IJCTA, 10(07), 2017, pp. 261-265

Linear Regression Model of Frequency Spectrum of Text

S. Lakshmisridevi* and R. Devanathan*

Abstract : One of the important characteristics of lexical statistics is the Large Number of Rare Events (LNRE) phenomenon.Various analytical models have been proposed to estimate frequency spectrum of LNRE distribution . In this paper, we approximate the logarithm of Zipf-Mandelbrot law for the frequency spectrum by a polynomial of arbitrary order of the index of the frequency class. Based on the approximation, a linear regression model is derived for the LNRE distribution. A maximum lilelihood solution in ,closed form, is provided for the proposed model. We have evaluated the proposed model against the existing Carroll, Sichel and Zipf models using frequency spectrum data taken from Alice in Wonderland.The results of proposed model compare favourably with those of the existing models.
Keywords : LNRE, frequency spectrum, Analytical linguitstics, Linear regression, Zipf-Mandelbrot law

1. INTRODUCTION

Frequency spectrum of a text corresponds to the number of word types with frequency m in a sample of N word tokens. For example, V(1, N) denotes the number of word types that occur only once in a text while V(*m*, N) corresponds to number of word types occurring exactly m times in the text .

Zipf [1] has formulated the rank frequency model as

$$f_z(z, \mathrm{N}) = \frac{\mathrm{C}}{z^{\alpha}},$$

where $f_z(z, \mathrm{N})$ is the frequency of a word of rank z in a text of N word tokens, α is a free parameter, ($a \approx 1$) and C is a normalizing constant. The words are ranked according to the frequency of occurrence in the text with rank 1 corresponding to the highest frequency and rank 2 the next highest and so on .

Considering two ranks z_1 and z_2 , such that

$$f(z_1, \mathrm{N}) = m + 1$$
$$f(z_2, \mathrm{N}) = m,$$
$$m > 0$$

it follows that

$$\mathrm{V}(m, \mathrm{N}) = (z_2 - z_1)$$
$$= \frac{\mathrm{C}}{m(m+1)} \quad (1)$$

V(*m*, N) is normally a non- increasing function in discrete form. But V(*m*, N) could be zero for some value of m and also that there could be some irregularity, such as,V(m + 1, N) > V(*m*, N) for some *m*.

To get a non increasing function, one can express an empirical structural type distribution as

$$g(m, \mathrm{N}) = \sum_{\omega \geq m} \mathrm{V}(\omega, \mathrm{N}) \quad (2)$$

* Hindustan Institute of Technology and Science, Chennai, India. *E-mail : lakshmi@hindustanuniv.ac.in*

$g(m, N)$ corresponds to the number of types of words with the frequency m or more in a sample of N tokens.The relation between $g(m, N)$ and $f_z(z, N)$ is that they are inverse of each other.

$$g(m, N) = z <=> f_z(z, N) = m$$

The characteristic of LNRE distribution is the presence of large numbers of words with vey low probability of occurrence . For example, in the Briticsh National Corpus more than half the types of words have sample relative frequencies of the order of 10^{-8} [2]. Khamaladze [3] has defined a sequence of words to be with LNRE if

$$\lim_{N \to \infty} \frac{E[V(1, N)]}{N} > 0$$

In other words, LNRE distribution is present in case the growth rate of vocabulary remains greater than zero even when N is increased indefinitely.

In the modelling of frequency spectrum , the structural type distribution $G(\pi)$ giving the number of types of words in a population with the probability greater than or equal to π is used. Using binomial distribution characterizing the number of patterns of m occurrences of a word in N tokens, it has been shown in Baayen [2] that the expectation of frequency spectrum is given by

$$E[V(m, N)] = \int \frac{(N\pi)^m}{m!} e^{-N\pi} dG(\pi) \quad (3)$$

where π is the success probability of occurrence of a word in N word tokens.

