INPUT,INTEGER]
Maximum sample size
currently NMAX cannot exceed 10000

NP [INPUT,INTEGER]
Number of covariates per observation
Currently NP $\leq$ 19
Also NP $\geq$ 1, to accommodate
at least a location paramter in the Gumbel model
or a scale parameter in the Weibull model.

RESP [INPUT,DOUBLE PRECISION,ARRAY]
Sample vector of responses, lifetimes or observations,
Need RESP(I) $>$ 0 for all I = 1, ..., N
if IDIST = 2 is specified.

COV [INPUT,DOUBLE PRECISION,ARRAY]
N$\times$ NP array of covariates
corresponding to the observation vector,

JFAIL [INPUT,INTEGER,ARRAY]
Vector of failure indicators with
JFAIL(I) = 1, if the I^{th} observation is a failure
JFAIL(I) = 0, if the I^{th} observation is a censored case,

COEF [OUTPUT,DOUBLE PRECISION,ARRAY]
Vector of maximum likelihood estimates for the
regression coefficients,

SIGMA [OUTPUT,DOUBLE PRECISION]
Maximum likelihood estimate for the Gumbel scale parameter
or the reciprocal of the Weibull shape parameter,
depending on IDIST = 1 or IDIST = 2.

IER [OUTPUT,INTEGER]
Success/error code

IER = 0	Success, maximum likelihood estimates computed.
IER =-1	N > 10,000 cases or NP > 19.
IER =-2	Maximum likelihood solution criterion not satisfied.
IER =-3	IDIST not 1 or 2 (Gumbel or Weibull distribution).
IER =-4	Not all response data are positive, while IDIST = 2.
IER =-5	Fewer uncensored data cases than NP.
IER =-6	Trouble evaluating log-likelihood at initial estimates.
IER > 0	Unexpected error returns from HDLSLE or HDNLPR or their subsidiaries. Here
	IER = 2000+JER, with JER = error return code from HDLSLE
	JER = 4000+JER, with JER = error return code from HDNLPR
	JER = 6000+JER, with JER = error return code from internal routine SHCENS

To raise the bound 10,000 on NMAX and the bound 19 on NP, one should adjust the first two PARAMETER statements in GMLE, namely

```
PARAMETER ( NN=10000 )
PARAMETER ( MAXDIM=20, MCON=0, MAXCON=1) .
```

In the second PARAMETER statement one changes `MAXDIM=20` (with `20 = 19+1`) to `MAXDIM = ICOV+1`, wher ICOV is the new maximum number of covariates. However, there may be complications with auxiliary arrays in the optimizer HDNLPR, which may require larger dimensions. In that case the optimizer will return an error message, stating how much space is needed. This may necessitate an appropriate change in the fifth parameter statement of GMLE, namely in

```
PARAMETER (NHOLD=7500,NIHOLD=500).
```

We refer to the documentation for HDNLPR for the interpretation of that error message and what changes may be indicated.