100(1-P)Th Percentile Remaining Survival Time

K. P. Singh[a], **S. Bae**[a] **and A. A. Bartolucci**[b]

[a] *Department of Epidemiology and Biostatistics, School of Public Health, University of North Texas Health Science Center, ME I, 3500 Camp Bowie Blvd., Ft. Worth, TX 76107-2699, USA, (ksingh@hsc.unt.edu, sbae@hsc.unt.edu)*

[b] *Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294-0008, USA, (albartol@uab.edu)*

Abstract: Given that a patient survives beyond time t, then there is a probability of (1-p) that the patient will survive beyond time $m_p(t)$. $m_p(t)$ is thus the time beyond which 100(1-p)% of all patients will survive given that each of them survive beyond time t. $m_p(t)$ is a very useful and interpretable index. Being a function of time elapsed it highlights different properties of survivorship frequently masked under $S(t,\underline{\theta})$ which may appear similar under a number of representatives. In this paper we derive maximum likelihood estimates $m_p(t)$ of using the log-logistic and three-parameter generalized log-logistic regression models for censored survival data. We also develop (1-α)100% estimates of confidence intervals for $m_p(t)$. We illustrate our results using a numerical example.

Keywords: Generalized Log-logistic Regression; Censored Data; Asymptotic Confidence Interval

1. INTRODUCTION

Let T be the lifetime of a patient and $0 < p < 1$.

$$P[T > m_p(t) \mid T > t] = 1 - p$$

Then define $m_p(t)$, the $100(1-p)th$ percentile remaining time, by:

$$S(m_p(t)) = (1-p)\,S(t)$$

Thus we may estimate $m_p(t)$ as:

$$\hat{S}(m_p(t)) = (1-p)\,\hat{S}(t)$$

$$\hat{m}_p(t) = F^{-1}[1-(1-p)\,\hat{S}(t)]$$

The following possible applications give pretty good motivation for $m_p(t)$:

a. Suppose it is known that a particular patient has a side effect, which is fatal for an unknown population of patients. It is also known that the effect of the treatment wears out with time. That is, given that a patient has survived by time t, the patient has a higher probability of survival beyond time $t+t'$. We would like to inform patients about the time they would survive beyond with probability of, say,.95, given that they survive beyond, say, two years from time by treatment, $m_{.95}(t)$ is precisely this time.

b. We might want to do the above for varying p.

c. $m_p(t)$ for varying p is a measure of future survivability of patient.

d. Radiation treatment may cause unknown cancer or disease.

e. Compare $m_p(t)$ when no side effect against $m_p(t)$ when there exists side effect.

f. Could be used to see if dose is too much or compare two treatments.

h(t; $\underline{\beta}$), the hazards function is an important function in estimating the most prognostic index, S(t; β). Cox [1972] defined hazards function as follows:

$h(t; \beta) = h_0(t) \exp(\beta X)$

In view of the high efficiency of the $\underline{\beta}$ estimates in the neighborhood of $\underline{\beta}=0$ [Oaks, 1977; Kalbfeisch, 1974; Efron, 1988] the approach is useful preferred in significant testing .

2. A GENERALIZED LOG-LOGISTIC MODEL FOR CENSORED SURVIVAL DATA

Let T be the survival time for an individual and let log (t) be a generalized logistic random variable with shape parameters ν and η. Then from Singh [1989] T is the generalized log-logistic random variable with shape parameters υ andη. Assume that the patients are grouped into N samples. For the kth sample, let T be the survival time for a patient and X_k be a q-dimensional vector of observed covariates from an individual

$$\underline{\beta'}\,\underline{X}_k = \beta_0 + B_1 X_{k1} + ... + \beta_q X_{kq}$$

then the *p.d.f.* of T is given by:

$$g_k(t) = \delta\,[F_k(t)]^{\nu}\,[1 - F_{k(t)}]^{\eta} / [tB(\nu,\eta)] \qquad (1)$$

where $B(\nu,\eta)$ is the complete beta function with parameters $\nu > 0$ *and* $\eta > 0$, t is the observed survival time for a patient in the kth sample, and

