European Journal of Statistics and Probability
Vol.3, No.1, pp. 1-12, March 2015
Published by European Centre for Research Training and Development UK (www.eajournals.org)

A GENERALIZED METHOD FOR ESTIMATING PARAMETERS AND MODEL OF BEST FIT IN LOG-LINEAR MODELS.

[1] **Okoli, C.N.;** [2] **Onyeagu, S.I. and** [3] **Osuji, G.A.**

Department of Statistics, Anambra State University, Nigeria. [2&3] Department of Statistics, Nnamdi Azikiwe University, Awka.Nigeria

Postal Address: Department of Statistics, Anambra State University, P.M.B 02, Uli, Nigeria.

ABSTRACT: *In this article, we proposed generalized method and developed algorithms for estimation of parameters and best model fit of log linear model for* q; $q \geq 2$ *dimensional contingency tables. For purpose of this work, the method was used to provide estimates of parameters of log –linear model for four- dimensional contingency table. Parameters of higher dimensional tables can in like manner be estimated. In estimating these parameters and best model fit, computer programs in R were developed for the implementation of the algorithms. The iterative proportional fitting was used to estimate the parameters and goodness of fits of models of the log linear model. A real life data was used for illustration and the result obtained showed the best model fit for four dimensional contingency table is [BSG, BGA]. This showed that the* best model fit must have sufficient evidence to fit the data without loss of information and must have the highest p-value and least likelihood ratio estimate.

KEYWORDS: Generalized Method, Algorithms, Contingency Table, Categorical Data, Parameters, Iterative Proportional Fitting.

INTRODUCTION

Contingency table (cross tabulation) is a table that displays the multivariate frequency distribution of the variables. Also, q-dimensional contingency table is a contingency table formed by cross-classification of more than two categorical variables. A categorical variable is one for which the measurement scale consists of a set of categories (Agresti, 2002). Categorical data are data consisting of counts of observation falling in different categories. Categorical data in contingency tables are collected in many investigations. In order to understand the type of structures in a prevailing data, appropriate log linear models are fitted. Log linear models are used to model the observed cell counts where the log of expected cell count is proportional to a linear combination involving number of model parameters. Each interaction term between two or more factors is associated with a set of model parameters and if these model parameters are non-zero it indicates that there exists an association between factors or variables. Log linear model is similar to more

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

familiar analysis of variance model except that it is applied to natural logarithm of expected frequencies (Jibasen and Lawal, 2004). Response observations in analysis of variance are assumed to be continuous and normal while in log linear modeling, observations are counts having Poisson distribution (lawal, 2003). Log linear analysis is used to determine whether there is any significant association or relationship(s) in q-dimensional contingency tables as well as determining the parsimonious model that explain the observed data .The saturated model contains all the variables being analyzed and all possible interactions between the variables. The variables investigated by log linear models are all response variables. In other words, there is no distinction made between response and explanatory variables. Indeed, Log linear model is a special case of generalized linear models (Mc Cullagh and Nelder, 1989).

REVIEW OF RELATED LITERATURE

Brown (1976) has shown that all interactions that are significant in at least one of the test of marginal and partial associations form the base model, He suggested that suppose model (W) is nested in model (V) , the hypothesis that the additional terms added to W to obtain V are equal to zero is examined using the statistic $G^2\left(W/V\right)=G^2\left(W\right)-G^2\left(V\right)$. In their work, Bishop et al (1975) traced the history of methods that have been proposed for iterative proportional fitting of log-linear models. They observed that Bartlett (1935) was the first to describe a method of getting MLE'S for a model that does not possess closed-form estimates. For a 3-dimensional contingency table (i x j x k), he noted that the cross-product condition specified by the model is

$$\frac{\hat{m}_{111}\hat{m}_{221}}{\hat{m}_{121}\hat{m}_{211}} \cdot \frac{\hat{m}_{122}\hat{m}_{212}}{\hat{m}_{112}\hat{m}_{222}} = 1$$

