Indian Journal of Science and Technology, *Vol 10(3), DOI: 10.17485/ijst/2017/v10i3/110616, January 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Modification of Zipf-Mandelbrot Law for Text Analysis using Linear Regression

S. Lakshmi Sridevi* and R. Devanathan

Hindustan Institute of Technology and Science, Chennai – 603103, Tamil Nadu, India;
lakshmi@hindustanuniv.ac.in

Abstract

Background: The application of Zipf's law is ubiquitous in linguistics and other fields. Mandelbrot proposed a modification of the law called Zipf-Mandelbrot law (ZM). An enhanced form of ZM law has been proposed. **Methods:** In this paper, we approximate the logarithmic form of ZM law into a linear regression form of arbitrary order of the inverse of the Zipfi an rank of words in a text. The maximum likelihood solution of the regression model is given in closed form. This is in contrast to the complex search for the optimum solution of the enhanced ZM models. **Findings:** The performance of the proposed model is shown to compare favorably with that of ZM law as well as other existing models using Chi-Square goodness of fit test. **Improvements:** The present work addresses mainly the lower ranks, so we propose to extend the work to higher order ranks using LNRE model in the future.

Keywords: Goodness of Fit, Linear Regression, Quantitative Linguistics, Zipf-Mandelbrot Law

1. Introduction

One of the most puzzling phenomenon in quantitative linguistics is Zipf's law[1-3]. Zipf, a pioneer in quantitative linguistics, studied frequency distributions of words in text. Zipf ranked the words in a text in decreasing order of frequency. He demonstrated an inverse relationship between frequency and rank. Let fr(r,N) denote the frequency of samples, in N tokens, of a word with Zipf rank, r(1,2,.....n) where n is the maximum rank considered. Zipf formulated the following relationship.

$$f_r(r,N) = \frac{C}{r^{\alpha}} \quad (1)$$

where □ is a free parameter, C is a normalizing constant and

$$N = \sum_{r=1}^{n} f_r(r,N).$$

(1) is known as Zipf's law. Based on empirical data, Zipf's law has been generalized to

$$r^{B} f = C \quad (2)$$

Where B □ 0.92. At the lower ranks, (2) are found to deviate from empirical data. To rectify this, Mandelbrot developed a further refinement of Zipf's law

$$(r + m)^{B} f = C \quad (3)$$

Addition of parameter m > 0 is useful for introducing a downward curvature at the lower ranks observed in empirical rank- frequency data. However, for higher ranks, a small positive value of m does not affect the frequency value[4,5]. (3) is called Zipf-Mandelbrot (ZM) law. Montemurro[6] argues that the written text actually has statistical properties that fall beyond the scope of (3). He demonstrates that corpora rank-frequency behavior at very high ranks does not follow ZM law. So he proposed a second power law regime for higher ranks. The phenomenon exhibited in[6] can be identified with the concept of LNRE distributions first introduced by Khmaladze[7]. LNRE stands of Large Number of Rare Events corresponding to large number of extremely low probability words that are present in lexical frequency distributions. The problem of LNRE distributions has been termed Empirical Structural Type Distributions which is the