$$F_K(t) = \{1 + e^{-[\delta \log(t) + \underline{\beta'}\,\underline{X}_k]}\}^{-1} \qquad (2)$$

is the log-logistic *c.d.f.*, with δ and β's *(q + 2)* unknown parameters. For notational simplicity let the generalized log-logistic model be denoted by $GLL(\nu,\eta)$. Note that if $\nu = \eta = 1$, $GLL(\nu,\eta)$ reduces to the log-logistic model. It is symmetric around

$$\log(t) = -\underline{\beta'}\,\underline{X} / \delta,$$

positive skewed if $\nu > \eta$ and negative skewed if $\nu < \eta$. The hazard function of T is given by:

$$h_k(t) = g_k(t) / S_k(t)$$

and $S_k(t) = 1 - G_k(t)$

Singh [1989] showed that the family of the generalized logistic models provides many of the different shaped hazard rates, for example, strictly increasing *(I)*, constant *(C)*, strictly decreasing *(D)*, bathtub shaped and upside down bathtub shaped.

Let n_{1k}, n_{2k}, *and* n_{3k} be respectively the number of uncensored, left censored and right censored observations in the kth sample, and let n be the number of uncensored observations in all N samples. Then:

$$n = \sum_{k=1}^{N} n_{1k}$$

The likelihood of the k_{th} sample may then be expressed as follows:

$$L_k = \prod_{i=1}^{n_{1k}} g_i \prod_{j=1}^{n_{2k}} G_K(t_j) \prod_{v=1}^{n_{3k}} S_k(t_v) \qquad (3)$$

The likelihood L of all N samples is simply the product of the L_k over all samples.

Denote the first derivative with respect to θ_1 by $D(\theta_1)$ and the second derivative with respect to θ_1 and θ_2 by $D^2(\theta_1,\theta_2)$. The *MLE* of the parameters δ, β_r *(r = 0, 1, 2,...,q)*,ν *and* η are obtained by solving $D(\underline{\theta}) = \underline{0}$. Note that these equations are nonlinear in $\underline{\theta}$ $(\delta, \underline{\beta}, \nu, \eta)$ and numerical iterative procedures such as the Newton-Raphson method should be used to fine the *MLE* of the parameters. In addition, computational difficulty occurs in evaluating $D(\nu)$ and $D(\eta)$ because of the flatness of the log-likelihood l over ν and η. An alternative is to consider the submodels $GLL(\nu,1)$ and $GLL(1,\eta)$. Note that the shape parameter still retains the property of measuring the structure of heavy tail. For the model $GLL(\nu,1)$,$\nu < 1$ reflects the heavy tail. For $GLL(1,\eta)$ also reflects such a tail.

3. MAXIMUM LIKELIHOOD ESTIMATE OF $m_p(t)$ FROM THE GENERALIZED LOG-LOGISTIC MODEL MODEL, GLL(V,1)

Consider the generalized log-logistic model describe in section 2 for v and η = 1.

$$G_k(t) = [\, F_k(t)\,]^{\nu} \qquad (4)$$

$$g_k(t) = \frac{\delta\nu}{t}\left[F_k(t)\right]^{m} [1 - F_k(t)] \qquad (5)$$

The log-likelihood is given by:

$$l(\underline{\theta}\,|\,t, \underline{x}\,) = n \log(\delta\nu\,) + \sum_{k=1}^{N} \qquad (6)$$

$$\sum_{i=1}^{n_{1k}} (\nu \log F_k(t_i) + \log\{1 - F_k(t_i)\}) +$$

$$\sum_{j=1}^{n_{2K}} \nu \log F_k(t_j) + \sum_{v=1}^{n_{3k}} \log\{1 - [\, F_k(t_r)\,]^{\nu}\}$$

where $\underline{\theta} = (\delta, \underline{\beta}, \nu)$.