Carroll [4] defined $G(\pi)$ in (3) in terms of log normal structure distribution. Sichel [5, 6] has given an inverse Gauss –Poisson Structural type distribution to derive $G(\pi)$. Baayen [3] has described a Zipfian family LNRE model based on relative spectrum elements defined as

$$\alpha(m, N) = \frac{E[(V(m, N)]}{E[V(N)]} \quad (4)$$

As given in Baayen [3], relative spectrum models $\alpha(m, N)$ of the form of (4) have been proposed by Zipf, Yule, Yule-Simon, Waring-Herdan-Muller, Karlin-Rounat and Zipf-Mandelbrot . Evert [7] has considered a structural type distribution of Zipf-Mandelbrot law and derived an expression of frequency spectrum as

$$E[(V(m, N))] = \frac{C}{m!} . N^{\alpha} \Gamma(m - \alpha)$$

where $\Gamma(.)$ is a Gamma function [8], α is a constant and C is a normalizing parameter.

The model (3) is further used for interpolation and extrapolation to predict expectation of $V(m, N)$ for $N < N_0$ and $N > N_0$ respectively, where N_0 is the corpus for which expectation of $V(m, N)$ has been modelled .

In this paper, we take (1) as a population model to estimate $V(m, N)$. Further, using (1) we derive a logarithmic type expression for the structural type distribution $g(m, N)$ given in (2). Approximating the logarithmic terms by a polynomial in $(1/m)$ of arbitrary order, we formulate a linear regression model for $g(m, N)$. Using the maximum likelihood solution of the linear regression model, we estimate the regression co-efficient corresponding to the best fit of the given empirical $g(m, N)$ data.

The rest of the paper is organised as follows. The following section describes the steps involved leading to the regression model development. Section III describes the simulation results for evaluating the proposed regression model. Section IV concludes the paper.

2. MODEL DEVELOPMENT

The discrete structural type distribution is given by

$$g(m, N) = \sum_{\omega \geq m} V_{\omega}(m, N) \quad (5)$$

Considering V(m, N) as a continuous function of m, $g(m, \mathrm{N})$ can be wriiten as

$$g(m, \mathrm{N}) = \int_m^{\mathrm{K}} \mathrm{V}(m, \mathrm{N}) dm$$

where K is the number of classes represented by m.Using (1) ,(5) can be written as

$$\begin{aligned} g(m, \mathrm{N}) &= \int_m^{\mathrm{K}} \frac{\mathrm{C}}{m(m+1)} dm \\ &= \int_{\mathrm{m}}^{\mathrm{K}} \frac{\mathrm{C}}{m} dm - \int_m^{\mathrm{K}} \frac{\mathrm{C}}{m+1} dm \\ &= \mathrm{C}\{\log m\}_m^{\mathrm{K}} - \mathrm{C}\{\log m + 1\}_m^{\mathrm{K}} \end{aligned}$$

Simplifying,
$$g(m, \mathrm{N}) = \mathrm{C}\left[\log\left(\frac{\mathrm{K}}{\mathrm{K}+1}\right) + \log\left(\frac{1}{m}+1\right)\right] \quad (6)$$

Proposition 1 : Using emprirical data, $g(m, \mathrm{N})$, given by (6), can be approximated to a linear regression model of order $p + 1$, $p > 0$

$$\mathrm{X}q = g + \varepsilon_0 \quad (7)$$

where $\varepsilon_0 \approx \mathrm{N}_n(0, \sigma_n)$ 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.

$$\begin{aligned} q &= [q_1, q_2, \ldots\ldots\ldots q_j \ldots\ldots\ldots q_p, q_0]^t \\ g &= [g(1, \mathrm{N}), g(2, \mathrm{N}), \ldots\ldots\ldots g(m, \mathrm{N}) \ldots\ldots\ldots g(n, \mathrm{N})]^t \\ \mathrm{X} &= [x_{i,j}] ; \\ i &= 1, 2, 3, \ldots\ldots\ldots n, \\ j &= 1, 2, 3, \ldots\ldots\ldots p + 1 \end{aligned}$$

$$x_{i,j} = \frac{1}{i^j};$$

$$\begin{aligned} i &= 1, 2, 3, \ldots\ldots\ldots\ldots\ldots n ; \\ j &= 1, 2, \ldots\ldots\ldots\ldots\ldots p \\ x_{i,\,p+1} &= 1, \\ \forall\, i &= 1, 2, 3\ldots \\ m &= 1, 2, \ldots n \end{aligned}$$

and

Proof : See Appendix

Solving (7) for maximum likelihood solution of regression co-efficients,we get,

$$\hat{q} = [(\mathrm{X}^t \mathrm{X})^{-1} \mathrm{X}^t] \quad (8)$$

where $\hat{q}$ is the estimate of q.

