Power Function Algorithm for Linear Regression Weights with Weibull Data Analysis

Ross, Robert

DOI
10.1007/978-3-030-95502-1

Publication date
2022

Document Version
Final published version

Published in
Advanced Computing

Citation (APA)
Ross, R. (2022). Power Function Algorithm for Linear Regression Weights with Weibull Data Analysis. In *Advanced Computing: 11th International Conference, IACC 2021* (pp. 529-541). Springer Nature. https://doi.org/10.1007/978-3-030-95502-1

Power Function Algorithm for Linear Regression Weights with Weibull Data Analysis

Robert Ross(✉)

IWO, Zweerslaan 46, 6711GG Ede, The Netherlands
r.ross@iwo.nl

Abstract. Weighted Linear Regression (WLR) can be used to estimate Weibull parameters. With WLR, failure data with less variance weigh heavier. These weights depend on the total number of test objects, which is called the sample size *n*, and on the index of the ranked failure data *i*. The calculation of weights can be very challenging, particularly for larger sample sizes *n* and for non-integer data ranking *i*, which usually occurs with random censoring. There is a demand for a light-weight computing method that is also able to deal with non-integer ranking indices. The present paper discusses an algorithm that is both suitable for light-weight computing as well as for non-integer ranking indices. The development of the algorithm is based on asymptotic 3-parameter power functions that have been successfully employed to describe the estimated Weibull shape parameter bias and standard deviation that both monotonically approach zero with increasing sample size *n*. The weight distributions for given sample size are not monotonic functions, but there are various asymptotic aspects that provide leads for a combination of asymptotic 3-parameter power functions. The developed algorithm incorporates 5 power functions. The performance is checked for sample sizes between 1 and 2000 for the maximum deviation. Furthermore the weight distribution is checked for very high similarity with the theoretical distribution.

Keywords: Asymptotic behavior · Power function · Similarity Index · Weighted linear regression

1 Introduction

The 2-parameter Weibull distribution is widely used for failure data where the lowest value represents the performance of a test object. E.g., the failure time of an electronic device in a destructive test is the time that a first failure is observed. The weakest path to breakdown determines the strength of the device.

Maximum Likelihood (ML) and Linear Regression (LR) can be applied to estimate the Weibull scale parameter α and shape parameter β from a set of observed failure times t_i ($1 \leq i \leq n$), where n is the sample size, i.e., the total number of tested objects. IWO aims at developing data analytics that can be implemented in office software and devices with limited computing capabilities, such as mobile phones and smart components. An advantage of LR over ML in this respect is that LR parameter estimators

D. Garg et al. (Eds.): IACC 2021, CCIS 1528, pp. 529–541, 2022.
https://doi.org/10.1007/978-3-030-95502-1_40

are analytical. An advanced LR method is weighted LR (WLR). It is preferred because of a faster declining bias and scatter in β, matches the ML standard deviation while keeping the analytical advantage of LR (6.3.2.7 in [1]). In general, the calculation of variances for WLR weights can be difficult [2]. Analytical methods do exist in case of Weibull, but are demanding for large sample sizes n and for censored data where data ranking indices i can become non-integer (see Sect. 3). The unavailability of weights for non-integer indices is recognized by IEC [3], recommending to use the weight for the nearest integer ranking instead. A precise weights algorithm would be very useful.

Here, the goal was set to develop an algorithm for the weights that on the one hand can be implemented on light-weight computing devices and in common office software and on the other hand can handle non-integer ranking indices. The development of this algorithm is based on a combination of power functions. The development of the Power model for the WLR weights for Weibull analysis is subject of the paper.

The structure of the discussion is as follows. The concept of asymptotic power functions is discussed in Sect. 2. The subject of WLR with censored data, variances and especially weights is introduced in Sect. 3. This leads to the demanding analytic expressions for WLR weights. Section 4 describes the development of a model for infinite sample sizes which serves as a basis for the algorithm for weights with finite sample sizes. Finally, the performance of the algorithm is tested for maximum deviation and agreement with the theoretical weight distribution employing a (dis)similarity index.

2 Asymptotic Power Functions

Parameter estimators are required to be consistent which means that with increasing sample size n the absolute bias (if not already zero) and scatter must asymptotically approach zero as described by Fisher [4]. This concept translates into the principle that collecting additional data is rewarded by achieving greater estimate accuracy.

A 3-parameter power function of the sample size n can be used as a model that complies with the consistency concept. E.g., it could describe the decline of the bias and the standard deviation of ML, LR and WLR estimators of the Weibull shape parameter [5]. The aimed weight model employs various asymptotic 3-parameter power functions D_n:

$$D_n = E_n - E_\infty = Q \cdot (n - R)^P \tag{1}$$

E_n is the expected value of an estimated parameter from a data set with sample size n. The asymptotic behavior of E_n is usually the study subject. E_∞ is the expected value of an estimated parameter with infinite sample size, i.e., the asymptotic value of E_n. D_n is the difference of E_n and E_∞ that asymptotically approaches zero with increasing n, the variable under consideration. P, Q and R are the 3 parameters of the power function of n that models the asymptotic behavior. In logarithmic form a linear relationship shows:

$$\log(D_n) = \log(Q) + P \cdot \log(n - R) \tag{2}$$

If $\log(D_n)$ is plotted against $\log(n\text{-}R)$, or linear regression applied, then R should be optimized to achieve a straight line. P follows from the slope, and Q from the inverse log of the intercept. Sometimes the power function will be an exact description of the asymptotic behavior, but in other cases it can be a reasonable approximation merely or just does not apply. The latter case can show as an S-shaped curve that cannot be straightened by varying R. Still, it may be worth the exercise.