**Author for correspondence*

inverse of Zipfian rank-frequency distribution. Zipf –Mandelbrot law is well formulated for random character sequences. It assumes an infinite vocabulary which is unrealistic in the realm of natural language corpora .In order to achieve better approximation of such frequency distribution, a finite Zipf-Mandelbrot law (fZM) has been introduced by Evert[8].Evert has evaluated fZM against ZM model and has shown that fZM outperforms the other models with the additional benefit of fast and robust numerical computation .The implementation of ZM for different applications has been discussed in[9,10]. Baayen[11] highlights two characteristics of lexical statistics as follows: 1. When dealing with words and their frequencies the usual law of large numbers may not apply. This is demonstrated, for example, by continuous increase of vocabulary size of corpus as a function of number of words in the corpus. 2. Authors of lexical text tend to use words reflecting lexical cohesion both at the level of sentence and discourse and hence it cannot be considered random. Baayen[11] shows that by randomizing a given text, for example, from Alice in Wonderland, the probability that a random pattern arises by chance is less than 0.05. In this paper, a modification of ZM model is proposed based on linear regression. Our approach is distinct from the modification of Zipf -Mandelbrot law proposed by Montemurro[6]. Montemurro proposes to model the complete corpus using a word frequency distribution that is characterized by multiple regions and with the ZM law characterizing the lower rank region. While Montemurro[6] takes an analytical approach to characterize the rank- frequency curve, in this paper, we propose a regressive approach. Also, while Montemurro model is quite complex requiring a search for optimum parameters with no apparent guarantee of convergence and global optimization, our approach, in contrast, is based on the common linear regression which has a maximum likelihood estimate solution in closed form .We begin with the Zipf-Mandelbrot law 3. and show that it can be modified into a linear regressive form of arbitrary order. In effect, our proposed linear regression model can be considered a generalization of Zipf-Mandelbrot law. The main contribution of the paper is the identification of linear regression formula as a generalization of Zipf-Mandelbrot law, provision of regression model for the rank-frequency distribution and demonstration of approximate solution of finite order to the rank- frequency curve using the maximum likelihood solution of linear regression. We illustrate the effectiveness of the model with corpora texts showing the goodness of fit of the proposed model. Also, the proposed model is compared with application of existing Zipf, ZM, Carroll and Sichel models. The rest of the paper is organized as follows: - The following section describes the development of the proposed model and the maximum likelihood solution of the model. Section 3 illustrates the use of the model to characterize the rank-frequency of a given text exhibiting the goodness of fit. We also evaluate the proposed model with an existing ZM application. The paper is concluded with summary and directions for further research.

2. Model Development

Taking natural logarithm of (3)

$$B\ln(r+m)+\ln(f)=\ln(C) \tag{4}$$

One can write (4) as

$$B\left[\ln(r)+\ln\left(1+\frac{m}{r}\right)\right]+In(f)=In(C) \tag{5}$$

We now state the following proposition.

Proposition 1: Equation (5) can be approximated to the following form

$$B\left[Q_0+\sum_{i=1}^{p}Q_i\left(\frac{1}{r^i}\right)\right]+In(f)=In(C) \tag{6}$$

where

$$Q_o=\sum_{n=1}^{p}\frac{1}{n}$$

$$Q_i=(-1)^i\left[\left(\frac{1-m^i}{i}\right)+\sum_{k=1}^{p-i}\left(\frac{1}{i+k}\right)(i+k)_{C_i}\right]$$

$$i=1,2,3.....p$$

and p is a finite integer.

Proof : Expanding ln(r) with r >1/2 and $\ln(1+\frac{m}{r})$, m/r > 0 [12,13] , one can write (5) as

$$\ln(r)+\ln\left(1+\frac{m}{r}\right)=\sum_{n=1}^{\infty}\frac{1}{n}\left[\left(\frac{r-1}{r}\right)^n+(-1)^{n+1}\left(\frac{m}{r}\right)^n\right] \tag{7}$$

Approximating upto p terms only

$$\ln(r)+\ln\left(1+\frac{m}{r}\right)\approx\sum_{j=1}^{p}\frac{1}{j}\left[\left(1-\frac{1}{r}\right)^j+(-1)^{j+1}\left(\frac{m}{r}\right)^j\right] \tag{8}$$

Using binomial expansion and collecting terms of the same degree one can write

$$\ln(r)+\ln\left(1+\frac{m}{r}\right)=Q_0+\sum_{i=1}^{p}Q_i\left(\frac{1}{r^i}\right) \quad (9)$$

where

$$Q_0=\sum_{n=1}^{p}\frac{1}{n}$$

and

$$Q_i=(-1)^i\left[\left(\frac{1-m^i}{i}\right)+\sum_{k=1}^{p-i}\frac{1}{i+k}(i+k)_{C_i}\right], i=1,2.....p$$