Which can be written in terms of the observed $\left\{x_{ijk}\right\}$ and unknown deviation θ as

$$\frac{\left(x_{111}+\theta\right)\left(x_{221}+\theta\right)}{\left(x_{121}-\theta\right)\left(x_{211}-\theta\right)} \cdot \frac{\left(x_{122}+\theta\right)\left(x_{212}+\theta\right)}{\left(x_{112}-\theta\right)\left(x_{222}-\theta\right)} = 1$$

This work by Bartlett was later extended to multiple categories and methods were also proposed for obtaining solutions to the constraining equations.

Roy and Kastenbaum (1956) and the review by Goodman (1964b). Most authors as observed by Bishop et al, suggested matrix inversion techniques or proposed methods of simplifying the necessary matrix inversion.

The use of Newton- Raphson techniques and their variates have also been suggested by many authors [see Bock (1970, 1972), and Haberman (1974)]. However, Deming and Stephan (1940) developed algorithm for proportional fittings of log linear models. The main disadvantage of all

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

these methods and algorithms developed is that for higher dimensional contingency tables, the procedures become more difficult to handle. However, in our own method a generalized method for the estimation of parameters and algorithms are developed and implemented in R programs for estimating the parameters and model(s) of best fit for four-dimensional contingency table. The advantage of the proposed method is that it saves the computer memory and can handle any higher dimensional contingency table(s) so long as the rules are observed.

METHODOLOGY

The iterative proportional fitting (IPF) in Deming and Stephan (1940) is used to estimate model parameters and best model fit of log linear model. This is to ensure that the expected values are obtained iteratively for model whose expected values are not directly obtainable from marginal totals of observed values. For example if we consider 4- factor model, without 4-factor interaction given as

$$\log_e(m_{ijkl}) = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{4(l)} + \mu_{12(ij)} + \mu_{i3(ik)} + \mu_{14(il)} + \mu_{23(jk)} + \mu_{24(jl)} + \mu_{34(kl)} + \mu_{123(ijk)}$$

$$+ \mu_{124(ijl)} + \mu_{134(ijl)} + \mu_{234(jkl)} \qquad (3.01)$$

To illustrate IPF algorithm for estimating expected frequencies (m_{ijkl}) we consider

Totals $m_{ijk.}, m_{ij.l}, m_{i.kl}$, and $m_{.jkl}$ are characterized to be equal to the corresponding observed marginal totals $n_{ijk.}$, $n_{ij.l}$, $n_{i.kl}$, and $n_{.jkl}$ respectively.

The procedure assumes the initial values $m_{ijkl}(0) = 1$ and proceed by adjusting these proportionally to satisfy the first marginal constraint $(m_{ijk.} = n_{ijk.})$, calculated from:

$$m_{ijkl}(1) = \frac{m_{ijkl(0)}\, n_{ijk.}}{m_{ijk.(0)}} \qquad (3.02)$$

Revise expected values $m_{ijkl}(1)$ to satisfy the second marginal's constraint $m_{ij.l} = n_{ij.l}$, using

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

$$m_{ijkl}(2) = \frac{m_{ijkl}(1) n_{ij.l}}{m_{ij.l}(1)} \tag{3.03}$$

Revise expected values $\hat{m}_{ijkl(2)}$ to satisfy the marginal constraint $\hat{m}_{i.kl} = \mathrm{n}_{\mathrm{i.kl}}$, using

$$m_{ijkl}(3) = \frac{m_{ijkl}(2) n_{i.kl}}{m_{i.kl}} \tag{3.04}$$

Complete the cycle by adjusting $m_{ijkl}(3)$ to satisfy the fourth marginal constraint $m_{.jkl} = \mathrm{n}_{.jkl}$, using

$$m_{ijkl}(4) = \frac{m_{ijkl}(3) n_{.jkl}}{m_{.jkl}(3)} \tag{3.05}$$

The steps are repeated until convergence to desired accuracy is attained. Note that for 2, 3 and 5 etc dimensional tables, the same steps could also be applied.

