SIMULTANEOUS CONFIDENCE BANDS FOR THE PTH PERCENTILE AND THE MEAN LIFETIME IN EXPONENTIAL AND WEIBULL REGRESSION MODELS

Ping Sa1 and S.J. Lee2

1Dept. of Mathematics and Statistics, U. of North Florida, Jacksonville, FL.32224
2Dept. of Mathematics and Statistics, U. of West Florida, Pensacola, FL.32514

ABSTRACT
Simultaneous confidence bands are developed for the pth percentile and the expected lifetime of the Weibull regression model, using the confidence region about the parameters and Lagrange multiplier procedure. An example of real data is analyzed to illustrate the procedure.

Key Words and Phrases: simultaneous confidence band; confidence region; maximum likelihood estimator; Lagrange multiplier procedure.

1. INTRODUCTION
The most commonly used Weibull regression model has the survival function for lifetime $T$, given the vector of regressor variables $x$, of the form

$$S(t, x) = \exp\left(-\left(\frac{t}{\theta_x}\right)^\nu\right), \quad t \geq 0$$

(1)

where $\nu > 0$ is a shape parameter which does not depend upon $x$, and $\theta_x$ is a function of $x$ that involves unknown parameters $\beta$. The most frequently used $\theta_x$ is $\theta_x = \exp(x'\beta)$, where $x'$ is the transpose of the vector $x$. Note that the exponential regression model is a special case of (1) as $\nu = 1$.

In reliability practice, one is often interested in obtaining simultaneous confidence intervals for mean or median response under various values of regressor variables when the model is Weibull regression with unknown parameters. In this paper, we construct confidence bands for the pth percentile and the mean life of the Weibull regression with $k$ regressor variables involved. That is, $\theta_x = \exp(x'\beta)$, where $x' = (1, x_1, ..., x_k)$, $x_i \in R$, $i = 1, 2, ..., k$ and $\beta = (\beta_0, \beta_1, ..., \beta_k)$. Scheffe' (1959, Appendix III) originated the idea
to project the ellipsoidal confidence region of $\beta$ onto a linear combinations of $\beta$'s in linear regression. Similar to Scheffe's idea, the construction of the confidence bands in this paper is based on an existing confidence region about the parameters along with Lagrange multiplier procedure which give a concise derivation and more general results than those using maximization and minimization procedure by Brand, Pinnock and Jackson (1973) and Khorasani and Milliken (1982). Khorasani and Milliken (1982) used the maximization and minimization approach to construct simultaneous confidence bands for the mean of the Michaelis-Menten Kinetic model with i.i.d. normal error terms and for the Logistic response curve. Brand, Pinnock and Jackson (1973) also constructed large sample confidence bands for the Logistic response curve and its inverse by the same technique. Both papers considered only one regressor variable in the models.

The following section provides detailed derivation of the confidence bands using Lagrange multiplier procedure. A numerical example with three regressor variables is presented in Section 3 to illustrate the method. The conclusion of this procedure is given in Section 4.

2. THE PROCEDURE

Consider model (1). Let $Q(p, \mathbf{x})$ be the $p$th percentile of the model and can be shown to be

$$Q(p, \mathbf{x}) = \theta_{\mathbf{x}} \left( - \log_{e}(1 - p) \right)^{1/\nu}, \ 0 < p < 1$$

(2)

where $\log_{e}$ is the natural logarithm. Note that $Q(p, \mathbf{x})$ is the median lifetime of the model when $p = 1/2$.

Let $\mu(\mathbf{x})$ be the mean lifetime of the model and can be derived to have the form,

$$\mu(\mathbf{x}) = \theta_{\mathbf{x}} \Gamma(1 + 1/\nu)$$

(3)

where $\Gamma(\cdot)$ is the Gamma function.

To find the simultaneous confidence bands of (2) or (3) for all $x_{i} \in R, i = 1,2,...,k$, one needs to find confidence band of $\theta_{\mathbf{x}} = \exp(\mathbf{x}'\beta)$, for all $x_{i} \in R, i = 1,2,...,k$. This implies that we need to find confidence band for $\mathbf{x}'\beta$, for all $x_{i} \in R, i = 1,2,...,k$, then apply the monotonicity property of the exponential function to it.