The *MLE* of the parameters $\delta, \beta_r, and \nu$ are obtained by solving the equations: $D(\delta) = 0, D(\beta_r) = 0,$ and $D(\nu) = 0.$ If $\hat{\delta}, \hat{\beta}_r\ (r = 0,1,\ldots,q),$ and $\hat{\upsilon}$ are the maximum likelihood estimates, then the survival function for the *GLL*(v,1) model can be estimated as:

$$\hat{S}_k(t) = 1 - [\, \hat{F}_k(t)\,]^{\hat{\nu}} \qquad (7)$$

and the hazard function is estimated as:

$$\hat{h}_k(t) = (\hat{\delta}\,\hat{\nu}\,/\,t)\,\frac{[\, \hat{F}_k(t)\,]^{\hat{\nu}}\,[1 - \hat{F}_k(t)]}{\hat{S}_k(t)} \qquad (8)$$

and MLE of $m_p(t)$ the *100(1-p)th* percentile remaining time, for an individual is given by:

$$\hat{m}_p(t) = \exp(-\,\underline{\hat{\beta}}'\,\underline{X}_k\,/\,\hat{\delta}\,)\,[w/(1 - w)\,]^{1/\delta} \qquad (9)$$

where:

$$w = [\,1 - (1 - p)\,\hat{S}_k(t))\,]^{1/\hat{\nu}}$$

Now we derive the $100(1-\alpha)\%$ confidence interval for $m_p(t)$ of the GLL (v,1).

The method of interval estimation we are using involves the asymptotic distribution of the *MLE* of $\underline{\theta}$. If $\underline{\theta}$ is interior to the parameter space, and since $l(\underline{\theta})$ is twice differentiable, it can be shown that the asymptotic distribution of $\underline{\hat{\theta}}$ is multivariate normal with mean $\underline{\theta}$ and the variance-covariance matrix which is the inverse of Fisher information matrix of second derivations evaluated at $\theta = \hat{\theta}$. Considering the *GLL(v,1)* model and using the derivatives (1-8) given in Appendix A, the asymptotic variance of $\hat{m}_p(t)$ approximately is given by:

$$Var(\hat{m}_p(t)) \cong (\hat{m}_p(t))^2\,[\,A^2\,var(\hat{\delta}\,) + \Sigma B_i^2$$

$$var(\hat{\beta}_i) + C^2\,var(\hat{\nu}\,) + A\,\Sigma B_i\,Cov(\hat{\delta}, \hat{\beta}_i)$$

$$+ AC\,Cov(\hat{\delta}, \hat{\nu}\,) + C\,\Sigma B_i\,Cov(\hat{\beta}_i, \hat{\nu}\,)$$

$$+ \Sigma\Sigma B_i B_j\,Cov(\hat{\beta}_i, \hat{\beta}_j)] \qquad (10)$$

where:

$$A = \frac{\partial}{\partial\hat{\delta}}\log(\hat{m}_p(t))$$

$$B_i = \frac{\partial}{\partial\hat{\beta}_i}\log(\hat{m}_p(t))$$

$$C = \frac{\partial}{\partial\hat{\nu}}\log(\hat{m}_p(t))$$

Hence the 100(1-α)% asymptotic confidence interval for the remaining survival time, $m_p(t)$, is defined as:

$$\hat{m}_p(t) \pm Z_{\alpha/2}\sqrt{Var(\hat{m}_p(t)} \qquad (11)$$

4. MAXIMUM LIKELIHOOD ESTIMATE OF $m_p(t)$ FROM THE GENERALIZED LOG-LOGISTIC MODEL, GLL(1,η)

Consider the generalized log-logistic model for v = 1 and η. The *cdf* and *pdf* of *T* respectively reduce to

$$G_k(t) = 1 - [1 - F_k(t)\,]^{\eta} \qquad (12)$$

$$g_k(t) = \frac{\delta\eta}{t}\,[F_k(t)]\,[1 - F_k(t)]^{\eta} \qquad (13)$$

The log likelihood function is given by:

$$l(\underline{\theta}) = n \log(\delta\eta) + \sum_{k=1}^{N}[$$

$$\sum_{i=1}^{n_{1k}} [\log F_k(t_i) + \eta \log(1 - F_k(t_i)) - \log(t_i)]$$

$$+ \sum_{j=1}^{n_{2k}} \log[1 - (1 - F_k(t_j))^{\eta}] \qquad (14)$$

$$+ \sum_{v=1}^{n_{3k}} \eta \log(1 - F_k(t_v))$$

where $\underline{\theta} = (\delta, \underline{\beta}, \eta)$. The first and second derivatives with respect to $\underline{\theta}$ can easily be obtained. The *MLE* of $\underline{\theta}$ is obtained by solving the equations:

$D(\delta) = 0, D(\mathrm{B}_r) = 0$ *and* $D(\eta) = 0$. Then the survival function, hazard function and the remaining survival time are respectively given by $\hat{S}_k(t) = [1 - \hat{F}_k(t)]^{\eta}$

$$\hat{m}_p(t) = \exp(-\underline{\hat{\beta}}_i \underline{X}_k / \hat{\delta}) \left[\frac{1 - w}{w}\right]^{1/\hat{\delta}} \qquad (15)$$

where

$$W^{\eta} = (1 - p)\,\hat{S}_k(t)$$

Now, we derive the $100(1-\alpha)\%$ asymptotic Confidence Interval for $m_p(t)$ using asymptotic distribution theory. The derivations are not difficulty to derive. Note that $\underline{\hat{\theta}} = (\hat{\delta}, \hat{\beta}, \hat{\eta})$ is the maximum likelihood estimate of $\underline{\theta}$. The asymptotic distribution of $\underline{\hat{\theta}} = (\hat{\delta}, \hat{\beta}, \hat{\eta})$ is approximately multivariate normal with mean $\underline{\theta}$ and the variance-covariance matrix, the inverse of Fishers information matrix of second derivatives evaluated at $\underline{\theta} = \underline{\hat{\theta}}$.

Considering GLL(1,η) model and using (1-8) given Appendix B, the asymptotic variance of $\hat{m}_p(t)$ approximately is as follows:

$$Var(\hat{m}_p(t)) \cong (\hat{m}_p(t))^2 [A^2\, var(\hat{\delta}) + \sum B_i^2\, var(\hat{\beta}_i) + C^2\, var(\hat{\eta}) + A \sum B_i\, Cov(\hat{\delta}, \hat{\beta}_i) + AC\, Cov(\hat{\delta}, \hat{\eta}) + C \sum B_i\, Cov(\hat{\beta}_i, \hat{\eta}) + \sum\sum B_i B_j\, Cov(\hat{\beta}_i, \hat{\beta}_j)] \qquad (16)$$

where

$$A = \frac{\partial}{\partial\hat{\delta}} \log(\hat{m}_p(t)), B_i = \frac{\partial}{\partial\hat{\beta}_i} \log(\hat{m}_p(t))$$

$$C = \frac{\partial}{\partial\hat{\eta}} \log(\hat{m}_p(t))$$

Hence the $100(1-\alpha)\%$ asymptotic confidence interval for $m_p(t)$ is given by

$$\hat{m}_p(t) \pm Z_{\alpha/2} \sqrt{Var(\hat{m}_p(t))} \qquad (17)$$

5. APPLICATION

Vogler et. al. [1992] undertook a randomized clinical trial to compare the therapeutic effectiveness of idarubicin (IDR) to daunorubicin (DNR). Both groups were given in combination with cytarabine (CA) in acute myelogenous leukemic (AML) patients. There were 105 patients on the IDR arm and 113 patients on the DNR arm. The randomization plan was generated prospectively and was restricted to incorporate stratification parameters for age (15 to 50, 51 to 60, and >60 years), and a history of an antecedent hematologic disorder (myelodysplastic syndrome with transformation to AML).

The groups were reasonably balanced, with no significant differences with regard to age, sex, FAB classification, antecedent hematologic disorder, performance status, presence of bleeding or infection at diagnosis, or median WBC count or hemoglobin concentration. The median platelet count was The likelihood of an earlier death increased with increasing age in both group (P<0.0004). The WBC counts at diagnosis were compared with regard to treatment arms. There were no significant differences between the treatment arms in patients with respect to WBC [Vogler et. al., 1992].