Sometimes the asymptotic power function must be adjusted. E.g., if the asymptotic behavior is with a declining n, the $n-R$ term may be replaced by $R-n$. If one function asymptotically approaches another function, then E_∞ may be taken not as a constant, but be replaced by that other function. In a more complicated case (like the presently studied WLR weights) various power functions may have to be combined.

The parameters P, Q and R of the power function can be interpreted as follows. The rate of the asymptotic approach is characterized by the power P in the limit $n \to \infty$. Because D_n approaches zero and provided $n\text{-}R > 1$, it is necessary that $P < 0$.

The parameter Q defines the deviation of E_n from E_∞ for small n, or to be more precise $Q = D_{n-R=1}$. It can be interpreted as an amplitude. If $D_n \neq 0$, then $Q \in \mathbb{R}$ having the same sign as D_n. If $D_n = 0$ by definition, then $Q = 0$.

Whereas the concept of consistency focuses on the limit $n \to \infty$ [4], the asymptotic 3-parameter power function introduces a parameter R that can be ignored in that limit. However, for small n, this parameter defines a singularity at $n = R$ with the negative P. The description of the bias for test sets with (very) small sample size n formed the background of the introduction of R [5].

To illustrate this, the variance σ^2 of an infinitely sized population and the expected estimated variance $\langle s_n^2 \rangle$ of an n sized sample drawn from that population are compared. The ratio $\langle s_n^2 \rangle / \sigma^2$ is well-known to be $(n-1)/n$ and asymptotically approaches $E_\infty = 1$ with increasing n. This ratio exactly follows a 3-parameter power function, namely:

$$\frac{s_n^2}{\sigma^2} - 1 = \frac{n-1}{n} - 1 = -1 \cdot (n-0)^{-1} \tag{3}$$

In terms of the 3-parameter function: $P = -1$, $Q = -1$ and $R = 0$.

3 Weibull Parameter Estimation by WLR

If a series of n devices are destructively tested, a series of failure times are observed. Let the failure times t_i (i = 1,..,n) be ranked such that for all i: $t_{i-1} < t_i$. If not all failure times become known in the test, some failure times may remain hidden. Such data are called *censored* or *suspended*. This may occur if a specific test object fails by another mechanism than which is studied.

E.g., devices may be tested destructively to assess the failure behavior of an on board diode. A failed device is short-circuited and has to be withdrawn from the test. If such devices also contain a transistor, some devices may fail due to transistor failure rather than diode failure. For those devices, the diode failure time then remains hidden. It is clear though that the diode failure time must be larger than the observed transistor failure time on that device. Moreover, the ranking of the unknown diode failure time among the next actually observed diode failure times remains unknown. As a consequence, part of the higher rankings of observed diode failure times becomes uncertain too. A method to deal with censored data in the ranking is the Adjusted Ranking Method [3]. Usually, this leads to non-integer expected rankings i of observed failure times.

3.1 Weighted Linear Regression

Two prominent families of Weibull parameter estimation are ML and (W)LR. The focus of the present paper is on WLR, and particularly on the weights calculation challenge, rather than the well-established regression method itself. With WLR parameter estimation a linear relationship is assumed between a plotting position Z and log-failure times ln(t) from which α and β are estimated. This plotting position Z is defined as:

$$Z(p) = \ln(-\ln(1-p)) \equiv \beta \cdot \ln(t) - \beta \cdot \ln(\alpha) \tag{4}$$

Here p is a probability, i.e. a value of the Weibull cumulative distribution $F(t;\, \alpha,\, \beta)$. In a Weibull plot, Z and ln(t) are plotted along the vertical respectively horizontal axis to form a straight graph granted the data are Weibull distributed indeed. Each observation $\ln(t_i)$ is assigned to the i^{th} expected plotting position, $\langle Z_{i,n} \rangle$ which are detailed in Sect. 3.2 below. The two WLR estimators a_{WLR} and b_{WLR} of respectively α and β are found as:

$$a_{WLR} = \exp\left(\overline{\ln(t)}_w - \frac{\overline{Z}_w}{b_{WLR}} \right) \tag{5}$$

$$b_{WLR} = \frac{\overline{\left(Z - \overline{Z}_w\right)^2}_w}{\overline{\left(\left(Z - \overline{Z}_w\right) \cdot \left(\ln t - \overline{\ln t}_w\right)\right)}_w} \tag{6}$$

