A Data Analysis Method and Its Applications in EXCEL

Jinlei Qin[1]
Information and Network Management Center, North China Electric Power University, Baoding, China
Email: jlqin717@163.com

Yuguang Niu and Zheng Li
State Key Laboratory of Alternate Electric Power System with Renewable Energy Sources,
North China Electric Power University, Beijing, China
Department of Computer Science and Technology, North China Electric Power University, Baoding, China
Email: nyg@ncepu.edu.cn, yeziperfect@163.com

***Abstract*—Data analysis is a quite important process and has been extensively employed in many areas. Especially in statistics, the distribution type test of data often needs to be handled. This paper presents that linear regression as a type of universal method can be applied to the distribution type test. In reliability engineering due to the failure data is commonly of non-linear relationship so that the linear regression method can not be directly employed. This difficulty can be resolved through linear transformation. Aimed for the linearization procedure of four kind of typical distributions, e.g., exponential distribution, normal distribution, logarithmic normal distribution and two-parameter Weibull distribution, their transformations are respectively different. After linearization transformations, correlation coefficients had been used as a criterion to choose the most matched distribution type. This method can be conveniently operated in MS EXCEL. Several samples and experiments illustrated the detailed transformation and the efficiency than other methods.**

***Index Terms*—data analysis, failure data, linear regression, correlation coefficient, distribution type test**

I. INTRODUCTION

Data analysis is a quite important process and has been extensively employed in many fields such as energy, software, communication and various engineering etc [1-3]. Some related methods had been researched by many scholars. Sun et al. [4] provides a robust data-recovery method based on functional data analysis to enhance the reliability of multichannel sensor system. Dauck et al [5] have developed an industrial data analysis (IDA) platform that automates the data analysis process to a large extend. The IDA platform uses fuzzy knowledge bases to match user requirements to features of analysis methods and to select, configure and execute IDA processes automatically. Lu et al [6] introduce a new lifetime evaluation method that can accomplish rapidly the long lifetime evaluation by using short-term aging data. In Ref [7], the author developed an efficient data analysis package for personal computer use in response to growing needs of the wind industry. Kovac et al [8] present a data analysis software package 'AFV-SOFT' that has been developed for the evaluation of alternative fuel vehicle performance. Jiang [9] has simulated the characters of diode by the NI Multisim platform and plotted conveniently voltage-ample characteristic curve in MS EXCEL according to the data analysis results. The curves plotted by EXCEL are more accurate and intuitive than the traditional hand-drawn curves. Shi [10] works out a EXCEL tool by taking advantage of the VBA language to accomplish the judgment function of data analysis. Some data analysis methods were also applied to software performance prediction and evaluation [11, 12]. The mimic situations arise frequently in reliability engineering and the rules or relationships among data can frequently arouse researcher's interests. Among a group of given failure data, the most concerned issue is which the distribution type belongs to.

Hypothesis test method is a traditional way for this issue. The general methods include Kolmogorov-Smirnov (K-S) test, Chi-square χ^2 test, Anderson-Darling (A-D) test and Crammer-von Mises (C-M) test [13-19]. However among those above mentioned methods, their computational process is always so complicated that more probability and statistical knowledge requires grasp for the data analyst. Furthermore, the computational results of hypothesis test may lead to not unique because that several distribution types can satisfy the conditions of hypothesis test. This situation often confuses the researchers and results in the imprecise judgment which can bring severe aftermath. Although some commercial software, e.g., MatLab and SAS, provide some mimetic functions, but more disk spaces are required meanwhile demanding pre-trainings are also necessary for the data analyzers.

Linear regression method had been put forward by scholar Goodman in the 1950s [20] and then been widely employed in many applications [21, 22]. The EXCEL as a kind of office software has been so broadly used in various fields that a number of researchers and analyzers

[1]Corresponding author. Email: jlqin717@163.com;
Manuscript received January 4, 2014; revised April 6, 2014; accepted June 18, 2014

doi:10.4304/jsw.9.12.2998-3004

can operate it expertly. The linear regression method as a tool had been embedded in EXCEL and the results of data analysis can be easily displayed on the chart by some simple operations like clicking on some correspond command buttons. In this paper, the applications of linear regression method had been developed for the judgment of failure data distribution type. Through the comparison of correlation coefficient in the EXCEL, the most matched distribution type can be quantitatively chosen so that the non-unique above mentioned would be avoid.

