Parameter estimation through weighted least-squares rank regression with specific reference to the Weibull and Gumbel distributions

J.M. van Zyl and R. Schall

Abstract: Least squares regression based on probability plots, also called rank regression, can be used to estimate the parameters of some distributions. Regression is performed between a function of the empirical distribution function and the order statistic as the independent variable. Using large sample properties of the empirical distribution function and order statistics, weights to stabilize the variance in order to perform weighted least squares regression are derived. Weighted least squares regression is then applied to the estimation of the parameters of the Weibull, the exponential and the Gumbel (extreme value type I) distributions. The weights are independent of the parameters of the distributions considered. Monte Carlo simulation shows that the weighted least-squares estimators outperform the usual least-squares estimators with respect to the mean square error, especially in small samples.

Keywords: Probability plot, Weighted Least-squares regression, Rank Regression, Weibull Distribution, Gumbel Distribution, Estimation.

1. Introduction

Least squares regression methods based on the relationship between the empirical cumulative distribution function (cdf) and the order statistics are frequently used to estimate parameters of distributions. In this paper we propose a weighted least squares regression method, where the weights are proportional to the inverse of the large sample variances of a function of the order statistics. The weighted least squares method will be applied to the problem of estimating the parameters of the Weibull, exponential and Gumbel distributions. The weights are of a simple form and independent of the parameters of the distribution. Simulation results show that the weighted least squares method outperforms the usual unweighted least squares regression with respect to the mean square error.

As a motivating example for our methodology, we consider the two-parameter Weibull distribution with cumulative distribution function

$$F(x;\alpha,\beta) = 1 - \exp(-(\frac{x}{\alpha})^{\beta}),\ x,\alpha,\beta \geq 0, \qquad (1)$$

where α is the scale parameter and β the shape parameter.

Methods for estimating the parameters α and β include the method of moments and maximum likelihood. A simple method of estimation (see Zhang, Xie and Tang, 2007) exploits the linearization of equation (1), namely

$$\log(-\log(1 - F(x;\alpha,\beta)) = \beta\log(x) - \beta\log(\alpha) \qquad (2)$$

Now let $x_1,...,x_n$ denote a sample of size n with corresponding order statistics $x_{(1)} \leq ... \leq x_{(n)}$. For the sample, equation (2) becomes

$$\log(-\log(1 - \hat{F}(x_{(r)};\alpha,\beta)) = \beta\log(x_{(r)}) - \beta\log(\alpha) \qquad (3)$$

where r is the order number and $\hat{F}_r$ is some non-parametric estimate of $F(x_{(r)};\alpha,\beta)$, such as $m_r = r/(n+1)$ or Bernard's median rank estimator (Bernard and Bosi-Levenbach, 1953) $m_r^b = (r-0.3)/(n+0.4)$.

Setting $y_r = \log(-\log(1-\hat{F}_r))$, $x_r = \log(x_{(r)}),\ r = 1,...,n$ equation (3) becomes

$$y_r = -\beta\log(\alpha) + \beta x_r \qquad (4)$$

Zhang, Xie and Tang (2007) consider simple least squares regression of Y against X (as suggested by equation (4)), as well as simple least squares regression of X against Y . Zhang, Xie and Tang (2007) give a detailed review and Monte Carlo study of the performance of these estimation techniques. The form where the logarithm of the order statistics are the independent variables (based on equation (4)), called regression of Y on X by Zhang, Xie and Tang (2007), will be further investigated in the present paper.

The order statistics $x_{(1)} \le ... \le x_{(n)}$ do not have constant variance, nor do the log transformed order statistics X, so that the regression model (4) is heteroscedastic. In this paper we derive approximate weights to stabilize the variances and we show by Monte Carlo simulation that the weighted regression outperforms the unweighted least-squares method.

2. Derivation of weights for least-squares from large sample variances

The weights for the regression will be derived as the inverse of the approximate variance of a scalar function Λ of an order statistic will be derived. It is assumed that the derivative of Λ is continuous at the expected value of the order statistic.

Let $x_1,...,x_n$ denote a sample of size n from a distribution F with corresponding order statistics $x_{(1)} \le ... \le x_{(n)}$. The weighted least squares expression to minimize with respect to the parameters is $\sum_{r=1}^{n} w_{_r} \left[E(\Lambda(x_{(r)})) - \Lambda(x_{(r)}) \right]^2$, where the weight for the r-th squared residual $u_r^2 = [\Lambda(\mathrm{X_r}) - \Lambda(x_{(r)})]^2$ is $w_r = 1/\operatorname{var}(\Lambda(x_{(r)})),\ r = 1,...,n$. The function Λ need not be a linear function of the order statistics.