Both groups were combined. 183 patients were

available for 100(1-p)th percentile remaining survival time. Of the 183 patients, 13 patients were censored. Log (Sur), Ons-Plat, Age and treatment were considered for GLL(v,η) where we fixed v=1 and η =1. The estimates for the regression coefficients for Log (Sur), Ons-Plat, Age and treatment are, respectively, 1.9447, 0.0079, 0.0382 and 0.9632.

The asymptotic estimate of variance-covariance matrix for the regression coefficient estimates is as follows:

	Log (Sur)	Ons-Plat	Age	Treat
Log (Sur)	.154657	.000303	.001805	.035484
Ons-Plat		.000023	-.000006	.000553
Age			.000512	.001604
Treat				.431110

95% asymptotic confidence intervals for 100 (1-p)th percentile remaining survival times for patients in Treatments A and B beyond 12 years are given by Figures 1 and 2.

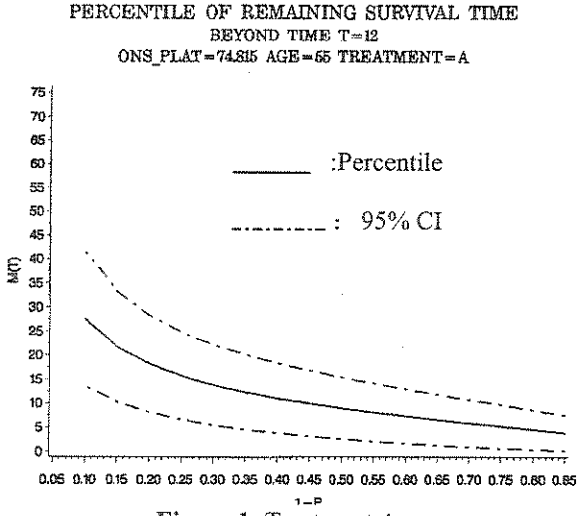

Figure 1: Treatment A.

From Figures 1 and 2, Treatment B gives narrower 95% confidence interval than Treatment A. Using the estimated regression coefficients and the asymptotic estimated variance-covariance matrix the *100(1-α)*% asymptotic confidence interval for $m_p(t)$ could easily be estimated. The authors are presently estimating v and η for the general case GLL(V,η).

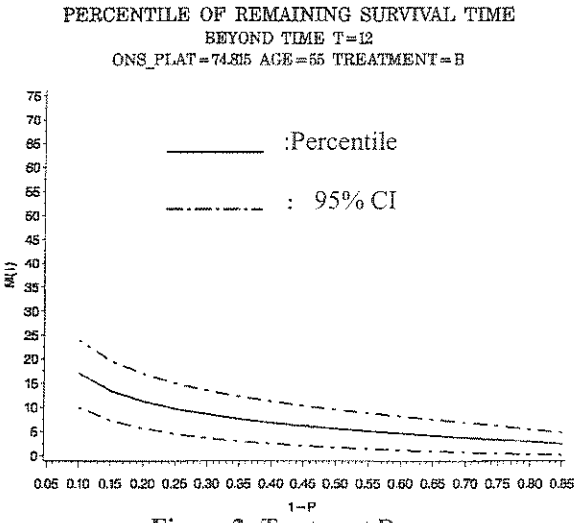

Figure 2: Treatment B.

6. REFERENCES

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

Efron, B., Logistic regression, survival analysis, and the Kaplan-Meier curve, *Journal of the American Statistical Association*, 83, 414-425, 1988.

Kalbfeisch, J. D., Some efficiency calculations for survival distributions, *Biometrika,* 61, 31-38, 1974.

Oaks, D., The asymptotic information in censored survival data, *Biometrika*, 64, 441-448, 1977.

Singh, K. P., A generalized log-logistic regression model for survival analysis: Hazard ratecharacteristics, Biom. Praxim., 29, 63-74, 1989.

Vogler, W. R., E. Velez-Garcia, R. S. Weiner, M. A. Flaum, A. A. Bartolucci, G. A. Omura, M. C. Gerber, and P. L. C. Banks, A phase III comparing idarubicin and daunorubicin in combination with cytarabine in acute myelogenous leukemia: A Southern Cancer Study Group study, *Journal of Clinical Oncology*, 10, 1103-1111, 1992.

