**Regression model of rank–frequency data of Tamil text**

S. Lakshmisridevi 1, a, 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 is 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 \( f_i(z, N) \) denote the frequency of word sample, where \( z \) denotes the rank and \( N \) corresponds to number of word tokens. According Zipf law,

\[
f_i(z, N) = \frac{C}{z^\alpha}
\]

(1)

Where \( \alpha \) is a free parameter, \( C \) is a normalizing constant and

\[
N = \sum_{i=1}^z f_i(z, N)
\]

(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

\[
(z + m)^\alpha f = C
\]

(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**

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

(3)

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

\[
B \ln(z + m) + \ln f = \ln C
\]

(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

\[
\ln(Q + \sum_{Q(z)} \frac{1}{z^\alpha}) + \ln f = \ln C
\]

(5)

Where
\[ Q_n = \sum_{n=1}^{p} \frac{1}{n^i} \]

\[ Q_i = (-1)^n \left( \sum_{k=1}^{i} \frac{1}{i+k} \right) \]

\[ i = 1, 2, 3, ..., p \]

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

\[ \log f = \sum_{i=1}^{p} q_i \rho^i \]

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

\[ Y = X^r + \varepsilon_0 \]

Where

\[ Y = [\ln f_1, \ln f_2, ..., \ln f_n] \]

\[ X = [x_i] ; i=1,2,3,.....n, j=1,2,3,.....p+1 \]

\[ x_{ij} = \frac{1}{z^j} ; z = 1,2,3,.....n ; j=1,2,3,.....p \]

\[ x_{ip+1} = 1, \forall i \]

\[ \varepsilon_0 \approx N_n(0, \sigma^2, \varepsilon_n) \] corresponds to the noise term assumed to be a multivariate normal i.i.d distribution of \( n \) variables with zero mean and variance \( \sigma^2 \).

Maximum likelihood solution for (7) is given as

\[ \hat{\gamma} = \left( X^T \cdot X \right)^{-1} X \]

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 in Table 1.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Square Value of Error</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Regression Model Order</td>
<td>Third</td>
</tr>
<tr>
<td>Mean Square</td>
<td>4.5E-05</td>
<td>2.0E-06</td>
</tr>
<tr>
<td>Table 2: Results of Chi-Square Test of Regression Model Order Eight.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>No. of ranks</td>
<td>Degrees of freedom</td>
</tr>
<tr>
<td>Eight</td>
<td>24</td>
<td>23</td>
</tr>
</tbody>
</table>

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