This remaining will be organized as follows. Section II presents the principle of linear regression and the property of correlation coefficient. In section III, we will show how to linearize respectively several typical distributions in order to apply the method in EXCEL. In section IV, a number of experiments are conducted in EXCEL to demonstrate the linear regression method and then to compare the results. Finally, the conclusions drawn from this study are given in Section V.

II. Linear Regression Method Description

Linear regression method is often used to describe the linear relationship among some random variables. The method can be depicted as following. Given random variables X and Y, their sample values $(x_i, y_i), i = 1, 2, \ldots, n$ are independent respectively. Then the following formula

$$y_i = a + bx_i + \varepsilon_i \quad \text{(i=1,2,…,n)} \tag{1}$$

can reveal the laws between variables X and Y. In (1), a, b are unknown parameters and named as regression coefficient. The random variable ε_i subjects to s-Normal distribution $N(0, \sigma^2)$ and stands for the errors on y_i. The main object of regression analysis is to calculate the estimation value $\hat{a}, \hat{b}$ of regression coefficient a, b according to the experimental data so that the value of y can be predicted on a given value of x.

Given x, the following formula

$$\hat{y} = \hat{a} + \hat{b}x \quad , \tag{2}$$

can be seen as estimation of $y = a + bx$. Equation (2) is also called as linear regression equation $\mu(x)$ and whose plots are named as regression lines. Once the estimation value $\hat{a}, \hat{b}$ could be found out, we will get the explicit form of (2) which could be used to calculate the value of variable $\hat{y}$ on the given variable x.

The least square method (LSM) is often employed to get the estimation value of regression coefficient. Generally speaking, we have $y_i \neq \hat{y}_i$ because the random variable ε_i is not always equal zero. The approach degrees of theoretical value y and factual value $\hat{y}$ can be represented by the following formula

$$Q(\hat{a}, \hat{b}) = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2 \quad . \tag{3}$$

So we may naturally think that the less value of $Q(\hat{a}, \hat{b})$, the more high of fit degree. Consequently, according to the principle of LSM, we will get the estimation value of regression coefficient when equation (3) reaches its minimum. It can be described by

$$Q(\hat{a}, \hat{b}) = \min Q(a, b) \ . \tag{4}$$

According to the principles of derivative, extreme values can be fetched by resolving the following equations

$$\begin{cases} \dfrac{\partial Q}{\partial a} = -2\sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i) = 0 \\ \dfrac{\partial Q}{\partial b} = -2\sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)x_i = 0 \end{cases} . \tag{5}$$

After the simplification of equation (5) and we will get the following equations

$$\begin{cases} \hat{b} = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ \hat{a} = \bar{y} - \hat{b}\bar{x} \end{cases} , \tag{6}$$

where $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$ .We may introduce the following notations

$$\begin{cases} S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \dfrac{1}{n}(\sum_{i=1}^{n} x_i)^2 \\ S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \dfrac{1}{n}(\sum_{i=1}^{n} y_i)^2 \\ S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \dfrac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) \end{cases} . \tag{7}$$

Then equation (6) can be changed into the following forms

$$\begin{cases} \hat{b} = \dfrac{S_{xy}}{S_{xx}} \\ \hat{a} = \bar{y} - \hat{b}\bar{x} \end{cases} . \tag{8}$$

As a consequence of the above analysis, the estimation values of regression coefficient can be gotten by the method of LSM.

Correlation coefficient is the most general index to measure the degrees of linear relationship among variables that are linearly related. Correlation coefficient is defined by

$$\gamma = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}. \quad (9)$$

γ is a dimensionless statistics and its absolute value is less than or equal 1. When $|\gamma| = 1$, it shows that all the sample data locate a straight line. Then we can come to a conclusion that the variables x, y are of linear relationship at the probability value 1. If $|\gamma| \neq 1$, then the bigger of the value of $|\gamma|$, the better of the linear related degree among the variables; the less of the value of $|\gamma|$, the worse of the linear related degree among the variables. In practical engineering, the square value of γ is often used in order to compute conveniently.