Proposed Method

We now proposed a generalized method for the estimation of parameters in log linear models.

For *any* $\left\|[j]\right\| = t \quad (t \neq 0)$ *and* $q < \infty \ (q \in \mathrm{N})$,

We define

$$\mu_{\{j_t\}(i_{j_t})} := \log\left\{\frac{\sum_{i_{j_q}/i_{j_t}} n_{(i_{j_t},+)}}{I_{i_{j_q}}/I_{i_{j_t}}}\right\} - \sum_{r=0}^{t-1} \sum_{\{j_s\}\in \mathrm{P}(\{j_t\}:|\{j_s\}|=r)} \mu_{\{j_s\}(i_{j_s})} \quad : \{j_t\} \in \mathrm{P}\left([q]:\left|\{j_t\}\right| = t\right) \ \mathrm{t} = 1,2,...,\mathrm{q}$$

Where $[q] = \{1,2,...,q\}$, $\mathrm{P}([q])$ is the power set of $[q]$

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

$\left\|\left[\underline{j}\right]\right\|$ is the level of interaction/cardinality of the set /length of element in the set. j is a member of collections of sequence $[q]$ and

$$\log\left(m_{i_{[q]}}\right)=\sum_{\{j\}\in P([q]:0\le|\{j\}|\le q)}\mu_{\{j\}(i_j)}$$

For example, q=4 the saturated log linear model for 4-diimensional contingency table and its parameters effects is given by

$$\begin{aligned}\log m_{i_1i_2i_3i_4} &= \mu+\mu_{1(i_1)}+\mu_{2(i_2)}+\mu_{3(i_3)}+\mu_{4(i_4)}+\mu_{12(i_1i_2)}+\mu_{13(i_1i_3)}+\mu_{14(i_1i_4)}+\mu_{23(i_2i_3)}+\mu_{24(i_2i_4)}+\mu_{34(i_3i_4)}\\ &\quad+\mu_{123(i_1i_2i_3)}+\mu_{124(i_1i_2i_3)}+\mu_{134(i_1i_3i_4)}+\mu_{234(i_2i_3i_4)}+\mu_{1234(i_1i_2i_3i_4)} \qquad (3.06)\end{aligned}$$

Conventionally, taking $\mu\phi(\phi)=\mu$ and

$\left\|\left[\underline{J}\right]\right\|=0$(i.e$\left[\underline{J}\right]=\phi$);

$$\hat{\mu}=\log\left(\frac{\sum_{i_1i_2i_3i_4}n_{i_1i_2i_3i_4}}{I_{i_1}I_{i_2}I_{i_3}I_{i_4}}\right)$$

$\left\|\left[\underline{J}\right]\right\|=1$;

$$\hat{\mu}_{[1](i_1)}=\log\left(\frac{\sum_{i_2i_3i_4}n_{(i_1,+)}}{I_{i_2}I_{i_3}I_{i_4}}\right)-\hat{\mu}$$

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

$$\hat{\mu}_{[2](i_2)} = \log\left(\frac{\sum_{i_1 i_3 i_4} n_{(i_2,+)}}{I_{i_1} I_{i_3} I_{i_4}}\right) - \hat{\mu}$$

$$\hat{\mu}_{[3](i_3)} = \log\left(\frac{\sum_{i_1 i_2 i_4} n_{(i_3,+)}}{I_{i_1} I_{i_2} I_{i_4}}\right) - \hat{\mu}$$

$$\hat{\mu}_{[4](i_4)} = \log\left(\frac{\sum_{i_1 i_2 i_3} n_{(i_4,+)}}{I_{i_1} I_{i_2} I_{i_3}}\right) - \hat{\mu}$$