Substituting (9) into (5),(6) follows. Hence the result. Equation (6) can now be put in the form

$$\left(\ln C-BQ_o\right)-\left[B\sum_{i=1}^{p}Q_i\left(\frac{1}{r^i}\right)\right]=\ln f$$

Or,

$$\delta_0+\left(\sum_{i=1}^{p}\delta_i\rho^i\right)=\ln f \quad (10)$$

where

$$\delta_0=\left(\ln C-BQ_o\right)$$

δ_i = -BQ_i, i= 1,2,……..p

Generalizing (10) into a regressive formula, we can write

$$Y=X\beta+\varepsilon_0 \quad (11)$$

where

$$Y=\left[\ln f_1,\ln f_2.....\ln f_i.....\ln f_n\right]^t$$

$$\beta=\left[\delta_1,\delta_2,.....\delta_j.....\delta_p,\delta_0\right]^t$$

t stands for transpose,

X = [$x_{i,j}$] ; i=1,2,3,……n, j= 1,2,3,……..p+1

$x_{i,j}=\frac{1}{i^j}$; i = 1,2,3,……..n ; j = 1,2,……..p

and

$x_{i,\,p+1}$ = 1, $\forall$ i=1,2,3…..n

$\varepsilon_0 \approx N_n(0,\sigma_n)$ corresponds to a noise term assumed to be a multivariate normal i.i.d distribution of n variables with zero mean and variance σ_n.

Maximum likelihood solution of β in (11) is given as

$$\hat{\beta}=\left[(X^tX)^{-1}X\right]Y \quad (13)$$

3. Evaluation of Proposed Model

The proposed model is evaluated using two text corpora written by Conrad[14] and Desai[15].The two corpora are of different genre in that the first one is a novel while the second one is an autobiography. For the corpora texts 1 and 2, a detailed simulation of the model is carried out starting from second order to eighth order, using Python programming and Microsoft Excel. Using the count function of Python, the frequency- rank data is generated. The data is truncated to the first fifty ranks only. The model order used is from second to eight. Various runs were made to fit the data. For each model order, we calculate the mean square error between the observed data and the model output. Figure 1(a) and Figure 1(b) give the plot of the rank-frequency data for corpora texts 1 and

Table 1. Root mean square error of proposed model applied to corpora texts 1 and 2.

Model order		Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth
Root Mean Square	Corpora Text 1	2325	635	208	141	116	113	111
	Corpora Text 2	1915	635	269	110	103	110	118

Table 2. Chi-Square statistic for proposed model applied to corpora texts 1 and 2

Corpora	Model Order	No.of ranks	Degrees of freedom	Chi-Square Statistics (CV)	Cumulative probability $P(\chi^2 \le CV)$
Text 1	Eighth	50	49	23	0.0006
Text 2	Seventh	50	49	32	0.03

2 respectively together with the fitted model output of order eight for corpora text 1 and order seven for corpora text 2 respectively. Root Mean square error values of all order models of corpora texts 1 and 2 are given in Table1. Table 2 gives the result of Chi- Square goodness of fit test[16] for both the corpora texts. Clearly the evaluation of goodness of fit using Chi-Square test for the both text corpora[14,15] is satisfactory as can be seen from the last column of Table 2.

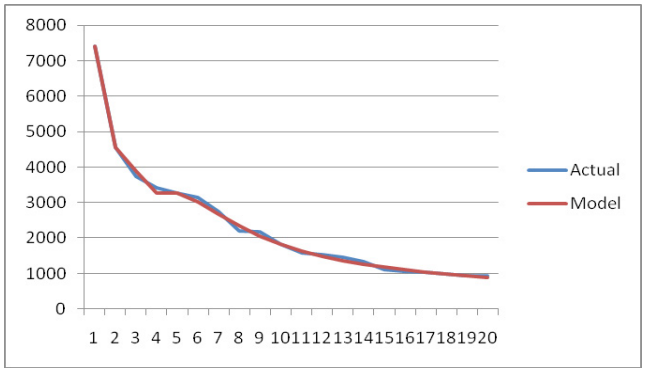

Figure 1. (a) Plot of word frequency vs. rank for proposed model and actual data (Corpora Text1)

