I J C T A, 9(41) 2016, pp. 1227-1234

An Analysis of Variation of Model Parameters of Text Rank -Frequency Data with Corpus Size

S. Lakshmisridevi[1] and R. Devanathan[2]

ABSTRACT

Rank frequency data of text has been modelled by Zipf and modified by Mandelbrot, but the model parameters are seen to vary with the corpus size. In this paper, using a linear regression model of rank - frequency data, we derive a a closed form relation between the maximum likelihood solutions of model parameters for two different corpus sizes. We further show that a closed form relation exists between the least squared errors of the maximum likelihood solutions for two different corpus sizes. Empirical data is used to verify the theory.

Keyword: Zipfs law, Zipf-Mandelbrot law, Regression, Chi-Square test, Maximum likelihood solution

I. INTRODUCTION

Zipf has been a pioneer in quantitative linguistic modelling of words in a text. Zipf [1,2] ranked words in a text in the order of decreasing frequency. Frequency of words is plotted against rank. Let $f_z(z, N)$ denote frequency of samples of N word tokens of a text for rank $z \in \{1,2,..... n\}$ where n is the maximum rank considered. Zipf formulated the following relationship

$$f_z(z,N) = \frac{C}{\tilde{z}}$$

where α is a free parameter, C is a normalizing constant and $N = \sum_{z=1}^{n} f_z(z,N)$. Equation (1) is known as Zipf's law. Mandelbrot [3] generalized (1) in order to correct deviation with respect to empirical data especially for the lower ranks. The modified form known as the Zipf –Mandelbrot (ZM) law is given as

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

where $m > 0$.

Baayen [4] has stated that lexical statistics does not seem to follow the traditional law of large numbers applicable to the theory of probability. For example, Baayen has demonstrated that vocabulary of a given corpus increases as a function of corpus size though the rate of increase of vocabulary decreases as the corpus size increases.

Taking lograthim, (1) can be put as

$$\ln f_z(z, N) = \ln C - a \ln z \tag{3}$$

By plotting $\ln f_z(z, N)$ vs $\ln z$, one can obtain a linear plot corresponding to equation (3) with the intercept on the lnfz(.) axis being ln C and the slope corresponding to $\alpha \neq 0.92$ as stated in [5]. Baayen has

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

shown in [4] that the intercept ln C increases with the increase of corpus size and also the slope decreases from - 0.92 to a more negative value as the corpus size increases.

Given the variation of model parameters with corpus size, research has been carried out to find characteristic constants of text with parameters not varying appreciably with corpus size [4]. Yule in [6] has characterized a constant K which is described in terms of frequency spectrum of words in a text given as V(m, N) corresponding to m occurrences of the words in the text. N corresponds to number of words in the given text. K is given as

$$K = 10^4 \frac{\sum_m m^2 V(m,N) - N}{N^2}$$

Simpson in [6] defined a measure denoted as D where

$$D = \sum_m V(m,N)\left(\frac{m}{N}\right)\left(\frac{m-1}{N-1}\right)$$

Baayen [4] has plotted characreristic constants K and D as functions of sample size taken from “Alice in Wonderland “. The plots given in Baayen [4] show that the value of K and D fall mostly outside 95% confidence interval based on Monte Carlo simulation of 5000 permutation runs of the text . This means that the variation in K and D cannot be attributed to a random nature of word occurrence in the text but perhaps to an underlying characteristic which is not captured by parameters K or D. Various other text characteristic constants have also been proposed. Guiraud [8] has proposed a measure called R given as,

$$B - \frac{V(N)}{\sqrt{N}}$$

where V(N) is the vocabulary size of corpus of N tokens. Brunet [9] gave a power relation

$$W = N^{V(N)-a}$$

where ‘a’ is a parameter .It is also shown by Baayen [4] that considering the text Alice in Wonderland , the constants R and W vary significantly with the corpus size. Sichel [10] has derived a constant called S given as

$$S = \frac{V(2, N)}{V(N)}$$

Honore [11] has proposed a constant

$$H - \frac{\ln N}{\left(\frac{1 - V(1,N)}{V(N)}\right)}$$

Both S and H are shown to vary in [4] with corpus size but within 95% confidence interval of Monte Carlo 5000 runs using the text Alice in Wonderland . Herdan [12] has related the vocabulary size against N as follows.

$$\ln V(N) = \log \beta + C \ln N$$

where β is a parameter. As per Herdan [12] , the plot of logV(N) vs log N where N is the corpus size is to be a straight line. However the slope of Herdan relationship, varies with corpus size as shown in [4].