$$\left|\left[\underline{J}\right]\right| = 2;$$

$$\hat{\mu}_{[12](i_1 i_2)} = \log\left(\frac{\sum_{i_3 i_4} n_{(i_1 i_2,+)}}{I_{i_3} I_{i_4}}\right) - \hat{\mu} - \hat{\mu}_{[1](i_1)} - \hat{\mu}_{[2](i_2)}$$

$$\hat{\mu}_{[13](i_1 i_3)} = \log\left(\frac{\sum_{i_2 i_4} n_{(i_1 i_3,+)}}{I_{i_2} I_{i_4}}\right) - \hat{\mu} - \hat{\mu}_{[1](i_1)} - \hat{\mu}_{[3](i_3)}$$

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

$$\hat{\mu}_{[14](i_1i_4)} = \log\left(\frac{\sum_{i_2i_3} n_{(i_1i_4,+)}}{I_{i_2} I_{i_3}}\right) - \hat{\mu} - \hat{\mu}_{[1](i_1)} - \hat{\mu}_{[4](i_4)}$$

$$\hat{\mu}_{[23](i_2i_3)} = \log\left(\frac{\sum_{i_1i_4} n_{(i_2i_3,+)}}{I_{i_1} I_{i_4}}\right) - \hat{\mu} - \hat{\mu}_{[2](i_2)} - \hat{\mu}_{[3](i_3)}$$

$$\hat{\mu}_{[24](i_2i_4)} = \log\left(\frac{\sum_{i_1i_3} n_{(i_2i_4,+)}}{I_{i_1} I_{i_3}}\right) - \hat{\mu} - \hat{\mu}_{[2](i_2)} - \hat{\mu}_{[4](i_4)}$$

$$\hat{\mu}_{[34](i_3i_4)} = \log\left(\frac{\sum_{i_1i_2} n_{(i_3i_4,+)}}{I_{i_1} I_{i_2}}\right) - \hat{\mu} - \hat{\mu}_{[3](i_3)} - \hat{\mu}_{[4](i_4)}$$

$\left|\left[\underline{J}\right]\right| = 3;$

$$\hat{\mu}_{[123](i_1i_2i_3)} = \log\left(\frac{\sum_{i_4} n_{(i_1i_2i_3,+)}}{I_{i_4}}\right) - \hat{\mu} - \hat{\mu}_{[1](i_1)} - \hat{\mu}_{[2](i_2)} - \hat{\mu}_{[3](i_3)} - \hat{\mu}_{[12](i_1i_2)} - \hat{\mu}_{[13](i_1i_3)} - \hat{\mu}_{[23](i_2i_3)}$$

$$\hat{\mu}_{[124](i_1i_2i_4)} = \log\left(\frac{\sum_{i_3} n_{(i_1i_2i_4,+)}}{I_{i_3}}\right) - \hat{\mu} - \hat{\mu}_{[1](i_1)} - \hat{\mu}_{[2](i_2)} - \hat{\mu}_{[4](i_4)} - \hat{\mu}_{[12](i_1i_2)} - \hat{\mu}_{[14](i_1i_4)} - \hat{\mu}_{[24](i_2i_4)}$$

$$\hat{\mu}_{[134](i_1i_3i_4)} = \log\left(\frac{\sum_{i_2} n_{(i_1i_3i_4,+)}}{I_{i_2}}\right) - \hat{\mu} - \hat{\mu}_{[1](i_1)} - \hat{\mu}_{[3](i_3)} - \hat{\mu}_{[4](i_4)} - \hat{\mu}_{[13](i_1i_2)} - \hat{\mu}_{[14](i_1i_4)} - \hat{\mu}_{[34](i_3i_4)}$$

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

$$\hat{\mu}_{[234](i_2i_3i_4)} = \log\left(\frac{\sum_{i_1} n_{(i_2i_3i_4,+)}}{I_{i_1}}\right) - \hat{\mu} - \hat{\mu}_{[2](i_2)} - \hat{\mu}_{[3](i_3)} - \hat{\mu}_{[4](i_4)} - \hat{\mu}_{[23](i_2i_3)} - \hat{\mu}_{[24](i_2i_4)} - \hat{\mu}_{[34](i_3i_4)}$$