III. Transformation of Typical Distribution

A. Data Preparation

The most important thing needed to do is the variable transformation so that the transformed variables are of linear correlation. In reliability engineering, a lot of failure level and failure data (such as time) are not of linear relationship. On the contrary, there are some nonlinear relationships, e.g., exponential distribution, normal distribution, logarithmic normal distribution and Weibull distribution, between variables failure level and failure time. Consequently we can't directly employ the correlation coefficient to judge which distribution type the law of failure data belong to. After variable transformation the relationships between failure level and failure time can be converted into the linear relationship. The detailed transformation of each distribution type will be treated in the next subsection B. Then correlation coefficient can be computed in EXCEL to judge which distribution type is more matched than the others by the comparison of correlation coefficient. Then the errors came from personal factor can be easily avoided, furthermore the quantification precision to choose the most matched distribution type will be dramatically improved.

The second important matter is how to determine the value of the failure level corresponding the given failure time. According to the random property of failure, failure data can be seen as a random variable T. Order all the failure data by ascendant sequence, e.g., $t_1 \le t_2 \le \cdots t_i \le \cdots \le t_n$. Each failure data t_i has a corresponding failure level $F(t_i)$ which can be described by the approximate median ranks formula [23]

$$F(t_i) = P(T \le t_i) = \frac{\mathrm{i}-0.3}{\mathrm{n}+0.4}, (\mathrm{n} \le 20). \quad (10)$$

Then above (10) is also named as experiential distribution function then the value of failure level can be determined for a given value of failure time t_i. The data points $(t_i, F(t_i)), i = 1, 2, \cdots, n,$ will be properly transformed into $(\sigma(t_i), \xi(F(t_i))), i = 1, 2, \cdots, n,$ so that the latter are of linear relationship. Then using the transformed failure data, correlation coefficient can be calculated and compared quantitatively so that a most matched distribution type will be chosen.

B. Transformation of Typical Distribution Type

As the above mentioned analysis, some data points $(t_i, F(t_i))$ belonging to a certain kind of distribution types aren't of linear relationship and special variable transformation need to be done. Aiming to different distribution type the linearization's method is also respectively different. Then what require do is to analyze different distribution type and find out its respective linearization method. The following will address detailedly the linearization process.

The cumulative distribution function (CDF) of exponential distribution is often depicted as follow

$$F(t) = 1 - e^{-\frac{t-\varphi}{\eta}}, (t \ge \varphi). \quad (11)$$

Parameters φ and η are the unknown and parameter φ is named as location parameter. Obviously, these data points $(t_i, F(t_i))$ aren't of linear relationship. However the linear equation (12) can be obtained by taking logarithm of (11) and simplifying:

$$\ln \frac{1}{1-F(t)} = \frac{1}{\eta} t - \frac{\varphi}{\eta}. \quad (12)$$

For (12), construct converted variables as follow

$$\left.\begin{aligned} y_i &= \ln \frac{1}{1-F(t_i)} \\ x_i &= t_i \end{aligned}\right\}. \quad (13)$$

From (12) and (13), a conclusion that the variables (x_i, y_i) is of linear relationship can be drawn.

Normal distribution's CDF can be written as follows

$$F(t) = \frac{1}{\sqrt{2\pi\sigma}} \int_0^t e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, t \ge 0. \quad (14)$$

The unknown parameter μ represents for expectation and σ for standard deviation. According to the relationship between normal distribution and standard normal distribution, any normal distribution $F(t)$ can be changed into standard normal distribution $\Phi(y)$ whose form is as same as (14) with the parameters $\mu = 0, \sigma = 1$ by the following linear transformation

$$y = \frac{1}{\sigma} t - \frac{\mu}{\sigma}. \quad (15)$$

Then $F(t_i) = \Phi(y_i)$ can be obtained on the base of (15). For each t_i, if supposing $x_i = t_i$ then x_i has a linear relationship with variable y_i based on the (15). The

value of y_i can be gotten by looking for standard normal distribution table.