The statistics $F(x_{(1)}),...,F(x_{(n)})$ are beta distributed with $F(x_{(r)}) \sim Beta(r, n-r+1)$. $E(F(x_{(r)})) = m_r = r/(n+1)$, $\operatorname{var}(F(x_{(r)})) = \dfrac{r(n-r+1)}{(n+2)(n+1)^2} = \dfrac{m_r(1-m_r)}{n+2}$.

Let X_r be such that $F^{-1}(X_r) = r/(n+1)$. Asymptotically $\sqrt{n}[x_{(r)} - X_r] \overset{d}{\rightarrow} N(0, \sigma_r^2)$ with $\sigma_r^2 = \frac{m_r(1-m_r)}{(F'(X_r))^2}, r = 1, ..., n$, provided $F'(m_r) = f(m_r)$ exists (DasGupta, 2008 p. 93). The delta method can now be applied to obtain the approximate variance of a scalar valued function Λ of the order statistics, where we assume that the first derivative of Λ is continuous at X_r and $\Lambda'(X_r) \neq 0$. Then

$$\Lambda(x_{(r)}) - \Lambda(X_r) \overset{d}{\rightarrow} N\left(0, \text{var}(x_{(r)})\left(\frac{d\Lambda(x_{(r)})}{dx_{(r)}}\right)^2_{x_{(r)}=X_r}\right), \text{ r=1,…,n.}$$

It follows that

$$\text{var}(\Lambda(x_{(r)})) \approx \frac{m_r(1-m_r)}{(n+2)(f(X_r))^2}\left(\frac{d\Lambda(x_{(r)})}{dx_{(r)}}\right)^2_{x_{(r)}=X_r} \qquad (5)$$

Furthermore if $\Lambda(\text{x}_{(\text{r})})$ is of the form as $\Lambda(\text{x}_{(\text{r})}) = \Lambda(F(x_{(r)}))$ it can be seen that

$$\left(\frac{d\Lambda(x_{(r)})}{dx_{(r)}}\right)^2_{x_{(r)}=X_r} = \left(\frac{d\Lambda(F(x_{(r)}))}{dF(x_{(r)})}\frac{dF(x_{(r)})}{dx_{(r)}}\right)^2_{x_{(r)}=X_r}$$

$$= (f(X_r))^2\left(\frac{d\Lambda(F(x_{(r)}))}{dF(x_{(r)})}\right)^2_{x_{(r)}=X_r},$$

so that the term $(f(X_r))^2$ cancels in approximation (6). It can be noted that the weights in such a case are not a function of the parameters of the distribution under consideration, and it is possible to apply this method with an explicit expression for the cdf, if a function Λ can be constructed which gives a relationship between the parameters and the cdf of the distribution. This need not be a linear function of the order statistics.

The weights calculated using Bernard's median rank estimator for E(x_(r)), namely $m_r^b = (r-0.3)/(n+0.4)$ (Bernard and Bosi-Levenbach, 1953) instead of $m_r = r/(n+1)$, were also tested.

Order statistics and thus also functions of order statistics are asymptotically independently distributed (Kendall, Stuart and Ord, 1987 p. 462). In this work we treat the residuals, $u_r = \Lambda(\mathrm{x}_{(\mathrm{r})}) - \Lambda(X_r)$ of the least squares regression as if they were independent.

An approximation for the bias term $\Lambda(X_r) - \mathrm{E}(\Lambda(\mathrm{x}_{(\mathrm{r})}))$ can be found by using the second order term of the Taylor expansion of $F(x_{(r)})$. Let $h_r = x_{(r)} - X_r$, $E(h_r) = 0$, the Taylor expansion of $\Lambda(\mathrm{x}_{(\mathrm{r})})$ up to the second order term is

$$\Lambda(\mathrm{x}_{(\mathrm{r})}) \approx \Lambda(X_r) + h_r\Lambda'(X_r) + \frac{1}{2}h^2\Lambda''(X_r) + o_p(1).$$

and $\Lambda(X_r) - \mathrm{E}(\Lambda(\mathrm{x}_{(\mathrm{r})})) \approx -\frac{1}{2}\Lambda''(X_r))E[(\mathrm{var}(\Lambda(x_{(r)}))]$.