ALGORITHMS FOR ESTIMATION OF PARAMETERS AND MODEL FITS IN LOG LINEAR MODELS FOR Q-DIMENSIONAL CONTINGENCY TABLES

The step by step approach for the estimation of parameters that will enable us to develop our R-programs for estimating parameters, expected values and goodness of log linear models are as follows:

Step 1: Given a set of observed data with variables say (2, 3, 4,…, q) . Identify the dimension of the

variable.

Step 2: Estimate the parameters of the model by the method of Iterative proportional fitting (IPF)

Step 3: Estimate the expected values by the method of IPF

Step 4: Compute the model estimates using IPF

Step 5: Using R- programs, the chi squared (χ^2), log –likelihood ratio (G^2) and all the parameters of

the model(s) involved are estimated.

ANALYSIS

In order to implement our developed programs, the data on B.Sc grade, State of Origin, gender and age forming a 5x2x2x2 contingency table were collected from retrieved files of 719 graduated students from Statistics department, Nnamdi Azikiwe University, Awka, Nigeria.

The variables and their categories are:

Variable 1: BSC grade

1. Pass
2. Third Class
3. Lower Division
4. Upper Division
5. First Class

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

Variable 2: State of Origin

1. Indigene
2. Non-indigene

Variable 3: Gender

1. Male

2. Female

Variable 4: Age

1. Under 26
2. 26 and over

The data is presented in the table below

Table 1: Table of data on cross classification of sample of 719 graduated students of Statistics department; Nnamdi Azikiwe University, Awka,Nigeria.

			BSC Grade	BSC Grade	BSC grade	BSC Grade	BSC Grade
Age	Gender	State of Origin	Pass	Third Class	Lower Grade	Upper Grade	First class
Under 26 26&over	Male	Indigene	3 5	41 35	45 17	20 4	1 1
Under 26 26& over	Male	Non-indigene	2 7	69 54	44 29	9 3	1 1
Under 26 26& over	Female	Indigene	I 2	31 31	58 21	32 7	1 2
Under26 26& over	Female	Non-Indigene	3 4	41 37	32 11	9 3	1 1

Source: Retrieved files of graduated students from Statistics department, NAU.

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

When the data was analyzed, we obtained the following table of output

TABLE 2: Summary of the results of model fit for 4-dimensional contingency table data

Model	G^2	χ^2	d.f	p-value
[B,S,G,A]	102.68	111.46	32	0
[S,G,BA]	62.30	65.53	28	0
[G,A,BS]	69.29	69.05	28	0
[S,A,BG]	92.66	94.96	28	0
[B,A,SG]	91.09	94.51	31	0
[B,G,SA]	99.30	103.93	31	0
[B,S,GA]	101.67	108.69	31	0
[G,BA,BS]	28.93	29.46	24	0.2225874
[S,BA,BG]	52.30	51.56	24	0.0007148
[BA,SG]	50.73	49.61	27	0.0037467
[G,BA,SA]	58.94	61.70	27	0.0003621
[S,BA,GA]	61.31	64.02	27	0
[BA,BS,BG]	18.93	19.58	20	0.525909
[BA,BS,SG]	17.36	17.51	23	0.791019
[G,BA,BS,SA]	28.49	28.79	23	0.1976995
[BA,BS,GA]	27.95	28.20	23	0.2178485
[A,BSG]	46.92	45.38	19	0.00366759
[G,BSA]	26.25	25.89	19	0.1235033
[S,BGA]	49.53	48.21	19	0.0001535
[B,SGA]	87.11	86.69	28	0

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

[BSG,BSA]	3.89	3.81	10	0.9529921
[BSG,BGA]	3.79	3.79	10	0.9562268
[BSG,SGA]	42.95	42.21	16	0.0002856
[BSA, SGA]	14.06	13.91	16	0.8945323
[BSGA]	0	0	0	1

P-value for G^2

The result shows that the best model fit which has sufficient evidence to fit the data without loss of information is [BSG, BGA]. This model must have highest p-value and the least likelihood ratio estimate. This model also implies that at 5% level of significance sex is independent of age given Bscgrade and gender.

