Dynamical Systems and Control 动力系统与控制, 2015, 4, 54-60
Published Online July 2015 in Hans. http://www.hanspub.org/journal/dsc
http://dx.doi.org/10.12677/dsc.2015.43007

Hans汉斯

A Refine Study on the Least-Squares Estimation for Two-Parameter Weibull Distribution

Ling Chen, Yanran Yu, Rongmei Ding, Cheng Li*, Changjin Yang

School of Urban Rail Transportation, Soochow University, Suzhou Jiangsu
Email: *licheng@suda.edu.cn

Received: Jun. 15th, 2015; accepted: Jul. 8th, 2015; published: Jul. 14th, 2015

Abstract

Weibull distribution is one of the common random variable distributions in reliability science and engineering. Despite the fact that log-linearizing the nonlinear reliability function with Weibull distribution contributes to solving the Weibull parameters, the precision of parameter estimation is reduced. Thus, a method by combining Taylor series expansion and least square method is proposed to improve fitting precision of the Weibull distribution. Contrastive analyses on Taylor series expansion-least square method, common least square method and weighted least square method are conducted to access the fitting effects via numerical simulation and calculation. The results show that the proposed method can reduce the Weibull curve fitting error effectively and thus can be reference for reliability test.

Keywords

Weibull Distribution, Linearization, Least-Squares Method, Taylor Series

二参数威布尔分布最小二乘法估计的优化研究

陈　玲，余衍然，丁荣梅，李　成*，杨昌锦

苏州大学，城市轨道交通学院，江苏 苏州
Email: *licheng@suda.edu.cn

*通讯作者。

收稿日期：2015年6月15日；录用日期：2015年7月8日；发布日期：2015年7月14日

摘　要

威布尔分布是可靠性科学和工程中常用的随机变量分布之一。将具有威布尔分布的非线性可靠度函数对数线性化，可方便威布尔参数的求解，但却降低了参数估计的精度。针对这个问题，提出了应用泰勒级数展开结合最小二乘法提高威布尔分布的拟合精度。通过数值模拟和实例计算，对比分析了泰勒级数展开－最小二乘法与普通最小二乘法及其加权处理的拟合效果。结果表明，该方法可以有效降低威布尔曲线拟合的误差，为可靠性试验提供参考。

关键词

威布尔分布，线性化，最小二乘法，泰勒级数

1. 引言

二参数威布尔分布模型广泛应用于机械零部件比如滚动轴承的故障诊断或寿命预测。考虑到产量与成本的因素，如何利用小样本的产品失效试验数据尽可能精确地估计出威布尔参数，成为当前可靠性研究的热点之一[1]。

常用的威布尔分布参数估计方法包括极大似然估计和线性化最小二乘法，其中后者由于计算简便直观而广泛应用于工程实践中，但估计精度不高。为提高线性化最小二乘法估计威布尔参数的精度，可以变换威布尔参数估计的方向[2] [3]，这主要是由于单方向的参数估计只考虑了单个方向的坐标误差。另一思路认为在非线性回归模型转化为线性模型时，模型中随机变量的分布特征会改变，不再满足 Gauss-Markov 假定，从而影响线性最小二乘法估计的精度[4]。因此，该思路通常对普通最小二乘法进行加权处理，如经典的 Bergman 加权最小二乘法[5]。然而，加权最小二乘法需要可靠度的估计值与实际值相差足够小，这样才能保证误差模型的转换成立。因此加权最小二乘法在小样本的情况下并非最佳。

本文从误差模型的角度切入，提出运用泰勒级数展开—最小二乘法(Taylor Series Expansion-Ordinary Least Squares Estimation, TSE-OLS) [6]来估计威布尔分布参数，利用 Monte-Carlo 方法考察完全失效数据的拟合效果，并通过实例计算分析 TSE-OLS 对比普通最小二乘法估计(Ordinary Least Squares Estimation, OLSE)及加权最小二乘法估计(Weighted Least Squares Estimation, WLSE) [5]的优点。

2. 精度问题的提出

一般地，威布尔分布最小二乘法估计采取反变换法。威布尔分布的累积分布函数有如下形式：

$$F(t)=1-\mathrm{e}^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{1}$$

如果 $F(t)$代表失效概率，则可靠度 $R(t)$可表示为

$$R(t)=1-F(t)=\mathrm{e}^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{2}$$