Clearly from the above brief survey, one can conclude that the parameters of a model fitting a text cannot be considered to be constant over a range of the corpus sizes. Then the question arises if the model

parameter vary with the corpus size, can we formulate a law of variation of parameters as a function of corpus size?. This then is the subject of the paper.

In this paper, we first formulate a linear regressive model based on Zipf -Mandelbrot law for a given text based on the author's earlier work [13] . We then show that the linear regressive model parameters vary with the corpus size and one can formulate variation of these parameters as a function of corpus size in closed form . This is in contrast with the existing results where the variation of the parameters with respect to the corpus size has not been modelled .The closed form formulation of model parameter variation is made possible in our case since our model being a linear regressive model, the maximum likelihood solution can be used to model the variation of parameters in closed form. The model of variation of parameters is then supported by empirical findings.

To summarize the rest of the paper, section II describes the regressive model. A relationship on the variation of the regressive model parameters is developed in section III .Section IV uses empirical data to support the theory developed in section III. Section V concludes the paper.

II. LINEAR REGRESSION MODEL

Based on an earlier work, we propose in this section a linear regressive model of Zipf- Mandelbrot law of a given text .Taking lograthim of (2)

$$\alpha \log(z+m) + \log f = \log C \tag{4}$$

By approximating logarithmic terms in (4) in terms of a polynomial of (1/z) of order p> 0 , (4) can be shown, as in [13] , to approximate a linear regression model as in the following proposition given without proof.

Proposition I

Equation (4) can be put in the form

$$\ln f = X\,\theta + \varepsilon_0 \tag{5}$$

where

$$\ln f = [\ln f_1, \ln f_2 \ldots\ldots \ln f_i \ldots\ldots \ln f_n]^t$$
$$\theta = [\delta 1, \delta_2, \ldots\ldots\ldots \delta_j \ldots\ldots \delta_p, \delta_0]^t,$$

t stands for transpose,

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

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

and

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

$\varepsilon_0 \neq N_n\ (0, \sigma_n)$ 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 θ in (5) is given as

$$\theta = [(X^t X)^{-1} X^t] \ln f \tag{6}$$

or,

$$\theta = \Phi \ln f$$

where

$$\Phi = (X^t X)^{-1} X^t$$

III. MODEL PARAMETER VARIATION

Considering normalized frequencies of data of two corpus sizes N_i, i = 1,2, let

$$f_i = [f_{i1}, f_{i2}, \ldots f_{ij} \ldots\ldots f_{in}]$$

where

$$f_{ij} = \frac{F_{ij}}{N_j}, i = 1,2 \;\; j = 1,2,\ldots\ldots n$$

F_{ij} corresponds to the absolute frequency of i-th corpus corresponding j-th rank.

One can write diagonal matrix of order n as

$$\Psi = \text{diag}\left[\frac{F_{21}}{N_2}/\frac{F_{11}}{N_1}, \frac{F_{22}}{N_2}/\frac{F_{12}}{N_1}, \ldots\ldots \frac{F_{2j}}{N_2}/\frac{F_{1j}}{N_1} \ldots\ldots \frac{F_{2n}}{N_2}/\frac{F_{1n}}{N_1}\right]$$

$$f_2 = \Psi \; f_1$$

Proposition II

$$\ln \Psi = \ln f_2 - \ln f_1$$

where

$$\ln \Psi = [\ln \Psi_{11}, \ln \Psi_{22}, \ldots\ldots \ln \Psi_{jj}, \ldots\ldots \ln \Psi_{nn}]^t$$
$$\ln f_i = [\ln f_{i1}, \ln f_{i2}, \ldots\ldots \ln f_{ij}, \ldots\ln f_{in}]^t, \; i = 1,2.$$

Proof :- See Appendix

Proposition III

Given two corpus sizes N_i i =1,2 of a text with rank frequency data f_i, i =1,2 and the maximum likelihood solution

$$\theta_i, \; i = 1,2$$
$$\theta_2 = \theta_1 + \ln \Psi$$

Proof :- See Appendix.