Note that model (1) with $\theta_{\mathbf{x}} = \exp(\mathbf{x}'\beta)$ can be written as
\[ Y = x' \beta + \sigma Z \]  

(4)

where \( Y = \log_e T \), \( \sigma = 1/\nu \) and \( Z \) has the standard Gumbel distribution with pdf \( \exp(z - \exp(z)) \), \( -\infty < z < \infty \). That is, the p.d.f. of \( Y = \log_e T \), given \( x \), is

\[ f(y|x) = \exp \left[ \frac{y-x' \beta}{\sigma} - \exp \left( \frac{y-x' \beta}{\sigma} \right) \right] / \sigma, \quad -\infty < y < \infty. \]

The estimates of parameters can be obtained in different ways. Equation (4) is essentially a linear regression for which exact methods are available (see Lawless, 1982) but are not usually feasible because of the intensive computations. The least squares method can be used, though it can be inefficient. For more details about the efficiency of least squares estimators, see Prentice and Shillington (1975) and Williams (1978). The more efficient methods are the likelihood ratio methods and the normality approximation of the maximum likelihood estimators in large sample. We can obtain the maximum likelihood estimates (m.l.e.) of \( \beta \) and \( \sigma \) as follows. Suppose a random sample of \( n \) observations is available, \((y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\), where \( y_j \) can be either a log lifetime or a log censoring time and \( x_j' = (x_{0j}, x_{1j}, \ldots, x_{kj}) \), \( x_{0j} = 1 \), for \( j = 1, \ldots, n \). The likelihood function is then

\[ L(\beta, \sigma) = \frac{1}{\sigma} \prod_{j \in A} \exp \left[ \frac{u_j-x_j' \beta}{\sigma} - \exp \left( \frac{u_j-x_j' \beta}{\sigma} \right) \right] \prod_{j \in B} \exp \left[ -\exp \left( \frac{u_j-x_j' \beta}{\sigma} \right) \right], \]

where \( A \) and \( B \) denote the sets for which \( y_j \) is a log lifetime and a log censoring time, respectively. The maximum likelihood equations for \( \beta \) and \( \sigma \) are, respectively,

\[ \frac{\partial \log L}{\partial \beta} = -\sum_{j \in A} x_{ij}/\sigma + \sum_{j=1}^{n} x_{ij} e^{z_j}/\sigma = 0, \text{ for } l = 0, 1, 2, \ldots, k, \]

and

\[ \frac{\partial \log L}{\partial \sigma} = -\tau/\sigma - \sum_{j \in A} z_j/\sigma \sum_{j=1}^{n} z_j e^{z_j}/\sigma, \]

where \( z_j = \frac{u_j-x_j' \beta}{\sigma} \), for \( j = 1, \ldots, n \), and \( \tau \) is the number of lifetimes in set \( A \). The m.l.e. of \( \beta \) and \( \sigma \) can be obtained by solving the above equations using the Newton-Raphson method or some other iteration procedures (see Lawless, 1982).

To find the explicit forms of maximum and minimum, which denoted by \( \lambda_U(x) \) and \( \lambda_L(x) \), respectively, of \( x' \beta \), we assume that sample size \( n \) is large and use the asymptotic
distribution of the maximum likelihood estimators or other asymptotically equivalent
estimators of $\beta = (\beta_0, \beta_1, ..., \beta_k)'$. For a sufficiently large $n$, the distribution of
$$(\hat{\beta} - \beta)' \Sigma^{-1}(\hat{\beta} - \beta)$$
is approximately a chi-square distribution with $(k + 1)$ degrees of freedom, where
$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)'$, $\Sigma^{-1}$ is $(k + 1)$ by $(k + 1)$ matrix and is obtained from the observed
information matrix $I_0$ below:

$$\Sigma^{-1} = -\left( \begin{array}{ccc} \frac{\partial^2 \log L}{\partial \beta_l \partial \beta_s} & \frac{\partial^2 \log L}{\partial \beta_l \partial \sigma} \\ \frac{\partial^2 \log L}{\partial \beta_s \partial \sigma} & \frac{\partial^2 \log L}{\partial \sigma^2} \end{array} \right),$$