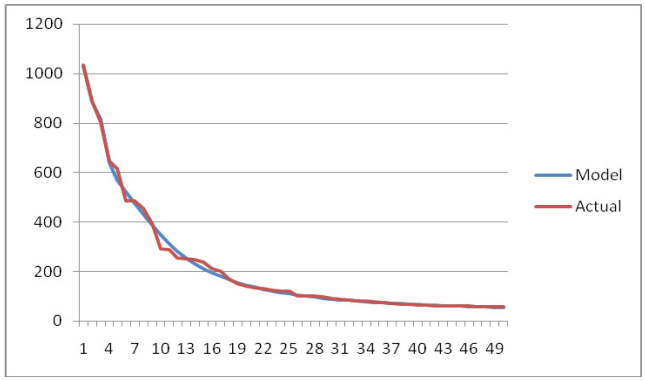

Figure 1. (b) Plot of word frequency vs Rank for model and actual data (Corpora Text 2).

The efficiency of proposed algorithm is compared with that of implementation of ZM. Figure 2(a) gives a plot of Bulgarian text data segment given in[9] together with ZM model fit. Figure 2 (b) gives the same empirical data together with the fit of the proposed regression model. Comparing Figures 2(a) and 2(b), it is clear that for the data considered the proposed regression model represents the data much more accurately than the ZM model[9]. This conclusion is again confirmed by the data given in Table 3 which gives the Chi-Square goodness of fit test for the ZM and the proposed regression model. Further our proposed regression model is compared with the implementation of Carroll, Sichel and Zipf against frequency spectrum data given in Baayen[11]. The results of Chi-Square test used for the comparison is given in Table 4. Clearly again the proposed regression model based on ZM law shows a better goodness of fit result compared to the fit using Carroll ,Sichel and Zipf models[4].

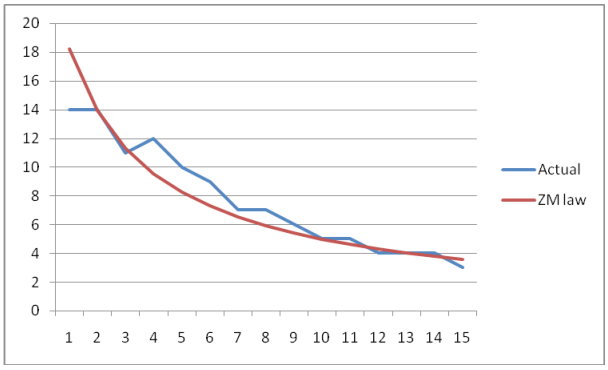

Figure 2. (a) Plot of word frequency vs. rank for model and actual data using ZM model.

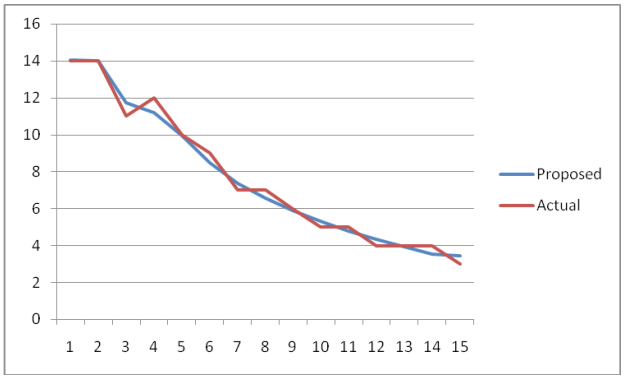

Figure 2. (b) Plot of word frequency vs. Rank for Model and Actual data using proposed model.