3. SIMULATION RESULTS

Taking the frequency data of Alice in Wonderland, as given in [2], the linear regression equation of (7) is used to fit the data for different orders of regression model using solution (8). Formulating different orders of regression model starting from order two to order eight ,the least square solution is found for each model order. The least square error for models of order four to eight are tabulated in Table 1. As seen in table 1, the regression model of order eight gives the best least mean square error fit with the empirical data . Figure 1 provides the plot of the empirical data together with eighth order model fit in terms $g(m, \mathrm{N})$ against class index m. Though the visual inspection of Figure 1 gives the impression of good fit of the model with empirical data , in order to evaluate the goodness of fit of Figure 1, a Chi-Square test was

performed . The results of Chi-Square test in Table 2 show that model order eight provides a good fit of the data [10]. Further , to evaluate the proposed regression model against the existing frequency spectrum models, we compare the results obtained with the proposed regression model against those of Sichel , Carroll and Zipf as given in Table 3 for the data taken from Baayen [2]. The data for V(*m*, N) of the proposed model given in Table3 corresponds to value derived from g(m,N) fit against the structural data adapted from Baayen [2]. It is seen from Table3 that the proposed model gives a favourable Chi-Square test result compared to other existing models for the data considered.

4. CONCLUSION

In this paper, we have proposed a model of a frequency spectrum of English text using the method of linear regression. The model is fit into a structured type distribution which is a non decreasing function without any irregularities. Using frequency data available its is shown that the proposed regression model fits well and satisfies the goodness of fit condition .Further the proposed regression model is evaluated by comparing its performance against the existing models of Carroll ,Sichel and Zipf.It is seen that the results of proposed model compares favourably with those of existing models.

As a future work, it is proposed to build a model which gives the expected values of spectral frequency distribution using Poisson distribution of word outcomes in a given text and the structural type of distribution of word frequency.

Appendix : 1

Proof of Proposition : 1

Rewriting (6),
$$g(m, \mathrm{N}) = \mathrm{C}\log\left(\frac{\mathrm{K}}{\mathrm{K}+1}\right) + \mathrm{C}\log\left(1+\frac{1}{m}\right)$$
$$= \mathrm{C}' + \mathrm{C}\log\left(1+\frac{1}{m}\right)$$

where
$$\mathrm{C}' = \mathrm{C}\log\left(\frac{\mathrm{K}}{\mathrm{K}+1}\right)$$