The suffix w indicates a weighted average. Such a weighted average $\overline{u}_w$ of a series observations u_i is calculated as:

$$\overline{u}_w = \frac{\sum_{i=1}^{n} \left(w_{i,n} \cdot u_i\right)}{\sum_{i=1}^{n} w_{i,n}} \tag{7}$$

The $w_{i,n}$ are the weights assigned to the data $\ln(t_i)$ in the weighted averaging. The weights do also depend on the sample size n, which is the reason for indicating both i and n in the suffix. The $w_{i,n}$ are the inverse of the variances $v_{i,n}$ of the respective plotting positions $Z_{i,n}$:

$$w_{i,n} = \frac{1}{v_{i,n}} = \frac{1}{\langle Z_{i,n}^2 \rangle - \langle Z_{i,n} \rangle^2} \tag{8}$$

The smaller $v_{i,n}$, the heavier weighs observation t_i in the estimation of a_{WLR} and b_{WLR}. The variances and weights calculations are very challenging and subject of this paper.

3.2 Variances of the Plotting Position Z

The variances $v_{i,n}$ of the Weibull plotting positions follow from the first and second moments of Z. An analytic expression for the first moment $\langle Z_{i,n} \rangle$ is:

$$\langle Z_{i,n} \rangle = \left[-\gamma + i \binom{n}{i} \sum_{j=0}^{i-1} \binom{i-1}{j} \frac{(-1)^{i-j} \cdot \ln(n-j)}{n-j} \right] \tag{9}$$

Here, γ is the Euler constant ($\gamma \approx 0.57722$). Rounding errors in the summation of alternately positive and negative terms can have a high impact. For the second moment:

$$\langle Z_{i,n}^2 \rangle = i \binom{n}{i} \sum_{m=0}^{2} \binom{2}{m} (-1)^{2-m} \frac{\partial^m}{\partial s^m} \Gamma(s+1) \Big|_{s=0} \sum_{j=0}^{i-1} \binom{i-1}{j} (-1)^{i-1-j} \frac{(\ln(n-j))^{2-m}}{n-j} \tag{10}$$

These expressions allow to calculate the variances for integer n, but become demanding with increasing n. Both are unsuitable for non-integer ranking indices i that occur with censored data. For light-weight computing these expressions do not suffice either.

Another approach is to determine the moments of Z in the probability domain by numerical integration, cf. (6). The expected j^{th} moment of $Z_{i,n}$ follows from:

$$\langle Z_{i,n}^j \rangle = \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n+1-i)} \int_0^1 [\ln(-\ln(1-p))]^j \cdot p^{i-1} \cdot (1-p)^{n-i} \mathrm{d}p \tag{11}$$

This expression is suitable for non-integer adjusted ranking as there is no summation involved. Singularities occur at $p = 0$ and 1. Numerical integration requires care therefore [6]. The expression as such is also demanding for light-weight computing, but it was used with Mathematica software and Gauss-Legendre quadrature to generate reference tables of $v_{i,n}$ values for n = 1(1)60, 75(1)80, 80(10)120, 125(25)250(250) 2000 as described in [6]. The notation '$A(B)C$' means a sequence from A to C with increment B. All $v_{i,n}$ and $w_{i,n}$ were determined with 10^{-9}–10^{-14} resolution and used for the model.

Figure 1 shows the $v_{i,n}$ for all integer i and sample size n = 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000. Also the $v_{i,n}$ is shown for i = 1(1)10000 and sample size $n \to \infty$.

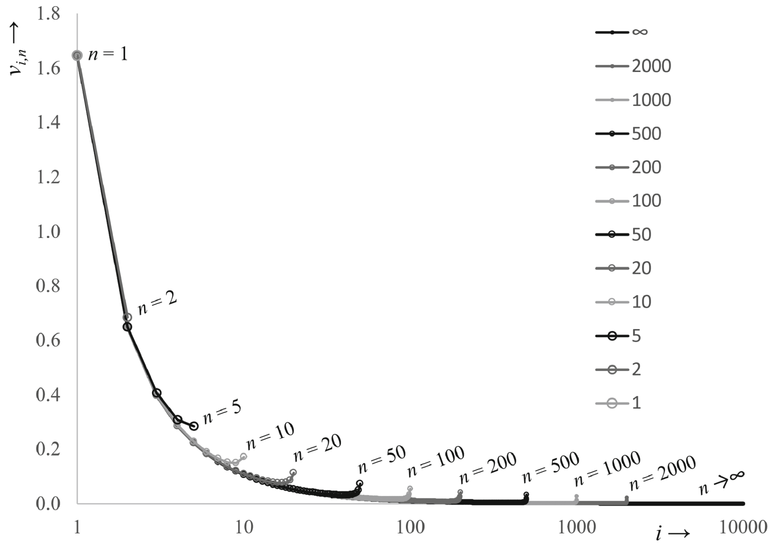

Fig. 1. The variances $v_{i,n}$ for all i with various finite n and $i \leq 10000$ for infinite n.