Table 3. Comparative evaluation of proposed model and ZM model

Model	No.of ranks	Degrees of freedom	Chi-Square Critical value(CV)	Cumulative probability $P(\chi^2 \leq CV$
ZM law	15	14	2.83145	0.0007
Proposed Regression Model (eighth order)	15	14	0.354258	0

Table 4. Evaluation of proposed model using Chi-Square Statistic with that of Sichel, Carroll and Zipf models.

Model	Chi-Square Statistics	Cummilative Probability
Proposed Regression Model	5.469781	0.02
Sichel	36.3309	0.9992
Carrol	8.110846	0.12
Zipf	8.223035	0.12

4. Conclusion

This paper has proposed a modification of Zipf-Mandelbrot (ZM) law in the form of a regression model. While the Zipf law identifies the frequency to a single power of the inverse of the rank, the proposed model assumes the frequency to be a polynomial of arbitrary order of the inverse of the rank. The proposed model can be seen to be a generalization of ZM model. The proposed model is shown to fit well to the data of text of different genre. Also, the proposed regression model has been evaluated with the implementation of the existing models such as Zipf, ZM, Sichel and Carroll. For the data given it is seen that the proposed regression model performs better than other models stated. The advantage of the proposed model is that it has a well known maximum likelihood solution in closed form. This is in contrast to the modification of ZM model proposed in[6,8] which involves complex analytical expression with attendant nonlinear search for the solution of optimum parameters with no apparent guarantee of convergence and global solution. One could cite that the proposed model in this paper needs to invert a matrix for the solution as seen from (12) possibly causing instability of the solution. This could happen for very large model order. Well known numerical solutions exist for solving linear system of equations of large order. Further, it is well known that the large number of low frequency words may be a matter of concern in word frequency modeling. We propose to apply our approach to LNRE model in our future work which addresses such concerns

5. References

1. Zipf GK. The Psycho–Biology of Language. Boston: Houghton Mifflin; 1935.
2. Zipf GK. Human behaviour and the principle of the least effort. A Introduction to Human Ecology. New York: Hafner; 1949.
3. Wyllys Ronald E. Empirical and theoretical bases of Zipf's law. Library Trends. 1981; 30(1):53-64.
4. Mandelbrot B. An information theory of statistical structure of language. In: Jackson WE, editor. Communication Theory. New York: Academic Press; 1953. P. 503-12.
5. Mandelbrot B. On the theory of word frequencies and on related Markovian models of discourse. In: Jakobson R. editor. Structure of language and its Mathematical Aspects, American Mathematical Society, Providence Rhode Island. 1962. p. 190-219.
6. Montemurro, Marcelo A. Beyond the Zipf–Mandelbrot law in quantitative linguistics. Physica A: Statistical Mechanics and its Applications. 2001; 300(3):567-78.
7. Khmaladze EV. The statistical Analysis of large number of rare events. Technical report MS-R8804,Dept of Mathematical Statistics, CWI. Amsterdam:Center of Mathematics and Compute Science(1987).
8. Evert S. A simple LNRE model for random character sequences. Proceedings of JADT. 2004; 2004.
9. Ioan-Iovi P. Word frequency studies. Walter de Gruyter. 2009; 64.
10. Riyal Manoj Kumar, et al. Rank-frequency analysis of characters in Garhwali text: Emergence of Zipf law. Current Science. 2016; 110(3): 429-34.
11. Baayen R. Harald. Word frequency distributions. Science & Business Media. 2001; 18.
12. Available from: www.math2.org/math/expansion/log.htm
13. Available from: www.wikepedia.org/wiki/Naturallograthim. 2016.
14. Conrad J. Heart of darkness. Black Wood Magazine. 3rd ed. 1899.
15. Mohandass Karamchand Gandhi. My Experiments with Truth. 5th ed. Sublime Books; 1921.
16. Available from: http://stattrek.com/chi-square-test/goodness-of-fit.asp