对等式两边取两次对数将其线性化为

$$\ln\left[\ln\left(\frac{1}{R(t)}\right)\right]=\beta\ln t-\beta\ln\eta \tag{3}$$

令 $Y=\ln\left[\ln\left(\frac{1}{R(t)}\right)\right]$，$X=\ln t$，$b=\beta$，$a=-\beta\ln\eta$，从而将(3)式表示成一元线性回归模型的形式：

$$Y=a+bX+e \tag{4}$$

其中，e 为随机误差。将故障间隔时间或失效时间数据从小到大排序 $t_1\le t_2\le\cdots\le t_n$ (n 为可靠性试验的样本量)，因此新的回归模型的自变量 $X_1,X_2,\cdots,X_n$ 也按从小到大排序。

对于无删失数据的试验样本，可以利用近似中位秩公式等方法初步估算出各个失效时间的失效概率

$$F_i(t_i)=\frac{i-0.3}{n+0.4} \tag{5}$$

其中，$i=1\sim n$。根据式(3)可获得因变量的序列 $Y_1,Y_2,\cdots,Y_n$，再由最小二乘原理估计出回归系数 $\hat{a}$ 和 $\hat{b}$：

$$\begin{cases}\hat{a}=\dfrac{\sum\limits_{i=1}^{n}X_i^2\sum\limits_{i=1}^{n}Y_i-\sum\limits_{i=1}^{n}X_i\sum\limits_{i=1}^{n}X_iY_i}{n\sum\limits_{i=1}^{n}X_i^2-\left(\sum\limits_{i=1}^{n}X_i\right)^2}\\ \hat{b}=\dfrac{\sum\limits_{i=1}^{n}Y_i-n\hat{a}}{\sum\limits_{i=1}^{n}X_i}\end{cases} \tag{6}$$

最后通过反变换获得形状参数和尺寸参数的估计值：$\hat{\beta}=\hat{b}$，$\eta=\mathrm{e}^{(-\hat{a}/\hat{b})}\eta$。

上述方法避免了非线性拟合和极大似然估计的复杂数值积分过程，同时解决了原方法可能无法得到最优解的难题，简化了计算，但却使参数估计的精度下降。下面从理论分析的角度，证明通过对数线性化最小后得到的回归系数，不再是最小二乘意义下的最优回归系数。

由最小二乘原理可知[7]，尽管可以使式(4)中的残差平方和取到最小，如式(7)所示，但无法保证原始回归模型的残差平方和最小。

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\left[Y_i-\left(\hat{a}+\hat{b}X_i\right)\right]^2=\min_{a,b}\sum_{i=1}^{n}\left[Y_i-\left(a+bX_i\right)\right]^2 \tag{7}$$

由于

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\left[\ln\left(\ln\left(1/R_i\right)\right)-\ln\left(\ln\left(1/\hat{R}_i\right)\right)\right]^2=\sum_{i=1}^{n}\left(\frac{\ln R_i}{\ln\hat{R}_i}\right)^2 \tag{8}$$

为了使 $\sum\limits_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$ 尽可能小，不妨取理想状态 $\sum\limits_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=0$，那么有 $Y_i=\hat{Y}_i$，即所有点都在拟合直线上。通过反变换，有 $\ln R_i=\ln\hat{R}_i$，代入式(8)则 $\sum\limits_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=n$，这与前述的理想状态是矛盾的。因此，对数线性化后的回归模型不能保证原始的非线性回归模型的残差平方和最小。

3. 对数线性化最小二乘法的加权处理

针对上述问题，Bergman 加权处理从原始的误差模型出发得到[8]

$$R_i-\hat{R}_i=\Delta R_i\approx\frac{\mathrm{d}R_i}{\mathrm{d}Y_i}\Delta Y_i=\frac{1}{\frac{\mathrm{d}Y_i}{\mathrm{d}R_i}}\Delta Y=\frac{1}{\frac{\mathrm{d}Y_i}{\mathrm{d}R_i}}\left(Y_i-\hat{Y}_i\right) \tag{9}$$

$$\sum_{i=1}^{n}\left(R_i-\hat{R}_i\right)^2=\sum_{i=1}^{n}\left[\frac{1}{\frac{\mathrm{d}Y_i}{\mathrm{d}R_i}}\left(Y_i-\hat{Y}_i\right)\right]^2 \tag{10}$$