7. APPENDIX A

The following partial derivatives are used in the derivation of asymptotic variance of $\hat{m}_p(t)$ using $GLL(\nu, 1)$:

Note that

$$W^{\hat{\nu}} = 1 - (1 - p)\,\hat{S}_k(t) \tag{1}$$

$$\frac{\partial}{\partial \hat{\delta}} W = \frac{(1 - p)[1 - \hat{S}_k(t)]\,[1 - \hat{F}_k(t)] \log(t)}{W^{\hat{\nu}-1}} \tag{2}$$

$$\frac{\partial}{\partial \hat{\beta}_i} W = X_{ki} \frac{\partial}{\partial \hat{\delta}} W \tag{3}$$

$$\frac{\partial}{\partial \hat{\nu}} W = \frac{(1 - p)\,[1 - \hat{S}_k(t)]}{\hat{\nu}\, W^{\hat{\nu}-1}} \quad \frac{\log(\hat{F}_k(t))}{\hat{\nu}\, W^{\hat{\nu}-1}} - \frac{W \log(W)}{\hat{\nu}} \tag{4}$$

Note that

$$\log(\hat{m}_p(t)) = \left[-\underline{\beta}' \underline{X}_k + \log\left(\frac{W}{1 - W}\right) \right] / \hat{\delta} \tag{5}$$

Thus

$$\frac{\partial}{\partial \hat{\delta}} \log(\hat{m}_p(t)) = \left[-\log(\hat{m}_p(t)) + \frac{1}{W(1 - W)} \frac{\partial}{\partial \hat{\delta}} W \right] / \hat{\delta} \tag{6}$$

$$\frac{\partial}{\partial \hat{\beta}_i} \log(\hat{m}_p(t)) = \left[-X_{ki} + \frac{1}{W(1 - W)} \frac{\partial}{\partial \hat{\beta}_i} W \right] / \hat{\delta} \tag{7}$$

$$\frac{\partial}{\partial \hat{\nu}} \log(\hat{m}_p(t)) = \frac{1}{\hat{\delta} W(1 - W)} \frac{\partial}{\partial \hat{\nu}} W \tag{8}$$

8. APPENDIX B

The following partial derivatives are used in the derivation of the asymptotic variance of $\hat{m}_p(t)$ using $GLL(1, \hat{\eta})$:

Note that

$$W^{\hat{\eta}} = (1 - p)\,\hat{S}_k(t) \tag{1}$$

$$\frac{\partial}{\partial \hat{\delta}} W = W\,\hat{F}_k(t) \log(t) \tag{2}$$

$$\frac{\partial}{\partial \hat{\beta}_i} W = W\,\hat{F}_k(t)\, X_{ki} \tag{3}$$

$$\frac{\partial}{\partial \hat{\eta}} W = \frac{w}{\hat{\eta}} \log\left[\frac{W}{1 - \hat{F}_k(t)}\right] \tag{4}$$

Note that

$$\log(\hat{m}_p(t)) = \left[\underline{\hat{\beta}}_i \underline{X}_k + \log\frac{(1 - W)}{W} \right] / \hat{\delta} \tag{5}$$

Thus,

$$\frac{\partial}{\partial \hat{\delta}} \log(\hat{m}_p(t)) = -\left[\log(\hat{m}_p(t)) + \frac{1}{W(1 - W)} \frac{\partial}{\partial \hat{\delta}} W \right] / \hat{\delta} \tag{6}$$

$$\frac{\partial}{\partial \hat{\beta}_i} \log(\hat{m}_p(t)) = -\left[X_{ki} + \frac{1}{W(1 - W)} \frac{\partial}{\partial \hat{\delta}} W \right] / \hat{\delta} \tag{7}$$

$$\frac{\partial}{\partial \hat{\eta}} \log(\hat{m}_p(t) = -\frac{1}{\hat{\delta} W(1 - W)} \frac{\partial}{\partial \hat{\eta}} W \tag{8}$$