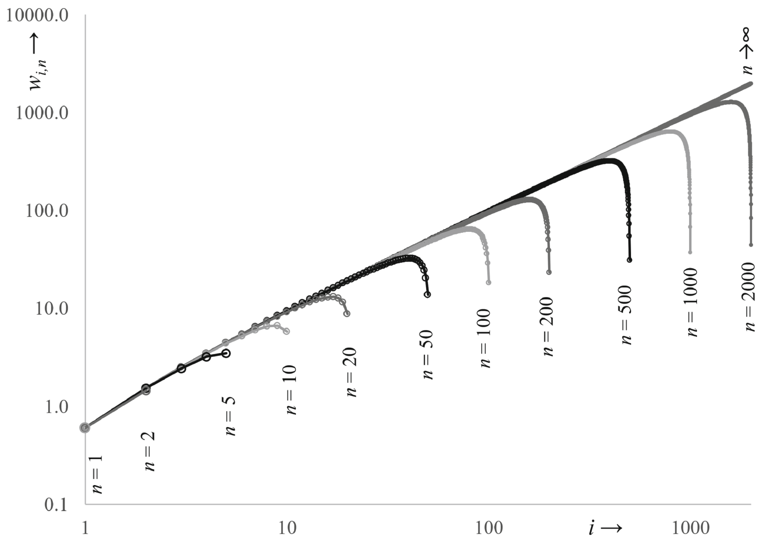

Fig. 2. The weights $w_{i,n}$ for all i with various finite n and $i \leq 1500$ for infinite n.

For all finite n and $i > 1$, the $v_{i,n}$ curve curls up from the curve for infinite n. This tail is small compared to $v_{1,n}$, but with the fast decay of $v_{i,n}$ with increasing i, the curl grows larger relatively. This has a significant impact on the weights $w_{i,n}$ as shown in Fig. 2. For finite n, the weights reach a maximum for some $i \geq n/2$ and then rapidly decrease.

In the following, an algorithm is developed for $v_{i,n}$ and $w_{i,n}$ based on 3-parameter power functions. The process steps to reach the algorithm are discussed in detail. Figure 1 suggests that the $v_{i,n}$ curves might be approached as an adaption of the curve for infinite n. For that reason, firstly the $v_{i,n}$ for infinite n is studied. Secondly, the adjustment for finite $v_{i,n}$ is studied.

4 Power Model for LR Weights

4.1 Infinite Sample Size

The Eqs. (9)–(11) are not easy to interpret for infinite n. However, for $i = 1$ a simple expression follows directly from (8) and (9) that is independent of n:

$$v_{1,n} = \frac{\pi^2}{6} \tag{12}$$

The variances $v_{i,n\to\infty}$ asymptotically approach 0 with increasing i. The same holds for the difference $v_{i,n\to\infty} - v_{i+1,n\to\infty}$. Both can be approximated with an asymptotic power function as in (1). The former gives a fairly good approximation, the latter appears to yield a power function that was later proven to be an exact solution and is elaborated here further. As the difference approaches 0 with $i \to \infty$, the asymptote $E_\infty = 0$.

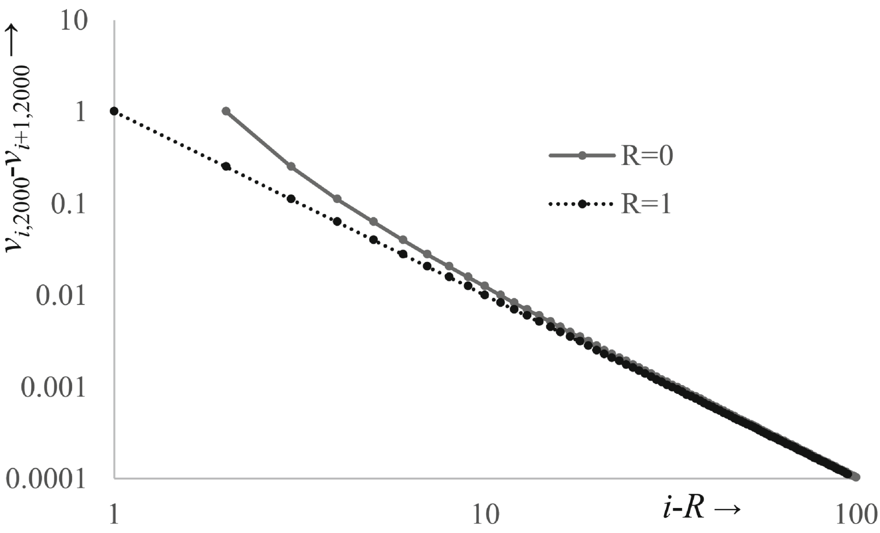

Fig. 3. The asymptotic behavior of $v_{i,n\to\infty} - v_{i+1,n\to\infty}$ for infinite n.

Figure 3 shows $v_{i,n\to\infty} - v_{i+1,n\to\infty}$ for $n = 2000$, i.e., the largest sample size. The log-log plot of D_i against i-R is straight for $R = 1$. The slope yields: $P = -2$ and the intercept $Q = 1$. With increasing n this asymptotic decay of D_i appears more and more accurate.