with the last row and last column deleted, where

$$\frac{\partial^2 \log L}{\partial \beta_l \partial \beta_s} = - \sum_{j=1}^{n} x_{lj} x_{sj} e^{j} / \sigma^2, \quad l, s = 0, 1, ..., k;$$

$$\frac{\partial^2 \log L}{\partial \beta_l \partial \sigma} = \sum_{j=1}^{n} x_{lj} / \sigma^2 - \sum_{j=1}^{n} x_{lj} e^{j} / \sigma^2, \quad l = 0, 1, ..., k$$

and

$$\frac{\partial^2 \log L}{\partial \sigma^2} = r / \sigma^2 + 2 \sum_{j=1}^{n} z_j / \sigma^2 / 2 \sum_{j=1}^{n} z_j e^{j} / \sigma^2 - \sum_{j=1}^{n} z_j e^{j} / \sigma^2.$$ 

The approximate $100(1 - \alpha)%$ confidence region for $\beta$ then satisfies,

$$(\hat{\beta} - \beta)' \Sigma^{-1}(\hat{\beta} - \beta) \leq \chi^2_{\alpha}(k + 1),$$

where $\chi^2_{\alpha}(k + 1)$ is the upper $\alpha$ percentage point of a chi-square distribution with $(k + 1)$
degrees of freedom.

Now, in order to construct confidence bands for $x' \beta$, we need to determine the
extreme values of $x' \beta$ over the boundary of the confidence region

$$(\hat{\beta} - \beta)' \Sigma^{-1}(\hat{\beta} - \beta) = \chi^2_{\alpha}(k + 1).$$

That is, optimize a linear function $f(\beta) = x' \beta$ subject to the constraint (5). Let
$g(\beta) = (\hat{\beta} - \beta)' \Sigma^{-1}(\hat{\beta} - \beta)$ and applying Lagrange multiplier technique, the solution is
found by solving

$$\nabla_\beta f = \lambda \nabla_\beta g.$$

That is,

$$x = -2\lambda \Sigma^{-1}(\hat{\beta} - \beta),$$

which gives

$$\beta = \hat{\beta} + \frac{1}{2\lambda} \Sigma x.$$
Substituting into the constraint equation (5) gives

\[ 1/\lambda = \pm 2 \sqrt{\frac{\chi^2_\alpha(k+1)}{x' \Sigma x}}. \]

Substituting into the expression for \( \beta \) above, we obtain the solutions

\[ \beta = \tilde{\beta} \pm \sqrt{\frac{\chi^2_\alpha(k+1)}{x' \Sigma x}} \left( \Sigma x \right). \]

So, the extreme values of \( f \) are

\[ x' \tilde{\beta} \pm \sqrt{\frac{\chi^2_\alpha(k+1)}{x' \Sigma x}}. \]  \hspace{1cm} (6)

It can be verified that the maximization and minimization procedure suggested in Brand, Pinnock and Jackson (1973) and Khorasani and Milliken (1982) produces the exact same extreme values as those of (6) after simplification when \( k = 1 \), but it involves more steps and is almost untractable to find extreme points when there are more than one regressor variable in the model.

Let \( \lambda_L(x) = x' \tilde{\beta} - \sqrt{\frac{\chi^2_\alpha(k+1)}{x' \Sigma x}} \)

and \( \lambda_U(x) = x' \tilde{\beta} + \sqrt{\frac{\chi^2_\alpha(k+1)}{x' \Sigma x}} \). Applying the monotonicity of the exponential function, the probability statement becomes,

\[ \Pr(\exp(\lambda_L(x)) < \exp(x' \beta) < \exp(\lambda_U(x)), \forall x_i \in R, i = 1,2,...,k) \approx 1 - \alpha. \]

Therefore, the large sample simultaneous confidence bands for \( Q(p, x) \) and \( \mu(x) \) are \((Q_L, Q_U)\) and \((\mu_L, \mu_U)\) which satisfy, respectively,

\[ \Pr(Q_L < Q(p, x) < Q_U, \forall x_i \in R, i = 1,2,...,k) \approx 1 - \alpha, \]

where \( Q_L = \exp(\lambda_L(x)) \left( - \log_e(1 - p) \right)^{1/\hat{\nu}}, \)

\[ Q_U = \exp(\lambda_U(x)) \left( - \log_e(1 - p) \right)^{1/\hat{\nu}} \] with \( \hat{\nu} \) the m.l.e. of \( \nu; \)

and

\[ \Pr(\mu_L < \mu(x) < \mu_U, \forall x_i \in R, i = 1,2,...,k) \approx 1 - \alpha, \]

where \( \mu_L = \exp(\lambda_L(x)) \Gamma(1 + 1/\hat{\nu}) \) and \( \mu_U = \exp(\lambda_U(x)) \Gamma(1 + 1/\hat{\nu}). \)