The distinction between logarithmic normal distribution and normal distribution is whether or not the variable has been taken logarithmic operation. Its CDF can be written down

$$F(t)=\int_0^t \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} dx . \qquad (16)$$

Be similar to the linearization of normal distribution, the $F(t_i)=\Phi(y_i)$ will be finished by the following variable transformation

$$y=\frac{1}{\sigma}\ln t-\frac{\mu}{\sigma} . \qquad (17)$$

From (17) it can be shown that data points (x_i, y_i) are of linear relationship by setting $x_i=\ln t_i$.

Weibull distribution is widely used to model the variability in the fracture properties of ceramics and metals where the concept of the weakest link has been applied. For the two-parameter Weibull distribution, its CDF can be depicted as following [24]

$$F(t)=1-e^{-\left(\frac{t}{\eta}\right)^m} . \qquad (18)$$

In (18), η is the scale parameter and m is the Weibull modulus alternatively referred to as the shape parameter. Taking logarithm of (18) twice, it yields a linear equation

$$\ln\ln\frac{1}{1-F(t)}=m\ln t-m\ln\eta . \qquad (19)$$

Setting

$$\left.\begin{aligned} y_i&=\ln\ln\frac{1}{1-F(t_i)}\\ x_i&=\ln t_i\end{aligned}\right\}, \qquad (20)$$

then the data points (x_i, y_i) are of linear relationship based on the (19).

IV. Application Method In Excel

Supposing there are 20 failure data in the following Table I. They have been arranged from small to large and from left to right in the table I.

TABLE I
FAILURE DATA

7.8	11.3	13.8	15.9	17.4
19.4	20.6	22.3	23.5	24.9
26.6	28.5	29.7	31.2	33.4
34.5	37.0	38.8	42.5	51.4

For each data in the Table I, it has a corresponding failure level according to the (10). Their failure levels have been listed in the following Table II.

TABLE II
FAILURE LEVEL OF EACH FAILURE DATA IN TABLE I

0.034314	0.083333	0.132353	0.181373	0.230392
0.279412	0.328431	0.377451	0.426471	0.47549
0.52451	0.573529	0.622549	0.671569	0.720588
0.769608	0.818627	0.867647	0.916667	0.965686

For exponential distribution, according to the (13) the values of variables x_i, y_i can be calculated by combining the data of Table I and Table II in EXCEL. The results (E and F column) have been shown in the following Fig. 1.

	A	B	C	D	E	F
1	i	ti	F(i)		xi	yi
2	1	7.8	0.034314		7.8	0.034916
3	2	11.3	0.083333		11.3	0.087011
4	3	13.8	0.132353		13.8	0.14197
5	4	15.9	0.181373		15.9	0.200126
6	5	17.4	0.230392		17.4	0.261874
7	6	19.4	0.279412		19.4	0.327687
8	7	20.6	0.328431		20.6	0.398139
9	8	22.3	0.377451		22.3	0.473933
10	9	23.5	0.426471		23.5	0.555946
11	10	24.9	0.47549		24.9	0.645291
12	11	26.6	0.52451		26.6	0.743409
13	12	28.5	0.573529		28.5	0.852212
14	13	29.7	0.622549		29.7	0.974315
15	14	31.2	0.671569		31.2	1.113427
16	15	33.4	0.720588		33.4	1.275069
17	16	34.5	0.769608		34.5	1.467972
18	17	37	0.818627		37	1.707202
19	18	38.8	0.867647		38.8	2.022283
20	19	42.5	0.916667		42.5	2.484907
21	20	51.4	0.965686		51.4	3.37221

Figure 1. Linear transformation for exponential distribution

Complying with the procedures of Charting in EXCEL, scattered diagram can be depicted for the E and F columns. First clicking the command "append trend line" in the chart menu, and then choosing "linear type" furthermore setting the display of formula. The result can be shown in the Fig. 2.