由式(3)可推导出权函数 $\omega_i=\dfrac{1}{\frac{\mathrm{d}Y_i}{\mathrm{d}R_i}}=R_i\ln R_i$，那么根据最小二乘原理，我们推得：

$$\begin{cases}\hat{b}=\dfrac{\sum_{i=1}^{n}\omega_i^2X_iY_i\sum_{i=1}^{n}\omega_i^2-\sum_{i=1}^{n}\omega_i^2Y_i\sum_{i=1}^{n}\omega_i^2X_i}{\sum_{i=1}^{n}\omega_i^2X_i^{\,2}\cdot\sum_{i=1}^{n}\omega_i^2-\left(\sum_{i=1}^{n}\omega_i^2X_i\right)^2}\\ \hat{a}=\dfrac{\sum_{i=1}^{n}\omega_i^2X_iY_i-\hat{b}\sum_{i=1}^{n}\omega_i^2X_i^{\,2}}{\sum_{i=1}^{n}\omega_i^2X_i}\end{cases} \tag{11}$$

同样地，通过第 1 节中的反变换得到形状参数和尺寸参数的估计值。要使加权函数成立，必须满足随机误差 e 的绝对误差足够小[6]。但在小样本($n\le 20$)的情况时，由于近似中位秩误差较大，很难满足该条件，此时加权处理并非符合严格意义上的最小二乘原理。

4. 泰勒级数展开－最小二乘法估计

文[7]提出利用泰勒级数展开结合最小二乘法的思想来提高对数线性化，他们的工作在幂函数的参数估计中取得了良好的效果。与之不同的是，这里我们主要结合威布尔函数介绍其改进算法。

设 $R_i=f\left(t_i;\beta,\eta\right)=\mathrm{e}^{-\left(\frac{t_i}{\eta}\right)^{\beta}}$，以 OLSE 得到的估计参数 β_0 和 η_0 作为初值，将其展开成一阶泰勒级数：

$$R_i=f\left(t_i;\beta_0,\eta_0\right)+\frac{\partial f}{\partial\beta}\left(\beta_0,\eta_0\right)\left(\beta-\beta_0\right)+\frac{\partial f}{\partial\eta}\left(\beta_0,\eta_0\right)\left(\eta-\eta_0\right)+e_i \tag{12}$$

上式可以写成

$$R_i=A_i+B_i\beta+C_i\eta+e_i \tag{13}$$

其中：

$$A_i=\mathrm{e}^{-\left(\frac{t_i}{\eta_0}\right)^{\beta_0}}\left[1+\beta_0\left(\frac{t_i}{\eta_0}\right)^{\beta_0}\ln\left(\frac{t_i}{\eta_0}\right)-\beta_0\left(\frac{t_i}{\eta_0}\right)^{\beta_0}\right]$$

$$B_i=\mathrm{e}^{-\left(\frac{t_i}{\eta_0}\right)^{\beta_0}}\left[-\left(\frac{t_i}{\eta_0}\right)^{\beta_0}\ln\left(\frac{t_i}{\eta_0}\right)\right]$$

$$C_i=\mathrm{e}^{-\left(\frac{t_i}{\eta_0}\right)^{\beta_0}}\left[\beta_0t_0^{\beta_0}\eta_0^{-\beta_0-1}\right]$$

该方法将非线性回归问题转化为多元线性回归，同时保持了非线性问题的最小二乘原理，而且回归系数没有发生变化，因此为复杂的非线性函数的参数估计带来了便利。这里，利用多元最小二乘法求出

形状参数和尺度参数的最佳估计值如下：

$$\begin{cases} \hat{\beta} = \dfrac{\left(\sum_{i=1}^{n} R_i B_i - \sum_{i=1}^{n} A_i B_i\right)\sum_{i=1}^{n} C_i^2 - \left(\sum_{i=1}^{n} R_i C_i - \sum_{i=1}^{n} A_i C_i\right)\sum_{i=1}^{n} B_i C_i}{\sum_{i=1}^{n} B_i^2 \cdot \sum_{i=1}^{n} C_i^2 - \left(\sum_{i=1}^{n} B_i C_i\right)^2} \\ \hat{\eta} = \dfrac{\sum_{i=1}^{n} R_i B_i - \sum_{i=1}^{n} A_i B_i - \sum_{i=1}^{n} B_i^2 \beta}{\sum_{i=1}^{n} B_i C_i} \end{cases} \tag{14}$$

对于任意小的 $\delta > 0$，若 $\left|\hat{\beta} - \beta_0\right| \le \delta$，$|\hat{\eta} - \eta_0| \le \delta$ (15)