Application 1: Weibull distribution

Consider a sample of size n from a two-parameter Weibull distribution with parameters α and β. The relationship $\log(-\log(1-F(x;\alpha,\beta)) = \beta\log(x) - \beta\log(\alpha)$ is used to perform rank regression. The approximate variance of $\log(-\log(1-F(x_{(r)};\alpha,\beta))$ is

$$\begin{aligned}\mathrm{var(log(\text{-}log(1\text{-}F(x_{(r)}))))} &\approx \frac{m_r(1-m_r)}{(n+2)(f(X_r))^2}\left(\frac{d\log(-\log(1-F(x_{(r)};\alpha,\beta)))}{dx_{(r)}}\right)^2_{x_{(r)}=X_r} \\ &= \frac{m_r(1-m_r)}{(n+2)(\log(1-m_r))^2(1-m_r)^2} \\ &= \frac{m_r}{(n+2)(\log(1-m_r))^2(1-m_r)} \\ &= \frac{r}{(n+2)(\log(\frac{n-r+1}{n+1}))^2(n-r+1)}. \qquad (6)\end{aligned}$$

The weighted least-squares regression equation is solved by letting

$$\mathbf{y}' = (\log(-\log(1-m_1),...,\log(-\log(1-m_n)), \quad X = \begin{pmatrix} 1 & \log(x_{(1)}) \\ \vdots & \vdots \\ 1 & \log(x_{(n)}) \end{pmatrix},$$

$W = diag(w_1,...,w_n),\ w_r = \dfrac{(n-r+1)}{r}\left(\log(\dfrac{n-r+1}{n+1})\right)^2,\ r = 1,...,n$.

$\hat{\boldsymbol{\theta}} = (X'WX)^{-1}X'W\mathbf{y}$, $\hat{\boldsymbol{\theta}}' = (-\hat{\beta}\log\hat{\alpha}, \hat{\beta}), \hat{\alpha} = \exp(-\hat{\theta}_1/\hat{\beta}), \hat{\beta} = \hat{\theta}_2$. Let $\mathbf{x}' = (\log(x_{(1)}),...,\log(x_{(n)}))$, then it follows that

$$\hat{\theta}_2 = \frac{\sum_{j=1}^{n} w_j(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n} w_j(x_j - \bar{x})^2},\ \hat{\theta}_1 = \bar{y} - \hat{\theta}_2\bar{x}.$$

Application 2: Exponential distribution

For the exponential distribution with cdf $F(x;\lambda) = 1-\exp(-\lambda x)$, the regression equation is $-\log(1-F(x;\lambda)) = \lambda x$, and

$$\text{var(-log(1-F(}x_{(r)};\lambda))) \approx \frac{r}{(n+2)(n-r+1)} \propto \frac{r}{(n-r+1)}.$$

By solving the least squares regression equations, it follows that

$$\hat{\lambda} = \frac{\sum_{j=1}^{n} -w_j x_j \log(1-m_j)}{\sum_{j=1}^{n} w_j x_{(j)}^2},\ w_j = \frac{n-j+1}{j} \propto 1/\operatorname{var}(-\log(1-F(x_{(r)};\lambda))) \ .$$

Application 3: Gumbel distribution

The Gumbel or extreme value distribution type I for the maximum has cdf $F(x) = e^{-e^{-(x-\mu)/\beta}}$ and the relationship $-\log(-\log(F(x;\mu,\beta)) = x/\beta - \mu/\beta$ is used to perform rank regression. The approximate variance for the transformation is

$\operatorname{var}(-\log(-\log(F(x_{(r)};\mu,\beta))) \approx \dfrac{1-m_r}{(n+2)(\log(m_r))^2 m_r}$, similar to that of the Weibull.

3. Simulation study.

3.1 Variance approximation

For the Weibull distribution, the variance approximation (7) was compared to the true variance by simulation. Residuals

$$u_r = \log(-\log(1-m_r)) - (\beta\log(x_{(r)}) - \beta\log(\alpha)),\ r = 1,...,n\ ,$$

were calculated for 5000 simulation samples and the approximated and true variances plotted against r. The approximation is good even for a relatively small sample size of n=30 (Figure 1). Both $m_r = r/(n+1)$ and the Bernard median ranks were used in the approximation of the variance.