With power function parameters {P,Q,R} = {−2,1,1}, a relationship is found:

$$n \to \infty : v_{i,n} - v_{i+1,n} = \frac{1}{(i-1)^2} \tag{13}$$

Combining (12) and (13) yields an algorithm for $v_{i,n}$ with infinite n:

$$n \to \infty : \quad v_{i,n} = \frac{\pi^2}{6} - \sum_{j=2}^{i} \frac{1}{(j-1)^2} \tag{14}$$

The sum in (14) with $i \to \infty$ is known to be equal to $\pi^2/6$ (cf. Equation 23.2.24 in [7]).

It is also interesting to do a similar exercise for the weights $w_{i,n\to\infty}$. From (12) follows for $i = 1$ and all n:

$$w_{1,n} = \frac{6}{\pi^2} \tag{15}$$

When listing the values $w_{i,n\to\infty} = 1/v_{i,n\to\infty}$, these appear to asymptotically approach to i−0.5. However, from (15) for $i \downarrow 1$ a deviation of $6/\pi^2 - 0.5 \approx 0.108$ is found. The asymptotic approach of $w_{i,n\to\infty}$ to i−0.5 can again be investigated in terms of an asymptotic power function. Figure 4 shows the difference $D_i = w_{i,n\to\infty} - (i-0.5)$ against i−R. The asymptotic power function parameters are {P,Q,R} = {−1.04,0.098,0.088}. This yields an algorithm for $w_{i,n\to\infty}$:

$$n \to \infty : \quad w_{i,n} \approx i - 0.5 + 0.098 \cdot (i - 0.088)^{-1.04} \tag{16}$$

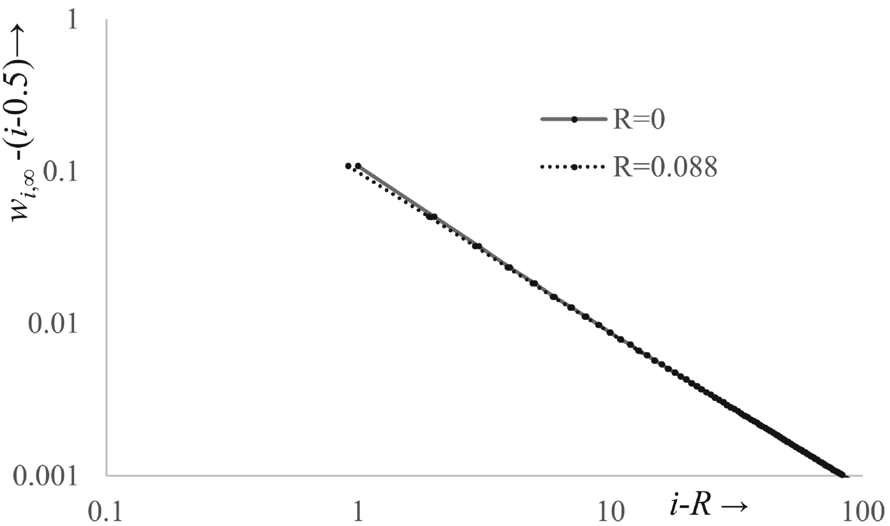

Fig. 4. The asymptotic behavior of $w_{i,n\to\infty} - (i - 0.5)$ for infinite n.

Also the variances for infinite n can be calculated through $v_{i,n\to\infty} = 1/w_{i,n\to\infty}$. This expression is suitable for non-integer i as can occur with censored failure data.

It appeared fruitful to first find an algorithm for $v_{i,n}$ and then convert to $w_{i,n}$. In the following, still, the asymptotic approach of $w_{i,n\to\infty}$ to i−0.5 is used as the foundation, because unlike (14) it is suitable for non-integer i. As $w_{i,n\to\infty}$ approaches i−0.5, likewise $v_{i,n\to\infty}$ approaches $(i-0.5)^{-1}$ with increasing i. This asymptote is a first power function with $\{P_1,Q_1,R_1\} = \{-1, 1, 0.5\}$. In a similar fashion as Fig. 4, the asymptotic behavior $D_i = v_{i,n\to\infty} - (i-0.5)^{-1}$ can be explored. The parameters of a second function are then found as $\{P_2,Q_2,R_2\} = \{-3, -0.1, 0.3445\}$. The algorithm for $v_{i,n\to\infty} = v_{i,\infty}$ is:

$$v_{i,\infty} = Q_1 \cdot (i - R_1)^{P_1} + Q_2 \cdot (i - R_2)^{P_2} \approx (i - 0.5)^{-1} - 0.1 \cdot (i - 0.3445)^{-3} \quad (17)$$

4.2 Finite Sample Size

The algorithm in (17) is suitable for the infinite n case and a good approximation for (severely) censored cases where $i \ll n$. For finite n and the limit $i \to n$, the variances increase (cf. Fig. 1) and the weights decrease sharply (cf. Fig. 2). Figure 5 shows the deviation $D_i = v_{i,n} - v_{i,\infty}$ as a function of i for various n.

