International Journal of Engineering & Technology, 7 (3) (2018) 1558-1560

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i3.11653
Research paper

Regression model of rank –frequency data of Tamil text

S. Lakshmisridevi [1] *, R. Devanathan [2]

[1] AP , Hindustan Institute of Technology and Science
[2] Professor Emeritus , Hindustan Institute of Technology and Science
**Corresponding author E-mail: lakshmi@hindustanuniv.ac.in*

Abstract

The application of Zipf's law is universal not only in linguistics but also in various other areas. Mandelbrot modified Zipf law as Zipf Mandelbrot law and it is further we proposed a modification of the ZM law for modeling rank frequency- data of linguistic text. Our model generalized ZM law into a linear regression model involving arbitrary order of Zipfian rank of words in a text .The performance of the proposed model is studied for an English text and it shown to compare favorably with that of Z-M law using Chi-Square goodness of fit test. In this paper we have applied to Tamil text and its performance is also up to the mark and it is been proved by the Chi-Square test and it addresses mainly the lower ranks, we propose to extend the work to higher order ranks using LNRE model in the future.

Keywords: *ZIPF Law, ZM Law, Chi-Square Text, Goodness of Fit.*

1. Introduction

One of the classical results of quantitative linguistics is Zipf's law [1], [2], who studied the frequency distribution of words in the text. Zipf arranged the words in the order of decreasing frequency. He showed that their exist inverse relationship of rank and frequency of a word in a given text. Let f_z (z, N) denote the frequency of word sample , where z denotes the rank and N corresponds to number of word tokens .According Zipf law.

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

Where α is a free parameter, C is a normalizing constant and

$$N=\sum_{i=1}^{n} f_z(z,N)$$

(1) is known as Zipf's law. However empirical data of English text show that the frequency of lower ranks tended to be smaller than those predicted by Zipf law . To account for this Mandelbrot[3, 4] introduced Modification of Zipf's law as

$$(z+m)^{\alpha} f = C \quad (2)$$

Where m is parameter, m>0
The effect of m is to reduce the effect of frequency at lower ranks as required by the empirical data. The parameter m does not affect the frequency of higher rank words.Montemurro[5] further generalizes Zipf –Mandelbrot law to account for different regions of variations of frequency with rank. .According to Montemurro Zipf-Mandelbrot law occupied only the lower ranks of his multiple region model.one of the difficulty in lexical statistics is that there are large number of very small frequency words which tend to elongate the tail of Zipf frequency rank curve .Since these small low frequency words are quite a significant part of tokens in the text, another approach to the modeling of text is taken in the form of Large Number of Rare Events (LNRE) models.Here the modeling is done based on the frequency spectrum , where the frequency spectrum is number of words having same number of occurrences in the text.
Baayen argues that lexical statistics is different from the common notations of probability in that theory of vocabulary of text corresponding to number of different word occurrences in text keep increasing as the corpus size increases. This is unlike the traditional probability that the probability tends to converge as the sample size increases

2. Development of the model

$$B\ [\ln (z) + \ln (1+\frac{m}{z})] + \ln (f) = \ln(C) \quad (3)$$

We now state the following proposition taking the natural logarithm of (2)

$$B\ln(z+m)+\ln f=\ln C \quad (4)$$

One can write (3) as

$$B\ln(z+m)+\ln f=\ln C$$

One can write (3) as
We now state the following Proposition 1:
Equation (4) can be written in the form

$$B[Q_0+\sum_{i=1}^{p} Q_i(\frac{1}{z_i})]+\ln f=\ln C \quad (5)$$

Where

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

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

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

Proof: See Appendix (5) can now be put in the form

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

Or

$$q_0 + \left(\sum_{i=1}^{p} q_i \rho^i \right) = \log f \qquad (6)$$

Where

$$q_0 = (\log C - BQ_o)$$

$$q_i = -BQ_i \text{ ;i= 1,2,..........p}$$

Generalizing (6) into a regressive formula, we can write

$$Y = X\gamma + \varepsilon_0$$

Where $Y = [\ln f , \ln f2,\ldots\ldots \ \ln fi,\ldots\ldots\ldots\ldots\ldots \ln fn]^t$