Proposition IV Given

$$\varepsilon_i = X\theta_i - \ln f_i, \; i = 1,2$$
$$E_i = \varepsilon_i^t . \varepsilon_i \; i = 1, 2, \tag{7}$$

It follows that

$$E_2 = E_1 + (\ln f_2 - \ln f_1)^t (I - \Phi^t X^t) (\ln f_1 + \ln f_2 - 2X \theta_1)$$

Proof :- See Appendix.

Remark : Proposition I proposes a linear regression model of arbitrary order of rank –frequency data with the maximum likelihood solution of the model being given by (6) . Using this solution, in closed form, Proposition III relates the model parameter solution for any two different corpus sizes . Further, Proposition IV shows that least square error of maximum likelihood solutions for of any two corpus sizes can also be related.

IV. SIMULATION RESULTS

Fig 1. plots the fit of rank-log normalized frequency data of the Bible text using proposed regression model of order eight. Chi-Squared test [14] on the fit yields a critical value CV= 0.07564 yielding a cummilative probability $P(\chi^2) \leq CV$ which is nearly zero .This shows that the fit of Fig. 1. satisfies the goodness of fit test.

Fig. 2 and Fig. 3 show the variation of regression model parameters with the corpus size using Bible text corpora. The model parameters δ_i = 1,2,.....8,0 are plotted along the horizontal axis .Model parameter values for different corpus sizes are plotted vertically in the different columns.

Table1 provides a verification of the relationship of model parameter solutions for two different corpus sizes Ni, i = 1, 2 using the result of Proposition III . The table shows that the prediction of the model parameter ,as per the theory , agrees with computational data .Table 2 gives verification of computation of least square error of model fit for two corpus sizes of Bible text corresponding to Ni , i = 1,2 according to the result of Proposition IV.Again the theory has been verified by computation.

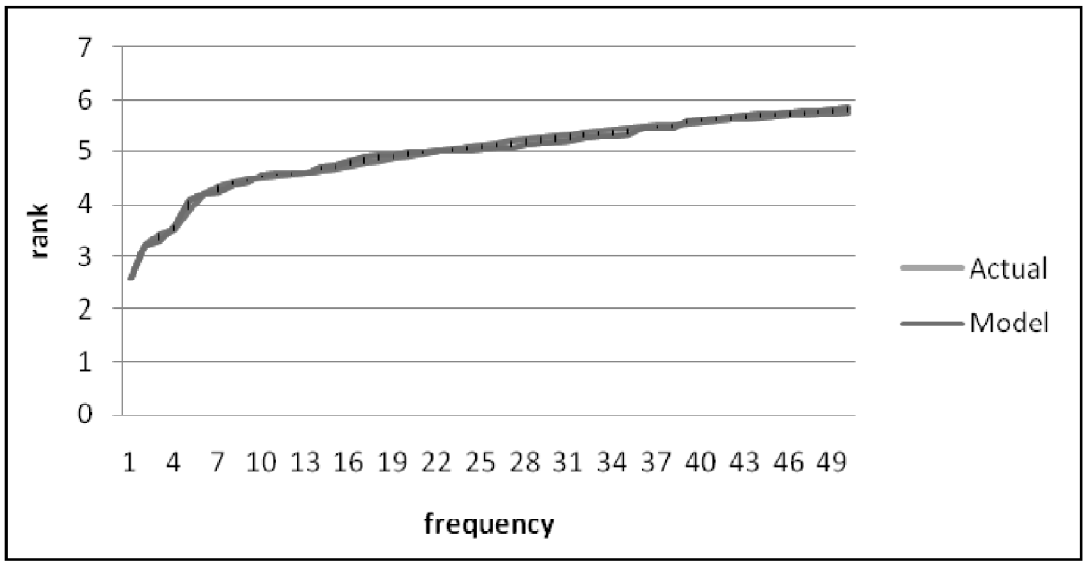

Figure 1: Eighth order regression model fit of ln normalized frequency –rank data of Bible text corpora of 1349 pages