For every curve, $D_i \downarrow 0$ for $i \downarrow 1$. This is due to (12). For $n \leq 9$ the curves reach higher maxima $v_{n,n}$ and for $n > 9$ the maxima $v_{n,n}$ are decaying. Noticing that the summation in (14) equals $\pi^2/6$ for $i \to \infty$ (cf. Equation 23.2.24 in [7]), the $v_{n,n}$ asymptotically approach 0. These observations led to the conclusion that this behavior may be described with the product of 3 power functions as a start. In the following the parameters of these three 3-parameter power functions will be indexed 3–5.

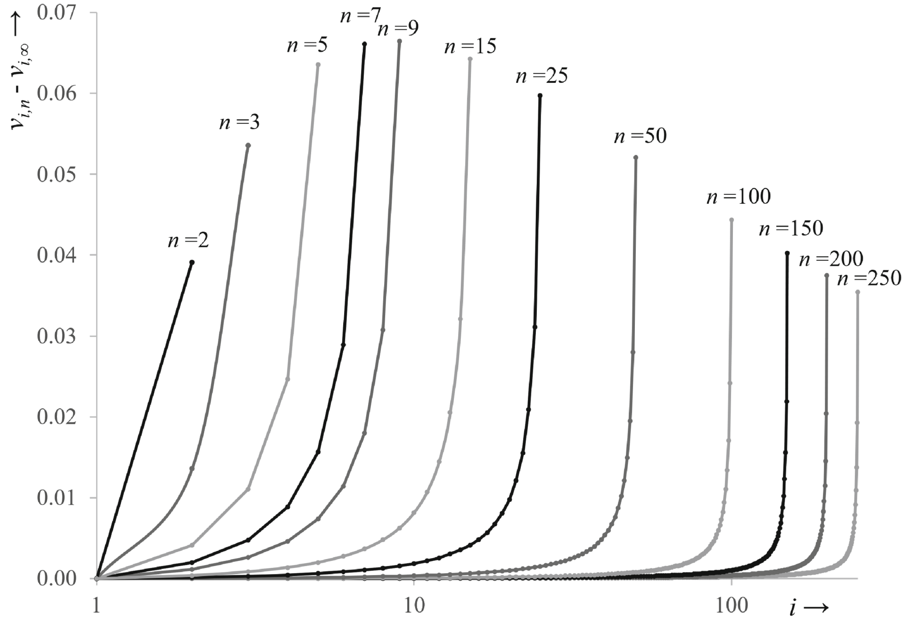

Fig. 5. The asymptotic behavior of $v_{i,n} - v_{i,\infty}$ for various sample sizes n.

The requirement to have $D_1 = 0$ can be met with a power function of i-1 (i.e., R_3 = 1) and a positive power P_3 (note: not asymptotic for $i \to \infty$). As for the amplitude Q_3, a product of 3 power functions will lead to a joint amplitude $Q = Q_3 \cdot Q_4 \cdot Q_5$ which is investigated as one single parameter Q. As all $D_i > 0$ for all $i > 1$ (cf. Fig. 5), the joint $Q > 0$.

The requirement of decaying $v_{n,n}$ can be met with a power function of n-R_4 and a negative power P_4. For $n \downarrow R_4$ this power function approaches its singularity on the one hand, while the power function of $i-1$ approaches 0. If D_i for all n and i can be described with the explored set of power functions, then $R_4 < 1$ and the singularity is not reached for any $n \geq 1$ and $i \geq 1$. It is noteworthy, that the power functions of $i-1$ and $n-R_4$ may very well describe the behavior of $v_{i,n}$ and particularly $v_{n,n}$ in the range $1 \leq i \leq 9$.

Finally, the individual curves for given n, can be described with a power function of $n-i-R_5$ that has a negative power P_5. Also here the singularity should not be reached for any $n - i \geq 0$ and therefore $R_5 < 0$. With $v_{i,\infty}$ as in (17), the variance model becomes:

$$v_{i,n} = v_{i,\infty} + Q \cdot (i-1)^{P_3} \cdot (n - R_4)^{P_4} \cdot (n - i - R_5)^{P_5} \qquad (18)$$

The WLR weight model follows as the inverse, i.e., $w_{i,n} = 1/v_{i,n}$.

The parameters P_3, P_4, P_5, Q, R_4 and R_5 were optimized with the reference data set mentioned above [6]. A mini-max criterion was applied, i.e. the largest ± relative deviation of any *weight* $w_{i,n}$ between the model and reference was minimized. A rounded off result was: {P_3, P_4, P_5, Q, R_4 and R_5} = {1.4, −1.656, −0.75, 0.125, −0.343, −0.8}. The algorithm for the variances $v_{i,n}$ thus becomes:

$$\begin{gathered} v_{i,n} \approx (i-0.5)^{-1} - 0.1 \cdot (i - 0.3445)^{-3} + \\ \left[0.125 \cdot (i-1)^{1.4} \cdot (n + 0.343)^{-1.656} \cdot (n - i + 0.8)^{-0.75}\right] \end{gathered} \qquad (19)$$