The best model fit in harmony with the hierarchy principle is written as:

$$\log m_{i_1 i_2 i_3 i_4} = \mu + \mu_{B(i_1)} + \mu_{S(i_2)} + \mu_{G(i_3)} + \mu_{A(i_4)} + \mu_{BS(i_1 i_2)} + \mu_{BG(i_1 i_3)} + \mu_{BA(i_1 i_4)} + \mu_{SG(i_2 i_3)} + \mu_{SA(i_2 i_4)} + \mu_{GA(i_3 i_4)} + \mu_{BSG(i_1 i_2 i_3)} + \mu_{BGA(i_1 i_3 i_4)}$$

TABLE 3: Goodness or Best fit Statistics

Model	Chi Squared & G^2 values	P-value	D.F
[BSG, BGA]	χ^2 =3.792102 G^2 = 3.792395	0.9562386 0.9562268	10 10

CONCLUSION

A generalized method and algorithms were developed for estimation of the parameters and best model fit of log linear model for q-dimensional contingency table. Four-dimensional contingency table was considered for this paper but the same method and algorithms could also be applied for higher dimensional tables. In estimating these parameters and best model fit, computer programs in R were developed for the implementation of the algorithms. The Iterative proportional fitting was used to estimate the parameters and models of the log linear model. The results of the data analysis showed that the best model fit has sufficient evidence to fit the data without loss of

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

information. The model also revealed that sex is independent of age given BScgrade and gender. From table 3, we observed that the results of the goodness of fit showed that the best model adequately fit the data set having highest p-value and the least likelihood ratio estimate.

REFERENCES

[1] Agresti, A. (2002). Categorical Data Analysis. New York. John Wiley & Sons Inc. 2nd edition pg 721-756.
[2] Bartlett , M.S. (1935). Contingency table Interactions. J.Roy. Statist. Soc. Suppl .2, 248-252.
[3] Bishop et al. (1975). Discrete multivariate analysis: Theory and Practice. Cambridge, Mass MIT press Pg 18-37.
[4] Bock, R.D.(1970) . Estimating multinomial response relations in Essays in Probability and Statistics edited by R.C. Bose et al, pg 453-479, Chapel Hill, Univ. of North Carolina Press.
[5] Bock R.D. (1972) .Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 37, 29-51.
[6] Roy, S.N. and Kastenbaum, M.A. (1976). On the hypothesis of no interaction in a multi-way contingency table. Annual Math. Statist. 27, 749-756.
[7] Brown, M.E. (1976). Screening effects in Multi-dimensional contingency table. Applied Statistics, 27, 37-46.
[8] Deming, W.E and Stephan F.F (1940) . On least square adjustment of a frequency tables when the expected marginal totals are known. Annuals of Mathematical stat. 11, 427-444.
[9] Goodman, L.A. (1964b). Simple methods of analyzing three factor interaction contingency tables. J. Dimer.Statist.Assoc.59, 319-352.
[10] Haberman, S.J. (1974). The Analysis of Frequency data. Chicago, Univ. of ChicagoPress.
[11] Jibasen ,D and Lawal H.(2004). Application of log linear models to prison data.Journal Nigerian Statistical Association, 17, 49-58.
[12] Lawal, H.B. (2003) .Categorical data analysis with SAS and SPSS applications. New Jersey. Pg 83-128.
[13] Mc Cullagh, P and Nelder J.A.(1989) .Generalised Linear Models. second Edition. Pg 485-510.

ISSN 2055-0154(Print), ISSN 2055-0162(Online)