则 $\hat{\beta}, \hat{\eta}$ 为较理想的估计值。若式(15)不成立，则令 $\beta_0 = \hat{\beta}, \eta_0 = \hat{\eta}$，作为新的初始估计值重复以上步骤，直到满足式(15)，从而保证了参数估计的精度。

5. 数值模拟与实例计算

利用 Monte-Carlo 法[9]模拟出若干组不同样本量(每组 N = 5000 次)的标准威布尔分布($\beta = \eta = 1$)的随机变量，以均方误差的期望[10] $E(MSE) = \frac{1}{N}\left[\frac{1}{n}\sum_{i=1}^{n}\left(R_i - \hat{R}_i\right)^2\right]$ 来评估三种方法的拟合效果。其模拟结果如图 1 所示。

从图 1 可见，WLSE 和 TSE-OLSE 方法的残差平方和都比 OLSE 小，随着样本量的增加，二者的 MSE 大小趋于一致。然而在小样本的情况下(n = 5~20)，TSE-OLSE 比 WLSE 下降得更明显，尤其在 n = 5 时，WLSE 的均方误差期望仅比 OLSE 低 4×10^{-4}，而 TSE-OLSE 的均方误差期望降低更显著(约 1.6×10^{-3})。因此，若可靠性试验限于成本等因素只有少量样本时，TSE-OLSE 方法更合适。

接下来通过 3 个实例数据(如表 1~3 所示)计算，进一步对比三种参数估计方法。计算结果如表 4 所示，其中 O 代 OLSE，W 代表 WLSE，T 代表 TSE-OLSE。从表中的均方误差对比可以看出，无论是小样本(n = 7)，中样本(n = 33)，还是大样(n = 105)，TSE-OLSE 都优于其它两种方法。

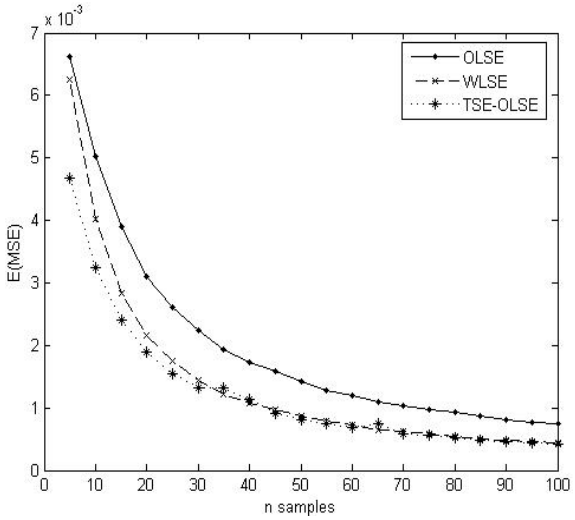

Figure 1. The expectation of mean square errors in different sample sizes
图 1. 不同样本量下的均方误差期望

Table 1. Time between failures of complex product A/h
表 1. 某复杂产品 A 的故障间隔时间/h [2]

t_1	t_2	t_3	t_4	t_5	t_6	t_7
59.18	234.66	240.43	284.91	596.18	738.90	1817.14

Table 2. Time between failures of numerical control system B/h
表 2. 某数控系统 B 的故障间隔时间/h [11]

t_1 - t_{11}	t_{12} - t_{22}	t_{23} - t_{33}
79.45	446.03	932.60
104.11	459.18	972.05
143.56	552.88	986.85
196.16	566.03	1261.37
248.77	695.89	1498.08
250.41	698.63	1590.14
276.71	735.34	1603.29
367.12	776.44	1708.49
395.07	814.25	1894.25
397.81	840.55	2326.58
408.22	919.45	2841.10

Table 3. Time between failures of complex product C/h
表 3. 某复杂产品 C 的故障间隔时间/h [2]