3. NUMERICAL EXAMPLE
The numerical illustration of the confidence band computations is the lung cancer data set studied by Prentice (1973). It is of interest that the confidence bands of mean survival time and median survival time (i.e., \( p = 1/2 \) in \( Q(p, x) \)) of patients are constructed under various values of some regressors studied in this subject. Weibull regression model with shape parameter \( \nu \) and \( \theta(x) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3) \), is considered for the data set, where \( x_1 \) is the performance status (a general medical status measure on a scale of 0 to 100: 10, 20, and 30 mean that the patient is completely hospitalized; 40, 50 and 60, partially confined to hospital; and 70, 80 and 90 able to care for himself), \( x_2 \) is patient's age; and \( x_3 \) denotes months from diagnosis to entry into the study.

Data set produced the maximum likelihood estimates of \( \beta \)s and \( \nu \) as

\[
\hat{\beta} = \begin{pmatrix} 0.9752 \\ 0.0615 \\ 0.00213 \\ 0.00334 \end{pmatrix} \quad \text{and} \quad \hat{\nu} = 1.0186,
\]

respectively. The estimated variance-covariance matrix of \( \hat{\beta} \) is:

\[
\Sigma = 10^{-3} \begin{pmatrix} 1598.605 & -7.049 & -19.879 & -2.120 \\ -7.049 & 0.103 & 0.01414 & 0.0303 \\ -19.879 & 0.01414 & 0.337 & -0.0166 \\ -2.120 & 0.0303 & -0.0166 & 0.0868 \end{pmatrix}.
\]

The above information along with the formulation established in Section 2 give the large sample simultaneous confidence bands for mean and \( p \)th percentile survival times. Since three regressors are involved in this example, the confidence bands of mean and median survival times are tabulated for selected values of regressors below to avoid complexity:

<table>
<thead>
<tr>
<th>Regressor values</th>
<th>95% confidence band for ( \mu(x) )</th>
<th>95% confidence band for ( Q(1/2, x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( x_2 )</td>
<td>( x_3 )</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>30</td>
<td>50</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>30</td>
<td>50</td>
<td>15</td>
</tr>
<tr>
<td>30</td>
<td>50</td>
<td>20</td>
</tr>
</tbody>
</table>

Note that Table 1 gives relatively lower bands for median lifetime than those for mean lifetime. This reflects the right skewness of a Weibull distribution (the median is smaller than the mean under a right skewed distribution). For different age groups, \( x_2 = 20 \)
and 50, the bands of the mean and median survival times under the same conditions of performance status ($x_1 = 30$) and months from diagnosis to entry into the study ($x_3 = 5, 10, 15, \text{ and } 20$) are significantly different. Note that the widths of the bands for age group 20 are wider than those of age group 50. This is due to that more uncertainty (or variations) is involved in estimating percentile and mean lifetimes corresponding to age levels outside the range of the data set ($x_2$, patient's age, is between 36 and 70 in the data set studied).

4. CONCLUSION

In this article, we have presented a procedure to construct simultaneous confidence bands for either the $p$th percentile or the mean lifetime when the regressor variables are unconstrained in Weibull regression. This confidence band depends strongly upon the confidence region of the parameters. Without an appropriate region, the confidence band would be meaningless. It is noted that the result applies to more than one regressor variable in the model.

In some applications, the user may be interested in $Q(p, \bm{x})$ or $\mu(\bm{x})$ when realistic constraint information of $\bm{x}$ is incorporated in the analysis. For example, the patients aged from 40 to 50 may be of interest only in lung cancer study. Tighter confidence bands are expected. The amount of improvement of confidence bands under constrained information of regressors is of interest and the authors are currently pursuing these avenues.

REFERENCES