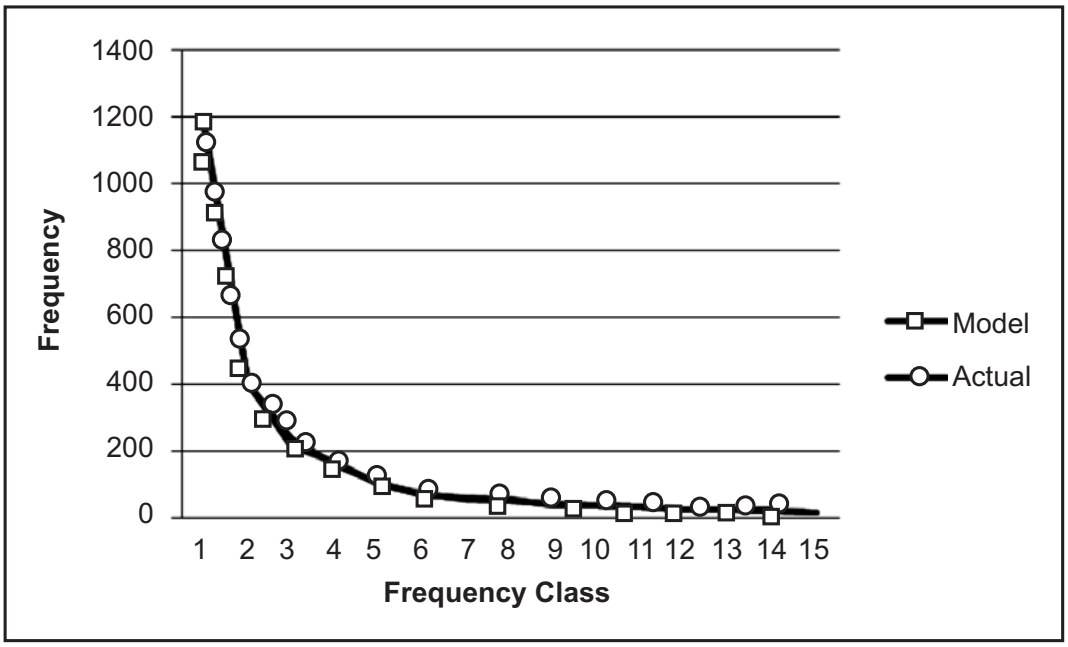

Figure 1: Plot of frequency spectrum for Model and Actual data(Alice in Wonderland)

Considering the expansion of $\log\left(1+\frac{1}{m}\right)$, $\left|\frac{1}{m}\right| \leq 1$ as a power series[10, 11], we approximate $g(m, N)$ using only $p + 1$ terms of the series, $p > 0$ as

$$g(m, N) \approx q_0 + \frac{q_1}{m} + \frac{q_2}{m^2} + \frac{q_3}{m^3} + ...\frac{q_p}{m^p} \quad (a.1)$$

Considering the empirical data of $g(m, N)$, $m = 1, 2, 3....n$, (*a*. 1) can be put in a linear regression form as

$$Xq = g + \varepsilon_0$$

as stated in Proposition 1.Hence the result.

Table 1

Root Mean Square Error of proposed model applied to Text in Alice in Wonderland

Model order	*Fourth*	*Fifth*	*Sixth*	*Seventh*	*Eighth*
Root Mean Square	431.917	290.4784	186.3808	214.4632	132.5941

Table 2

Chi-Square Statistic for proposed model applied to Text in Alice in Wonderland

Model Order	*Fifth order*	*Sixthh*	*Seventh order*	*Eighth*
Chi-Square Statistic (CV)	5.853487	5.48	5 .29	4.6834
Cumulative probability $P(X^2<=CV)$	0.03	0.02	0.01	0.005

Table 3

Evaluation of proposed model against Sichel, Carroll and Zipf models using Chi- Square Statistic

Model	*Chi-Square Statistic*	*Cumulative Probability* $P(X^2<=CV)$
Proposed Model	4.68347	0.01
Sichel	16.3309	0.9992
Carrol	8.110846	0.12
Zipf	8.223035	0.12

5. REFERENCES

1. Zipf.G.K, ,The Psycho –Biology of Language ,Houghton Mifflin, Boston(1935)
2. Baayen, R. Harald. *Word frequency distributions*. Vol. 18. Springer Science & Business Media,(2001)
3. 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 Compute Science(1987).
4. Carroll, John B. "A rationale for an asymptotic lognormal form of word-frequency distributions." *ETS RESEARCH BULLETIN SERIES* 1969.2 (1969): i-94.
5. Sichel, Herbert S. "On a distribution law for word frequencies." *Journal of the American Statistical Association* 70.351a (1975): 542-547.
6. Sichel, H. S. "Word frequency distributions and type-token characteristics." *Mathematical Scientist* 11.1 (1986): 45-72.
7. Evert, Stefan. "A simple LNRE model for random character sequences." *Proceedings of JADT*. Vol. 2004. (2004)
8. Weisstein, Eric W. "Eric Weisstein's world of mathematics." (2002).
9. www. http://stattrek.com/chi-square-test/goodness-of-fit.asp
10. www.wikepedia.org/wiki/Natural-lograthim.
11. www.math2.org/math/expansion/log.htm.