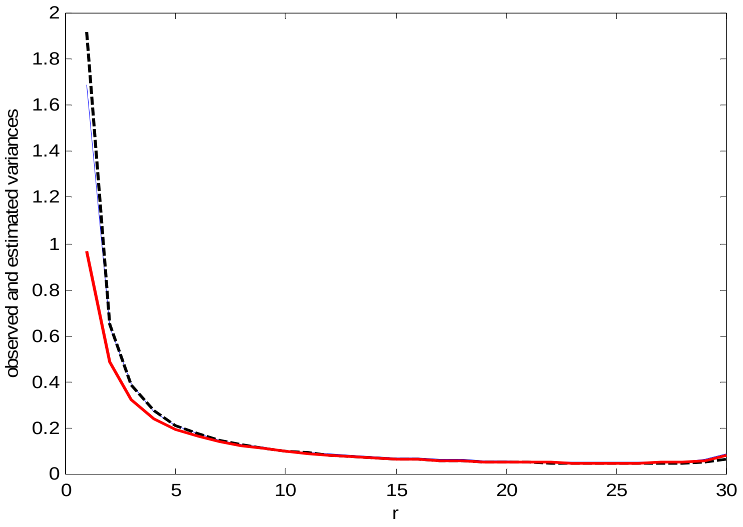

Figure 1. Variance of 5000 residuals, r=1,…,n, in a sample of size n=30, from a Weibull distribution with $\alpha = 1, \beta = 0.5$. The solid line denotes the observed variance, the dashed line the estimated variances using the Bernard method and the dashdot line the usual estimated variances.

For the exponential distribution let $u_r = \log(1 - m_r) - \lambda \log(x_{(r)}), r = 1,...,n$, where sample size used in the simulation is n=15. The true and approximate variances are shown in figure 2.

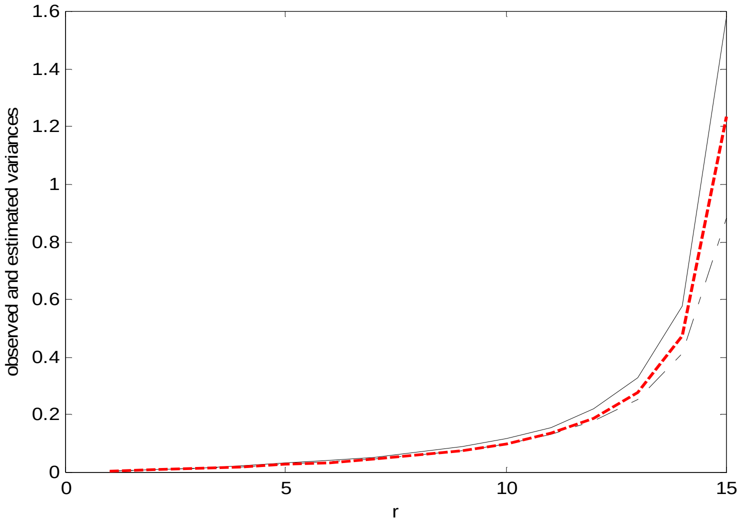

Figure 2. Variance of 5000 residuals, r=1,…,n, in a sample of size n=15, from an exponential distribution with $\lambda = 0.5$. The solid line denotes the observed variance, the dashed line the estimated variances using the Bernard method and the dashdot line the usual estimated variances.

For the Gumbel distribution with parameters μ, β, let $u_r = -\log(-\log(m_r)) - x/\beta + \mu/\beta$. The true and approximated variances calculated from 5000 samples of size 15 are shown in figure 3.