t_1 - t_{15}	t_{16} - t_{30}	t_{31} - t_{45}	t_{46} - t_{60}	t_{61} - t_{75}	t_{76} - t_{90}	t_{91} - t_{105}
18.29	81.81	149.47	234.66	380.96	571.55	843.02
21.19	82.43	152.45	240.43	381.87	574.80	846.21
23.74	97.62	152.49	248.83	383.69	582.03	855.10
23.84	107.09	154.02	249.78	409.06	595.91	858.08
25.39	117.69	154.43	250.52	413.52	596.18	901.78
32.94	118.60	159.52	254.25	417.62	599.37	977.26
35.26	119.63	165.86	260.70	440.96	519.23	1001.10
35.51	127.62	167.10	275.24	442.37	629.93	1001.59
35.84	129.11	185.80	276.10	497.07	642.07	1095.94
47.17	130.60	194.33	283.84	500.55	562.36	1096.44
47.17	130.60	216.47	284.91	509.98	591.23	1202.71
47.42	142.52	221.97	285.65	511.98	726.49	1502.14
51.31	142.77	223.05	331.71	519.97	738.90	1691.17
59.18	143.14	226.44	357.29	525.38	746.60	1786.68
67.49	144.34	227.93	377.89	547.72	832.26	1817.13

Table 4. Calculation results of instance A/B/C
表 4. 实例 A/B/C 计算结果

算例	$\hat{\beta}$			$\hat{\eta}$		
	O	W	T	O	W	T
A (n = 7)	0.965	0.898	0.996	607.09	559.44	557.85
B (n = 33)	1.359	1.262	1.285	922.29	906.12	903.35
C (n = 105)	1.12	1.016	1.03	423.70	423.67	423.90

算例	MSE		
	O	W	T
A (n = 7)	4.42×10^{-3}	4.34×10^{-3}	4.01×10^{-3}
B (n = 33)	8.98×10^{-4}	6.93×10^{-4}	6.84×10^{-4}
C (n = 105)	6.72×10^{-4}	4.19×10^{-4}	4.14×10^{-4}

6. 结论

本文介绍了威布尔分布参数估计的对数线性化最小二乘法及其加权处理，提出了利用泰勒级数展开结合最小二乘法的迭代思想以提高参数估计的精度。通过对标准威布尔分布的完全样本数据进行数值模拟和实例计算，对比分析了泰勒级数展开－最小二乘法、普通最小二乘法、加权最小二乘法等三种方法的拟合效果。结果表明泰勒级数展开－最小二乘法比其它两种方法的均方误差更小，尤其在小样本的情况下更能体现其优势。基于本文工作，可进一步探讨删失样本、不同的形状参数和尺寸参数组合的拟合问题，为工程可靠性试验及零部件寿命预测提供参考。

致　谢

感谢国家自然科学基金(11202145，11202144)和江苏省自然科学基金(BK2012175，BK20130303)对本工作开展提供的支持。

参考文献 (References)

[1] 丛伟，陈晓阳，王志坚，顾家铭 (2013) Weibull 分布产品小样本定时截尾试验方案下的可靠性评估. *中国机械工程*, **24**, 1891-1896.

[2] 王晓峰，申桂香，张英芝，陈炳锟，郑珊，刘葳 (2011) 可靠性模型参数估计方法的对比. *华南理工大学学报(自然科学版)*, **39**, 47-52.

[3] Zhang, L.F., Xie, M. and Tang, L.C. (2007) A study of two estimation approaches for parameters of Weibull distribution based on WPP. *Reliability Engineering and System Safety*, **92**, 360-368. http://dx.doi.org/10.1016/j.ress.2006.04.008

[4] Bergman, B. (1986) Estimation of Weibull parameters using a weight function. *Journal of Materials Science Letters*, **5**, 611-614. http://dx.doi.org/10.1007/BF01731525

[5] Hung, W.L. (2001) Weighted least-squares estimation of the shape parameter of the Weibull distribution. *Quality and Reliability Engineering International*, **17**, 467-469. http://dx.doi.org/10.1002/qre.423

[6] 张大克，王玉杰 (2007) 非线性回归模型线性化后的参数估计精度问题. *天津科技大学学报*, **2**, 68-71.

[7] 陶菊春，吴建民 (2003) 可线性化非线性回归预测模型的剖析与改进. *数学的实践与认识*, **2**, 7-12.

[8] 赵增炜，刘岭，王文昌 (2008) 非线性回归的线性拟合加权最小二乘估计. *中国医院统计*, **1**, 1-2.

[9] 张仙凤，吕志鹏 (2006) 基于 MATLAB 的蒙特卡罗方法在可靠性设计中的应用. *装备制造技术*, **4**, 76-77.

[10] Al-Fawzan, M.A. (2000) Methods for estimating the parameters of the Weibull distribution. King Abdulaziz City for Science and Technology, Riyadh.

[11] 张海波，贾亚洲，周广文 (2005) 数控系统故障间隔时间分布模型的研究. *哈尔滨工业大学学报*, **37**, 198-200.