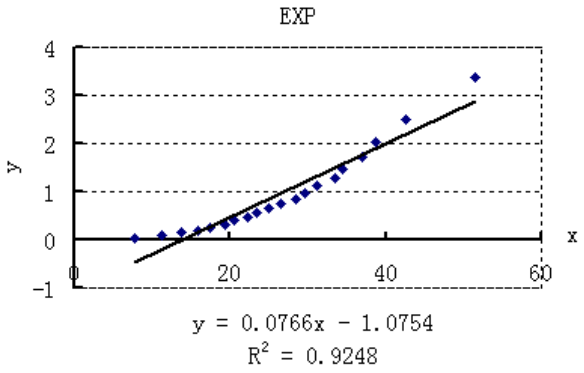


Figure 2. Scattered diagram and trend line for exponential distribution

The parameter R^2 is equal to the square of γ. In other words the R^2 represents the linear correlation degree. At the same time, the linear regression equation is also shown in the Fig. 2.

For normal distribution, the first thing needed to do is to get the value of y_i. Using the $F(t_i)=\Phi(y_i)$, y_i can be gotten by the inverse function of standard normal distribution. In EXCEL the function is NormInv whose syntax format is NORMINV (probability, mean, standard_dev).

The meanings of those parameters are as follows:

probability: the probability value of normal distribution.

mean: the arithmetic average of normal distribution.

standard_dev: standard deviation of normal distribution.

From the (15), we can easily know that the value of x_i is equal to the t_i .The values of x_i (E column) and y_i (F column) have been calculated and shown in the following Fig. 3.

	A	B	C	D	E	F
1	i	ti	F(i)		xi	yi
2	1	7.8	0.034314		7.8	-1.82086
3	2	11.3	0.083333		11.3	-1.38299
4	3	13.8	0.132353		13.8	-1.11534
5	4	15.9	0.181373		15.9	-0.91015
6	5	17.4	0.230392		17.4	-0.73756
7	6	19.4	0.279412		19.4	-0.58459
8	7	20.6	0.328431		20.6	-0.44425
9	8	22.3	0.377451		22.3	-0.31218
10	9	23.5	0.426471		23.5	-0.18537
11	10	24.9	0.47549		24.9	-0.06148
12	11	26.6	0.52451		26.6	0.061476
13	12	28.5	0.573529		28.5	0.185367
14	13	29.7	0.622549		29.7	0.312182
15	14	31.2	0.671569		31.2	0.444249
16	15	33.4	0.720588		33.4	0.58459
17	16	34.5	0.769608		34.5	0.737556
18	17	37	0.818627		37	0.910147
19	18	38.8	0.867647		38.8	1.115337
20	19	42.5	0.916667		42.5	1.382994
21	20	51.4	0.965686		51.4	1.820865

Figure 3. Data after transformation for normal distribution

The scatter diagram and trend line shown in the following Fig. 4 can be depicted in the similar manner.

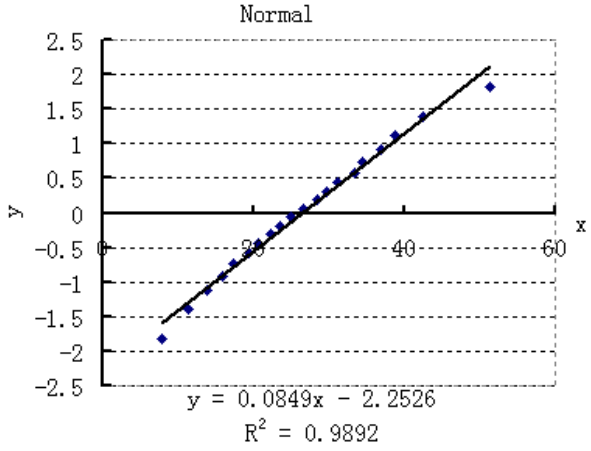


Figure 4. Scattered diagram and trend line for normal distribution

For the logarithmic normal distribution, the analysis method is similar to the normal distribution. The only distinguish lie in the value of x_i is equal to $\ln t_i$ according to the (17). The data after transformation can be also depicted as the following Fig. 5.