From (19) follows the algorithm for the WLR weights $w_{i,n}$:

$$\begin{gathered} w_{i,n} = 1/v_{i,n} \approx 1/\{(i-0.5)^{-1} - 0.1 \cdot (i - 0.3445)^{-3} + \\ \left[0.125 \cdot (i-1)^{1.4} \cdot (n + 0.343)^{-1.656} \cdot (n - i + 0.8)^{-0.75}\right]\} \end{gathered} \qquad (20)$$

This algorithm enables to calculate weights for WLR not only for integer ranking indices i, but also for non-integer ranking indices. As mentioned before, failure data from tests with random censoring is often treated with adjusted ranking that usually yields non-integer indices. The present algorithm is built with 3-parameter power functions. The algorithm is suitable for light-weight computing and can conveniently be embedded in common office software like spreadsheets. The performance is tested below.

5 Evaluation of the Algorithm

The variance model of (17)–(19) and weight algorithm of (20) were evaluated in two ways. The first was part of the optimization, namely, the error of each model weight $w_{i,n,mod}$ calculated with (20) compared to the theoretical or reference weight $w_{i,n,ref}$. Figure 6 shows the maximum ± relative error in $w_{i,n}$ for each sample size n that was involved in the optimization process which was the sequence detailed in Sect. 3.2 above. For all tested $n \leq 2000$ the absolute value of maximum error was < 2.8% which occurred for i = 2000, n = 2000. For all $n \leq 500$ the absolute value of max error was $\leq 1\%$. Relative to the max $w_{i,n} = w_{n,max}$ for each n, the relative error was $\leq$ 0.34% (occurring for n = 2). The model parameters may be fine-tuned further. However, for the purpose of WLR, the approximation is very satisfactory with the $\leq$ 0.34% max error.

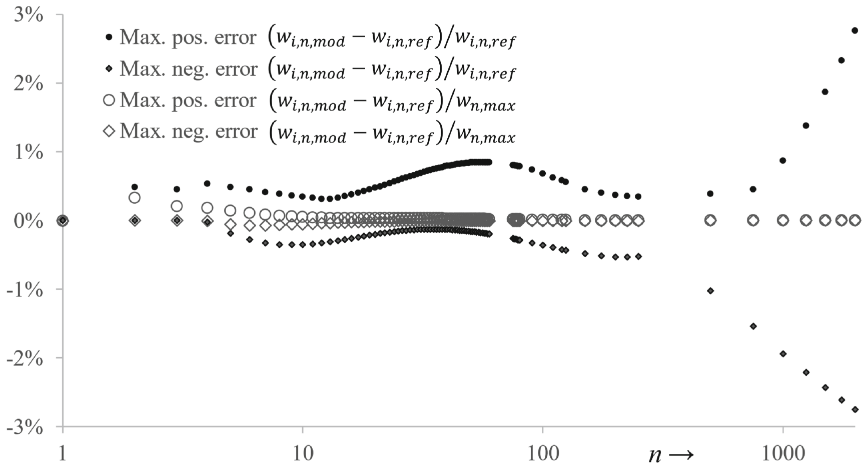

Fig. 6. The maximum relative deviations in the model $w_{i,n,mod}$ from the reference $w_{i,n,ref}$.

The second evaluation tested the shape of the weight distribution. If all model weights would deviate from theory by the same factor, then the *relative* model and theoretical weights would remain the same. The WLR results by model and theory would also be exactly the same. A Similarity Index S_{fg} [1] was employed to test the overall shape. This is a measure for how similar two distributions F and G are, judged by their respective distribution densities f and g. If the distribution densities are identical, then S_{fg} = 1. If f and g have nothing in common, then S_{fg} = 0. The general definition is:

$$S_{fg} = \frac{f \cdot g}{f \cdot f + g \cdot g - f \cdot g} \qquad (21)$$

The terms $f \cdot g, f \cdot f$ and $g \cdot g$ are inner products that can be defined in various ways. In the present case, the weights can be defined as discrete distribution densities:

$$f_{i,n} = \frac{w_{i,n,mod}}{\sum_{i=1}^{n} w_{i,n,mod}} \qquad g_{i,n} = \frac{w_{i,n,ref}}{\sum_{i=1}^{n} w_{i,n,ref}} \tag{22}$$

The S_{fg} of the model and reference can then be determined for each sample size n:

$$S_{fg} = \frac{\sum_{i=1}^{n} \left(f_{i,n} \cdot g_{i,n}\right)}{\sum_{i=1}^{n} \left(f_{i,n} \cdot f_{i,n}\right) + \sum_{i=1}^{n} \left(g_{i,n} \cdot g_{i,n}\right) - \sum_{i=1}^{n} \left(f_{i,n} \cdot g_{i,n}\right)} \tag{23}$$

Additionally, a Dissimilarity Index D_{fg} can be defined as $D_{fg} = 1 - S_{fg}$. D_{fg} can be regarded as $L2$ (sum of squared deviations) normalized by the union of F and G:

$$D_{fg} = 1 - S_{fg} = \frac{f \cdot f + g \cdot g - 2f \cdot g}{f \cdot f + g \cdot g - f \cdot g} = \frac{(f - g) \cdot (f - g)}{f \cdot f + g \cdot g - f \cdot g} \tag{24}$$

For all $n \leq 2000$, S_{fg} appeared $> 0.9999885 \approx 1$. D_{fg} was found largest for $n = 6$, namely, $1.02 \cdot 10^{-5} \approx 0$. Figure 7 shows S_{fg} of the model and theory as a function of sample size n.