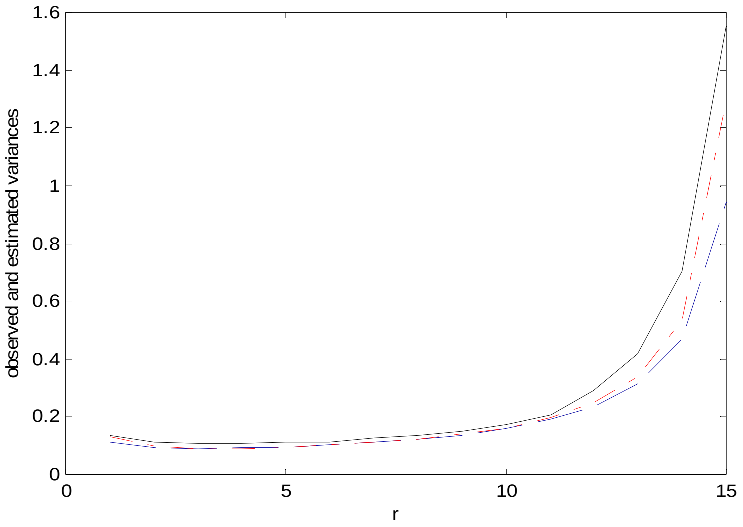

Figure 3. Variance of 5000 residuals, r=1,…,n, in a sample of size n=15, from a Gumbel distribution with $\mu = 0.5, \beta = 2.0$. The solid line denotes the observed variance, the dashdot line the estimated variances using the Bernard method and the dashed line the usual estimated variances.

It can be seen that the variance approximation is reasonable even for relatively small sample sizes, and that Barnard's median ranks result in better approximations of the variances for the Weibull and Gumbel distributions than the usual expected ranks.

3.2 Performance of weighted least squares estimators

In tables 1 and 2 the performance (MSEs) of the weighted least squares and the usual unweighted least squares method for estimating the parameters of the Weibull distribution are compared.

$\beta = 0.5$ $\alpha = 1.0$	MSE LS (β) (Bernard)	MSE LS (α) (Bernard)	MSE Weighted LS (β)	MSE Weighted LS (α)	MSE Weighted LS (β) (Bernard)	MSE Weighted LS (β) (Bernard)
n=10	0.8838 (0.4875)	0.0273 (1.2900)	0.7870 (0.4341)	0.0226 (1.2550)	0.6529 (0.4759)	0.0214 (1.2912)
n=15	0.5432 (0.4846)	0.0178 (1.2310)	0.4736 (0.4472)	0.0145 (1.1913)	0.3991 (0.4788)	0.0128 (1.2120)
n=30	0.2199 (0.4818)	0.0084 (1.1221)	0.1882 (0.4656)	0.0064 (1.0861)	0.1686 (0.4803)	0.0060 (1.1143)
n=100	0.0568 (0.4880)	0.0027 (1.0506)	0.0511 (0.0027)	0.0018 (1.0309)	0.0491 (0.4881)	0.0018 (1.0408)

Table 1. MSE (and Mean) of estimated parameters of the Weibull distribution with $\alpha = 1.0, \beta = 0.5$ (5000 simulated samples).

$\beta = 1.5$ $\alpha = 1.0$	MSE LS (β) (Bernard)	MSE LS (α) (Bernard)	MSE Weighted LS (β)	MSE Weighted LS (α)	MSE Weighted LS (β) (Bernard)	MSE Weighted LS (β) (Bernard)
n=10	0.0550 (1.4517)	0.2364 (1.0416)	0.0519 (1.2952)	0.1978 (1.0333)	0.0520 (1.4487)	0.1856 (1.0321)
n=15	0.0383 (1.4233)	0.1265 (1.0126)	0.0351 (1.3364)	0.1262 (1.0191)	0.0347 (1.4414)	0.1178 (1.0359)
n=30	0.0188 (1.4366)	0.0777 (1.0200)	0.0170 (1.3887)	0.0585 (1.0091)	0.0170 (1.4388)	0.0529 (1.0233)
n=100	0.0053 (1.4695)	0.0250 (1.0102)	0.0049 (1.4648)	0.0165 (1.0042)	0.0052 (1.4674)	0.0159 (1.0093)

Table 2. MSE of estimated parameters of the Weibull distribution with $\alpha = 1.0, \beta = 1.5$. Estimated using weighted least squares and the usual regression method based on 5000 simulated samples.

For the samples sizes investigated, the MSE of the weighted methods outperforms the usual least squares method with respect to MSE, and the use of the Bernard weights decreased the bias too.

Results for the exponential distribution are given in table 3 and 4.

$\lambda = 0.5$	MSE MLE	MSE Weighted LS	MSE Weighted LS (Bernard)
n=10	0.0399 (0.5532)	0.0285 (0.4592)	0.0292 (0.4608)
n=15	0.0231 (0.5358)	0.0189 (0.4673)	0.0195 (0.4688)
n=30	0.0098 (0.5165)	0.0094 (0.4792)	0.0097 (0.4801)
n=50	0.0056 (0.5104)	0.0058 (0.4875)	0.0059 (0.4881)
n=100	0.0027 (0.5051)	0.0029 (0.4935)	0.0030 (0.4937)

Table 3. MSE (and mean) of estimates of the parameter of the exponential distribution with $\lambda = 0.5$ (25000 simulated samples).