	A	B	C	D	E	F
1	i	ti	F(i)		xi	yi
2	1	7.8	0.034314		2.054124	-1.82086
3	2	11.3	0.083333		2.424803	-1.38299
4	3	13.8	0.132353		2.624669	-1.11534
5	4	15.9	0.181373		2.766319	-0.91015
6	5	17.4	0.230392		2.85647	-0.73756
7	6	19.4	0.279412		2.965273	-0.58459
8	7	20.6	0.328431		3.025291	-0.44425
9	8	22.3	0.377451		3.104587	-0.31218
10	9	23.5	0.426471		3.157	-0.18537
11	10	24.9	0.47549		3.214868	-0.06148
12	11	26.6	0.52451		3.280911	0.061476
13	12	28.5	0.573529		3.349904	0.185367
14	13	29.7	0.622549		3.391147	0.312182
15	14	31.2	0.671569		3.440418	0.444249
16	15	33.4	0.720588		3.508556	0.58459
17	16	34.5	0.769608		3.540959	0.737556
18	17	37	0.818627		3.610918	0.910147
19	18	38.8	0.867647		3.65842	1.115337
20	19	42.5	0.916667		3.749504	1.382994
21	20	51.4	0.965686		3.939638	1.820865

Figure 5. Data after transformation for logarithmic normal distribution

The scatter diagram and trend line for logarithmic normal distribution can be depicted by the similar manner in the following Fig. 6.

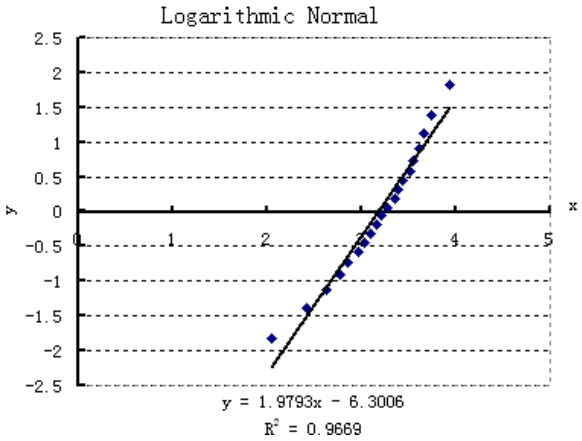


Figure 6. Scattered diagram and trend line for logarithmic normal distribution

For two-parameter Weibull distribution the data can be transformed according to the (20) in the similar method and the values of x_i (E column) and y_i (F column) can be depicted as the following Fig. 7.

	A	B	C	D	E	F
1	i	ti	F(i)		xi	yi
2	1	7.8	0.034314		2.054124	-3.3548
3	2	11.3	0.083333		2.424803	-2.44172
4	3	13.8	0.132353		2.624669	-1.95214
5	4	15.9	0.181373		2.766319	-1.60881
6	5	17.4	0.230392		2.85647	-1.33989
7	6	19.4	0.279412		2.965273	-1.1157
8	7	20.6	0.328431		3.025291	-0.92095
9	8	22.3	0.377451		3.104587	-0.74669
10	9	23.5	0.426471		3.157	-0.58708
11	10	24.9	0.47549		3.214868	-0.43805
12	11	26.6	0.52451		3.280911	-0.29651
13	12	28.5	0.573529		3.349904	-0.15992
14	13	29.7	0.622549		3.391147	-0.02602
15	14	31.2	0.671569		3.440418	0.107443
16	15	33.4	0.720588		3.508556	0.243
17	16	34.5	0.769608		3.540959	0.383882
18	17	37	0.818627		3.610918	0.534856
19	18	38.8	0.867647		3.65842	0.704227
20	19	42.5	0.916667		3.749504	0.910235
21	20	51.4	0.965686		3.939638	1.215568

Figure 7. Data after transformation for two-parameter Weibull distribution

The scatter diagram and trend line can be shown in the following Fig. 8 by the similar method.

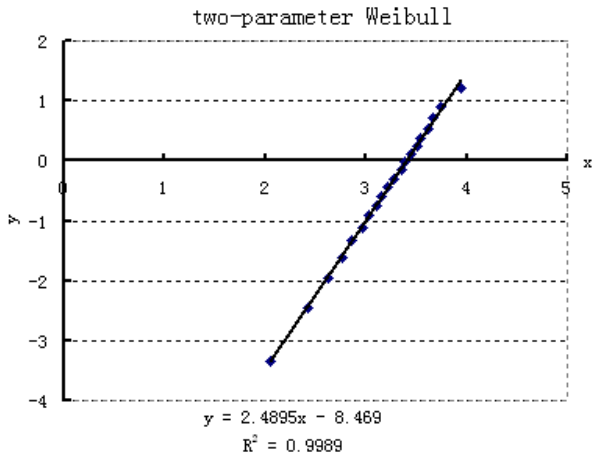


Figure 8. Scattered diagram and trend line for two-parameter Weibull distribution

For the sake of convenience, the above computational result of the data analysis can be listed in the following Table III.