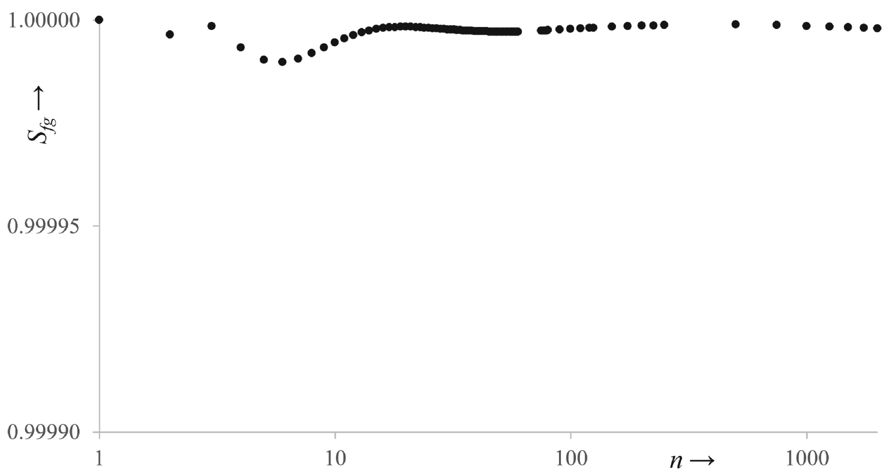

Fig. 7. The similarity of $w_{i,n,mod}$ and $w_{i,n,ref}$, i.e. between the WLR weight algorithm and theory.

6 Discussion and Conclusion

The present paper discussed the making of an algorithm to calculate the weights for WLR with two challenging boundary conditions. The first condition is that the algorithm must be able to handle failure data that have non-integer rankings i $(1 \leq i \leq n)$.

The second condition comes from an on-going project that aims to develop widely accessible data analytics in support of asset management that translates into a requirement of 'light-weight computing'.

The approach of the subject is to combine asymptotic 3-parameter power functions. In earlier research, such functions were applied to cases like the bias and standard deviation of Weibull and other distribution parameters that are efficient and therefore show often a smooth asymptotic behavior. Some aspects of the present challenge, i.e., the WLR weights case, can be associated with asymptotic behavior, but it is more complex. The trajectory to reach the present results required much trial and error. The paper shows the line along which the algorithm was successfully developed.

The parameter values as in (20) are the result of an optimization process. It may be possible to find roots that make the algorithm even more accurate. The performance with the present parameter values is tested for maximum deviation of the model from theory for a wide range of sample sizes, namely, n = 1(1)60, 75(1)80, 80(10)120, 125 (25)250(250)2000 as reported in [6]. The deviation of all weights is within 1% for tested sample sizes $n \leq 500$ and within 2.8% for $n \leq 2000$. For all n, relative to the max $w_{i,n,max}$, the error < 0.34%. Secondly, the distributions of model and reference weights are tested for their similarity. The similarity index S_{fg} was practically 1, i.e. the model is almost identical to theory. The dissimilarity $D_{fg} \leq 1.02 \cdot 10^{-5}$ for all n.

This work on theory and ultimate model is supported by the Netherlands Ministry of Economic Affairs and Climate through the RvO agency, Grant ref. nr. TEUE418008, TKI Project FINDGO and by the EU H2020 R&I program and RvO under ECSEL grant agreement No 826417 (Project Power2Power).

References

1. Ross, R.: Reliability Analysis for Asset Management of Electric Power Grids. Wiley-IEEE Press, Hoboken, NJ (2019)
2. Penn State. https://online.stat.psu.edu/stat501/lesson/13/13.1.Accessed 25 Oct 2021
3. IEC TC112. IEC 62539(E):2007 Guide for the Statistical Analysis of Electrical Insulation Breakdown Data. International Electrotechnical Committee, Geneva (2007)
4. Fisher, R.A.: Theory of statistical estimation, In: Proceedings of the Cambridge Philosophical Society vol. 22(5), pp. 700–725 (1925)
5. Ross, R.: Bias and standard deviation due to weibull parameter estimation for small data sets. IEEE Trans. Dielectr. Electr. Insul. **3**(1), 28–42 (1996)
6. Ross, R., Ypma, P.A.C., Koopmans, G.: Weighted linear regression based data analysis for decision making after early failures. In: Proceedings of IEEE PES ISGT-Asia 2021, paper 148, Brisbane (2021)
7. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions, 9th edn. Dover Publications Inc., New York (1970)