$$\gamma = [q_1, q_2, \ldots\ldots\ldots\ldots\ldots q_p, q_0]$$

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

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

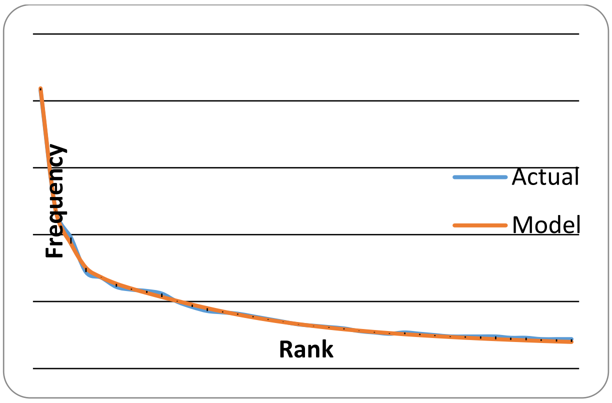
$x_{i\ p+1} = 1, \forall$ i

$\varepsilon_0 \approx N_n(0, \sigma_n, z_n)$ corresponds to the 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 for (7) is given as

$$\hat{\gamma} = [(X^t X)^{-1} X] Y$$

3. Implementation of modified ZIPF Mandelbrot law in Tamil text.

The proposed model is verified for its efficiency using the text corpora , the famous tamil novel ponniyinselvan written by Kalki [9]. The detailed simulation of the model is carried out starting from second order to eighth order , using Python programming and Microsoft Excel for the text corpora. Using the count function, the data is generated on the frequency rank till 24 ranks . The model order used is from four to eight . Various runs were made to fit the data .For each model order we calculate mean square error between the observed data and model output for different model orders Fig.1 gives the plot of the data together with fitted model output of order eight. Root Mean square error result of the all order models are given inTable.1.



Model output vs Frequency data of Corpora text 1

Rank	Model	Actual
1.	0.120554	0.120557
2.	0.074157	0.074002
3.	0.05973	0.060678
4.	0.056907	0.055628
5.	0.053757	0.053071
6.	0.048997	0.050969
7.	0.043696	0.044682
8.	0.038647	0.035674
9.	0.034182	0.035217
10.	0.030369	0.029549
11.	0.027159	0.025656
12.	0.024467	0.024727
13.	0.022207	0.023717
14.	0.020301	0.021567
15.	0.018684	0.017788
16.	0.017302	0.017234
17.	0.016114	0.016794
18.	0.015086	0.015866
19.	0.01419	0.015182
20.	0.013404	0.015084
21.	0.01271	0.012624
22.	0.012096	0.011484
23.	0.011548	0.010979
24.	0.011057	0.010311

Table 1: Mean Square Value of Error

Regression Model order	Third	Fourth	Fifth	Sixth	Seventh	Eighth
Mean Square	4.5E-05	2.0E-06	6.6E-07	6.6E-07	6.5E-07	6.7E-07

Table 2: Results of Chi-Square Test of Regression Model Order Eight.

Model	No. of ranks	Degrees of freedom	Chi-Square Critical value(CV)	Cumulative probability P($\chi^2 \leq$ CV)
Eight	24	23	1.98912E-05	0

4. Conclusion

This paper has given a applied a regression model of ZM While the Zipf law identifies the model used assumes the frequency to be a polynomial of arbitrary order of the inverse of the rank.The advantage of the model used is that it has a well known maximum likelihood solution in closed form. The model used models the tamil text in a very good manner and it can be witnessed by results of Chi-Square test.

References

[1] Zipf. G. K, the Psycho –Biology of Language, Houghton Mifflin, Boston (1935).

[2] Zipf. G. K, Human Behaviour and the Principle of the Least Effort. A introduction to human Ecology, Hafner, New York. (1949 [3] Wyllys, Ronald E. "Empirical and theoretical bases of Zipf's law." *Library Trends* 30.1 53-64(1981).

[3] Mandelbrot, B An information theory of Statistical Structure of language, in W. E. Jackson (e. d.), Communication theory, Academic press, New York (1953), pp 503-512.

[4] Mandelbrot, B On the theory of word frequencies and on related Markovian models of Discourse, in R. Jakobson (ed.),Structure of language and its Mathematical Aspects ,American Mathematical Society ,Providence Rhode Island(1962) ,pp.190-219.

[5] Montemurro, Marcelo A. "Beyond the Zipf–Mandelbrot law in quantitative linguistics." Physica A: Statistical Mechanics and its Applications 300. 3 (2001): 567-578. https://doi.org/10.1016/S0378-4371(01)00355-7.

[6] Khmaladze, E. V.: The statistical Analysis of large number of rare events, Technical report MS-R8804, Dept. of Mathematical Statistics, CWI. Amsterdam: Center of Mathematics and Computer Science (1987).

[7] Evert, Stefan. "A simple LNRE model for random character sequences." *Proceedings of JADT*. Vol. 2004. (2004).

[8] Popescu, Ioan-Ioviț. *Word frequency studies*. Vol. 64. Walter de Gruyter, 2009.

[9] https://book.ponniyinselvan.in/.