TABLE II
COMPARISON TABLE OF LINEAR REGRESSION RESULT

Distribution Type	Linear Regression Equation	Correlation Coefficient (R^2)
Exponential	y = 0.0766x - 1.0754	0.9248
Normal	y = 0.0849x - 2.2526	0.9892
Logarithmic Normal	y = 1.9793x - 6.3006	0.9669
Two-parameter Weibull	y = 2.4895x - 8.469	0.9989

From the comparison table, we can draw a conclusion that the most matched distribution of failure data is subjected to the two-parameter Weibull distribution because the correlation coefficient is the highest. At the same time, the linear regression equation had been found out which can be used to estimate the related parameters of the two-parameter Weibull distribution according to the (19) and (20). Because the correlation coefficient is computed quantitatively, the personal observation error will be avoided efficiently. The failure data can be treated as the two-parameter Weibull distribution for prediction or some other intents.

V. CONCLUSOINS

In this paper, the linear regression principle had been put forward as a type of data analysis method for the distribution type test. The method can be employed for the distribution type inference of failure data. For each general distribution, the linear regression method can't be implemented directly. First linearization needs to do and the detailed procedures are also respective different. The correlation coefficient had been examined for the comparison of several probable distribution types.

The detailed steps of the application for the linear regression in EXCEL have been illustrated. A group of data has been used for the distribution type test by the comparison of the correlation coefficient. The result has shown that the linear regression method is validated and convenient for the analyst. Furthermore, the method is also adapted for developing related software.

ACKNOWLEDGMENT

The authors wish to thank the reviewers and the editor for their constructive comments that have helped to improve this article. This work was supported in part by a grant from Key Program of National Nature Science Foundation of China (51036002), National Basic Research Program of China (973 Program) (2012CB215203) and Hebei Province Natural Science Foundation (F2014502081).

REFERENCES

[1] J. Han and M. Song, "Efficiency evaluation information system based on data envelopment analysis," *Journal of Computers*, vol. 6, pp. 1857-1861, 2011.

[2] L. Lan, X. Gou, Y. Xie, and M. Wu, "Intelligent GSM cell coverage analysis system based on GIS," *Journal of Computers*, vol. 6, pp. 897-904, 2011.

[3] Z. Liu and C. Gao, "A hybrid optimization algorithm to evaluate the CCWPE based on DEA sampled by FCE," *Journal of Computers (Finland)*, vol. 7, pp. 2836-2841, 2012.

[4] J. Sun, H. Liao, and B. R. Upadhyaya, "A Robust Functional-Data-Analysis Method for Data Recovery in Multichannel Sensor Systems," *Cybernetics, IEEE Transactions on*, vol. PP, pp. 1-1, 2013.

[5] D. D. Nauck, M. Spott, and B. Azvine, "Fuzzy methods for automated intelligent data analysis," in *Fuzzy Systems, 2004. Proceedings. 2004 IEEE International Conference on*, 2004, pp. 487-492 vol.1.

[6] L. Guoguang, H. Yun, and E. Yunfei, "Aging data analysis methods based on short-term aging test," in *Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR2MSE), 2012 International Conference on*, 2012, pp. 905-908.

[7] T. L. Olsen and S. M. Hock, "Data analysis method for wind turbine dynamic response testing," in *Energy Conversion Engineering Conference, 1989. IECEC-89., Proceedings of the 24th Intersociety*, 1989, pp. 2035-2040 vol.4.

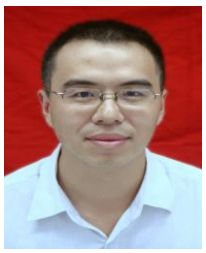
[8] M. Kovac, E. Stefanakos, and T. Arbogast, "AFV-soft: advanced data analysis software for electric and other alternative fuel vehicles," in *Southeastcon '96. Bringing Together Education, Science and Technology., Proceedings of the IEEE*, 1996, pp. 58-61.