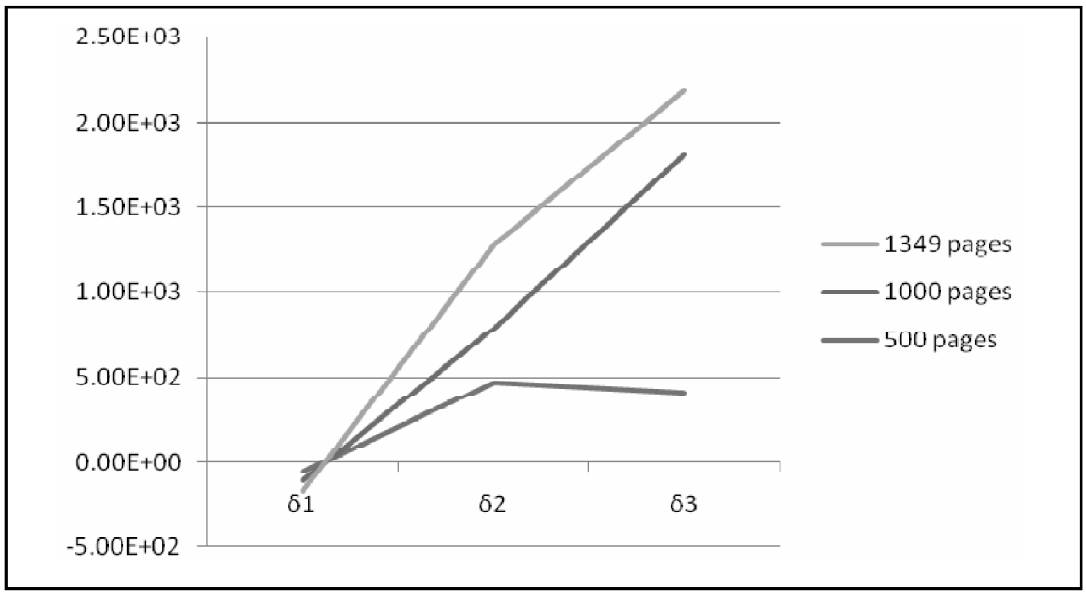

Figure 2: Variation of Model parameters –δ_1 to δ_3 for change in corpus size

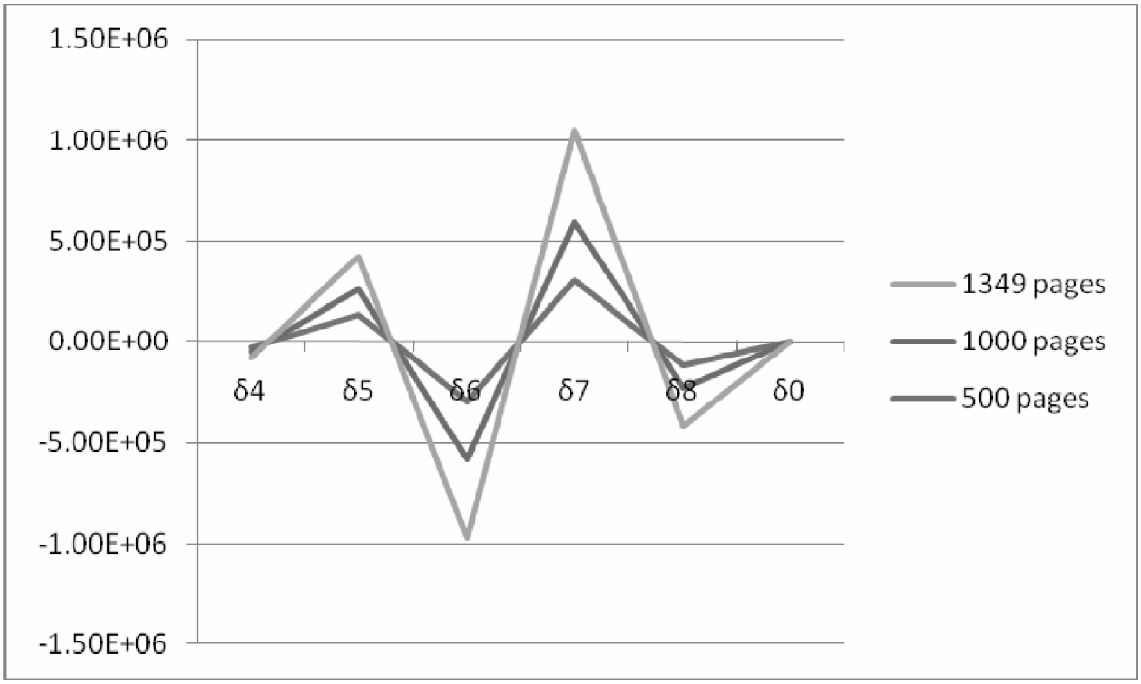

Figure 3: Variation of Model parameters- δ_4 to δ_8 and δ_0 for change in corpus size

Table1
Verification of relationship of model parameter solutions for two Corpus sizes (Ni,i =1,2)

Model Parameters	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_0
θ(Using (6) for N2.)	- 4.96e+01	3.6e+02	1.40e+03	-2.71e+04	1.29e+05	-2.e+05	2.911e+05	-1.120e+05	6.58e+00
q by (Using (6) for N2.)	6.02e+01	-4.92e+02	- 3.762e+02	2.80e+04	-1.60e+05	3.97e+05	-4.55e+05	1.879e+05	-6.77e+00
θ (Using 2 Proposition–III)	6.03E+01	-4.93E+02	-3.78E+02	2.80E+04	-1.61E+05	3.97E+05	-4.52E+05	1.88E+05	-6.78E+00

Table 2
Computation of Squared error for two different Corpus sizes

Corpus Size	*Squared error calculation*	
First 1000 pages of Bible (Corpus Size N1) directly from the model fit	$E_1 = \varepsilon^t_1 \varepsilon_1$	0.349732
First 1349 pages of Bible (corpus Size N2) directly from the model	$E_2 = \varepsilon^t_2 \varepsilon_2$	0.304560
First 1349 pages of Bible (corpus Size N2 using the result of Proposition- IV)	E_2	0.307264

V. CONCLUSION

Zipf-Mandelbrot law has been used to fit rank frequency data of text for given corpus size.However, the parameters of the fit are shown to vary with the corpus size .In this paper, we show that a linear regression model can provide a good fit of rank-frequency data along with the provision that the variation of model parameters with corpus size can be modelled as well.