$\lambda = 1.5$	MSE MLE	MSE Weighted LS	MSE Weighted LS (Bernard)
n=10	0.3795 (1.6693)	0.2654 (1.3851)	0.2722 (1.3900)
n=15	0.2121 (1.6084)	0.1709 (1.4032)	0.1756 (1.4080)
n=30	0.0885 (1.5517)	0.0846 (1.4398)	0.0867 (1.4426)
n=50	0.0503 (1.5319)	0.0516 (1.4628)	0.0525 (1.4643)
n=100	0.0233 (1.5162)	0.0257 (1.4807)	0.0260 (1.4812)

Table 4. MSE (and mean) of estimates of the parameter of the exponential distribution with $\lambda = 1.5$ (25000 simulated samples).

The weighted estimate outperforms the MLE estimator in smaller samples sizes but in this case the Bernard estimate of expected rank is not the best performer, and the usual estimate of expected rank should be used to derive the weights.

Results for the Gumbel distribution are given in table 5 and 6.

$\beta = 2.0$ $\mu = 0.5$	MSE LS (β) (Bernard)	MSE LS (μ) (Bernard)	MSE Weighted LS (β)	MSE Weighted LS (μ)	MSE Weighted LS (β) (Bernard)	MSE Weighted LS (μ) (Bernard)
n=10	0.9672 (2.5459)	0.4906 (0.3960)	0.7333 (2.5128)	0.4675 (0.4808)	0.4724 (2.2902)	0.4755 (0.5404)
n=15	0.6340 (2.4407)	0.3187 (0.4129)	0.4223 (2.3823)	0.3038 (0.4908)	0.2909 (2.2252)	0.3085 (0.5312)
n=30	0.2766 (2.2742)	0.1553 (0.4298)	0.1526 (2.2011)	0.1444 (0.4890)	0.1186 (2.1167)	0.1455 (0.5098)
n=100	0.0694 (2.1277)	0.0495 (0.4675	0.0340 (2.0644)	0.0465 (0.5015)	0.0309 (2.0375)	0.0467 (0.5081)

Table 5. MSE (and mean) of estimates of parameters of the Gumbel distribution with $\mu = 0.5, \beta = 2.0$ (5000 simulated samples).

$\beta = 4.0$ $\mu = 0.5$	MSE LS (β) (Bernard)	MSE LS (μ) (Bernard)	MSE Weighted LS (β)	MSE Weighted LS (μ)	MSE Weighted LS (β) (Bernard)	MSE Weighted LS (μ) (Bernard)
n=10	4.2569 (5.1190)	1.8820 (0.2898)	3.1528 (5.0409)	1.7819 (0.4655)	2.0452 (4.5980)	1.8134 (0.5853)
n=15	2.5688 (4.8731)	1.3531 (0.3144)	1.6485 (4.7417)	1.2625 (0.4750)	1.1341 (4.4267)	1.2799 (0.5554)
n=30	1.0968 (4.5495)	0.6447 (0.3569)	0.6066 (4.3989)	0.6064 (0.4748)	0.4711 (4.2301)	0.6106 (0.5163)
n=100	0.2762 (4.2518)	0.1978 (0.4189)	0.1381 (4.1263)	0.1819 (0.4856)	0.1259 (4.0725)	0.1822 (0.4987)

Table 6. MSE (and mean) of estimates of parameters of the Gumbel distribution with $\mu = 0.5, \beta = 4.0$ (5000 simulated samples).

The weighted estimate outperforms the usual least squares estimator and Bernard's median ranks are best to use when calculating the weights.

4. Discussion

The weighted least squares method outperforms the usual unweighted least squares method, especially for small sample sizes, and the weights are very easy to calculate.

References

Zhang, L.F., Xie M,Tang, L.C. (2007), “A study of two estimation approaches for parameters of Weibull distribution based on WPP,” *Reliability Engineering & System Safety*, 92, 360 -368.

Bernard, A., Bosi-Levenbach, E.C. (1953), ‘The plotting of observations on probability paper”, *Stat. Neederlandica*,7, 163 – 173.

DasGupta, A. (2008), Asymptotic Theory of Statistics and Probability, Springer Texts in Statistics, NY.

Kendall, M., Stuart, A. and Ord, J.K. (1987), “Kendall’s Advanced Theory of Statistics,” Charles Griffin and Company, London.