[9] J. Youyong, "Simulation research and contrastive analysis of the volt-ampere characteristics of resistor and diode based on Mulitisim and Excel," in *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*, 2010, pp. V4-232-V4-235.

[10] H.-b. Shi, "Applying Excel VBA to Implement Comparison among Physical Experiment Data," in *Management and Service Science (MASS), 2011 International Conference on*, 2011, pp. 1-3.

[11] J. Kontio, "A case study in applying a systematic method for COTS selection," in *Software Engineering, 1996., Proceedings of the 18th International Conference on*, 1996, pp. 201-209.

[12] G. Bontempi and W. Kruijtzer, "A data analysis method for software performance prediction," in *Design, Automation and Test in Europe Conference and Exhibition, 2002. Proceedings*, 2002, pp. 971-976.
[13] E. Reschenhofer, "Generalization of the Kolmogorov-Smirnov test," *Computational Statistics and Data Analysis*, vol. 24, pp. 433-441, 1997.
[14] F. A. Andrade, I. Esat, and M. N. M. Badi, "A new approach to time-domain vibration condition monitoring: Gear tooth fatigue crack detection and identification by the Kolmogorov-Smirnov test," *Journal of Sound and Vibration*, vol. 240, pp. 909-919, 2001.
[15] Z. Drezner, O. Turel, and D. Zerom, "A modified kolmogorov-smirnov test for normality," *Communications in Statistics: Simulation and Computation*, vol. 39, pp. 693-704, 2010.
[16] T. Hauschild and M. Jentschel, "Comparison of maximum likelihood estimation and chi-square statistics applied to counting experiments," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 457, pp. 384-401, 2001.
[17] Y.-T. Chen and M. C. Chen, "Using chi-square statistics to measure similarities for text categorization," *Expert Systems with Applications*, vol. 38, pp. 3085-3090, 2011.
[18] A. W. Grace and I. A. Wood, "Approximating the tail of the Anderson–Darling distribution," *Computational Statistics & Data Analysis*, vol. 56, pp. 4301-4311, 2012.
[19] S. N. Chiu and K. I. Liu, "Generalized Cramé–von Mises goodness-of-fit tests for multivariate distributions," *Computational Statistics & Data Analysis*, vol. 53, pp. 3817-3834, 2009.
[20] T. P. Goodman, "Technique for Approximate Measurement of Correlation Coefficients," *Journal of Applied Physics*, vol. 27, pp. 773-775, 1956.
[21] C. Jen-Tzung, "Quasi-Bayes linear regression for sequential learning of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 268-278, 2002.
[22] C. Jen-Tzung, "Linear regression based Bayesian predictive classification for speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, pp. 70-79, 2003.
[23] M. Cacciari, G. Mazzanti, and G. C. Montanari, "Comparison of maximum likelihood unbiasing methods for the estimation of the Weibull parameters," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 3, pp. 18-27, 1996.
[24] M. T. l. David Hudak, "On estimating percentiles of the Weibull distribution by the linear regression method," *Journal of Materials Science* vol. 44, pp. 1959-1964, 2009 February.



Jinlei Qin received his B. S. degree and M. S. in department of computer science and technology at north china electric power university. He is now a Ph. D. graduate student in the school of control and computer. His research interests include condition-based maintenance, system reliability analysis and life data analysis.

Yuguang Niu obtained his Ph. D. degree in the department of thermal energy and power engineering at north china electric power university. He is a professor and doctoral tutor in the school of control and computer. He works as a deputy director of the state key laboratory of alternate electric power system with renewable energy sources. His research interests include modeling and control of industry process, fault detection and diagnosis and system reliability. He has participated and completed several research projects of national level and ministry level and obtained recognition of peers and experts.

Zheng Li received her B. S. degree and M. S. in department of computer science and technology at north china electric power university. She is now a Ph. D. graduate student in the school of control and computer. Her research interests include the research and application of intelligence algorithms in electric power system.