It has been shown that an eighth order linear regression model fits the empirical data of a Bible text corpora accurately. It is also shown that the parameters of model for maximum likelihood solution vary with the corpus size. However exploiting the closed form solution of the regression model we are able to predict the variation of parameters as a function of corpus size.Using the Bible text corpora, the predictions are verified .

Further, the closed form solution of the regressive model allowed us to predict the relationship between least squared error for two different corpus sizes. The squared error predictions were also verified using actual computation .

In future work, we propose to use a Poisson distribution to predict the rank-frequency value of any corpus size given the maximum likelihood solution of the model for a particular corpus size.

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 huma 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 Statstical Structure of language,in W. E. Jackson (e.d.), Communication theory, Academic press, New York (1953), pp 503-512.

[4] Baayen, R. Harald. *Word frequency distributions*. Vol. 18. Springer Science & Business Media, (2001).

[5] Wyllys, Ronald E. “Empirical and theoretical bases of Zipf’s law.” *Library Trends* 30.1 (1981):53-64.

[6] Yule, C. Udny. *The statistical study of literary vocabulary*. Cambridge University Press, 2014.

[7] Simpson, Edward H. “Measurement of diversity.” *Nature* (1949).

[8] Guiraud, Pierre. *Les caractères statistiques du vocabulaire*. Presses universitaires de France, 1954.

[9] Brunet, Etienne. *Le vocabulaire de Jean Giraudoux structure et évolution*. Éditions Slatkine, 1978.

[10] Honoré, Antony. “Some simple measures of richness of vocabulary.” *Association for literary and linguistic computing bulletin* 7.2 (1979): 172-177.

[11] Sichel, H. S. “Word frequency distributions and type-token characteristics.” *Mathematical Scientist* 11.1 (1986): 45-72.

[12] Herdan, Gustav. “Quantitative linguistics.”,Butterworths (1964).

[13] R.Devanathan,S.Lakshmisridevi .”Modified Zipf-Mandelbrot law using LinearRegression(Accepted for Presenatation, RTCSE’16 Kulalampur Malaysia, Jan, 02 .2017).

[14] www. http://stattrek.com/chi-square-test/goodness-of-fit.asp

Proof of Proposition II:-

$$f2 = \mathrm{Y} \ln f1 \qquad (A.1)$$

Taking logarithm element wise

$$(\ln f_2)_j = \ln \Psi_{jj} + (\ln f_i)_j$$

where $(\ln f_i)_j$ corresponds to the j-th element of $\ln f_i$, $i = 1,2$, $j = 1,2,...n$.

That is,

$$\ln f_2 = \ln \Psi + \ln f_1$$

where

$$\ln \Psi = [\ln \Psi_{11}, \ln \Psi_{22},.....\ln . \Psi_{jj},.....\ln \Psi_{nn}]t$$

Hence

$$\ln \Psi = \ln f_2 - \ln f_1 \qquad (A.2)$$

Proof of Proposition III

$$\theta_1 = \Phi \ln f_1 \qquad (A.3)$$

$$\theta_2 = \Phi \ln f_2 = \Phi(\ln f_1 + \ln \Psi) = \Phi \ln f_1 + \Phi \ln \Psi$$

$$\theta_2 = \theta_1 + \Phi \ln \Psi \qquad (A.4)$$

Proof of Proposition IV

$$\begin{aligned}\varepsilon_2 &= X\theta_2 - \ln f_2 \\ &= X(\theta_1 + \Phi \ln \Psi) - \ln f_2 \\ &= X\theta_1 + X\Phi \ln \Psi - \ln f_2 \\ &= X\theta_1 - \ln f_1 + \ln f_1 + X\Phi \ln \Psi - \ln f_2\end{aligned}$$

Using (*A*.2), and since

$$\varepsilon_1 = X\theta_1 - \ln f_1,$$

we can have

$$\varepsilon_2 = \varepsilon_1 + (X\Phi - I)(\ln f_2 - \ln f_1)$$

$$E_2 = \varepsilon_2^t \varepsilon_2 = [\varepsilon_1 + (X\Phi - I)(\ln f_2 - \ln f_1)]^t [\varepsilon_1 + (X\Phi - I)(\ln f_2 - \ln f_1)]$$

$$E_2 = \varepsilon_1^t \varepsilon_1 + 2\varepsilon_1 (X\Phi - I)(\ln f_2 - \ln f_1) + (\ln f_2 - \ln f_1)^t (X\Phi - I)^t (X\Phi - I)(\ln f_2 - \ln f_1)$$

But

$$(X\Phi - I)^t (X\Phi - I) = (I - \Phi^t X^t)$$

Using (7) we have

$$E_2 = E_1 + 2\varepsilon_1^t (X\Phi - I)(\ln f_2 - \ln f_1) + (\ln f_2 - \ln f_1)^t (I - \Phi^t X^t)(\ln f_2 - \ln f_1)$$

$$E_2 = E_1 + 2(\ln f_2 - \ln f_1)^t (\Phi^t X^t - I)\varepsilon_1 + (\ln f_2 - \ln f_1)^t (I - \Phi^t X^t)(\ln f_2 - \ln f_1)$$

$$= E_1 + (\ln f_2 - \ln f_1)^t (I - \Phi^t X^t)(1 \quad f_2 - \ln f_1 - 2\varepsilon_1)$$

n

$$= E_1 + (\ln f_2 - \ln f_1)^t (I - \Phi^t X^t)\{(\ln f_2 - \ln f_1) - 2(X\theta_1 - \ln f_1)\}$$

That is,

$$E_2 = E_1 + (\ln f_2 - \ln f_1)^t (I - \Phi^t X^t)(\ln f_2 + \ln f_1 - 2X\theta_1)$$

Hence the result.