LOG LINEAR MODELING

Bolarinwa, I.A. & Bolarinwa, B.T.
Department of Mathematics & Statistics
The Federal Polytechnic, P.M.B. 55, Bida, Nigeria

Abstract

Data on gender, school attended for the National Diploma (ND), ND grades and HND grades of Higher National Diploma (HND) statistics graduates of the Federal Polytechnic, Bida, were examined for presence of association using log linear model. HND grade was found to be associated with ND grade but not associated with school attended for ND. Gender differences were observed in both ND and HND grades. All two-factor interactions but School*HND were found to be significant; so also are all the three-factor interactions but Gender*ND*HND. The need to address gender imbalance in performance was observed.

Keywords: Gender, Log linear, Grade, Interaction.

1. INTRODUCTION

Occasions often arise, when there is need to study association between two or more categorical variables, particularly in the social and medical sciences. One may be interested in studying association between smoking and cardiac arrest or cancer. Association in contingency tables is traditionally investigated by the Chi-square statistic due to Pearson (1900). Later, new method likened to conventional ANOVA was developed. This method of modeling is the log linear modeling.

The log linear model is a specialized case of generalized linear model for Poisson distributed data and is more commonly used for analyzing multidimensional contingency tables that involve more than two variables, although it can be used to analyze two-way contingency tables too (Jeansonne, 2002). A log linear model is similar to the more familiar ANOVA model except that it is applied to the natural logarithm of the expected frequencies (Jibasen, 2004; Lawal, 2003). Response observations in ANOVA are assumed to be continuous normal while in log linear modeling; observations are counts having Poisson distribution (Lawal, 2003).

In the context of log linear modeling, the main effects are usually not of interest (Jibasen, 2004). In fact, Everitt and Dunn (1991) described the main effects parameters in a log linear model as “nuisance parameters”. Bishop, Fienberg, and Holland (1975), Goodman (1964, 1968, 1970, 1971), Haberman (1978), Everitt (1977), Agresti (1996, 2002), Knoke and Burke (1980) are a few of the contributions to the literature on log linear model.

This research is aimed at developing an appropriate log linear model for examining interactions among gender, school attended for National Diploma (ND), ND grade, and Higher National Diploma (HND) grade of HND statistics graduates of the Federal Polytechnic, Bida.

The work is organized as follows: section 2 presents the methods; section 3 presents the results and discussion while the last section presents the conclusion.

2. METHODS

Data

Data on gender, school attended for ND, ND grade, and HND grade were retrieved from files of 424 HND Statistics graduates of the Federal Polytechnic, Bida.

The schools attended for ND were classified into two as follows:

Federal Polytechnic, Bida- Group 1; other schools- Group 2

ND and HND grades were classified as: Pass; Lower credit; Upper credit; Distinction

Model

The log linear model to be considered is the one with four dimensions since four variables are involved. The general log linear model for a contingency table with four variables is given as:

$$\begin{aligned}\log_e m_{ijkl} &= \mu + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{12(ij)} + u_{13(ik)} + u_{14(il)} + u_{23(jk)} + u_{24(jl)} \\ &\quad + u_{34(kl)} + u_{123(ijk)} + u_{124(ijl)} + u_{234(jkl)} + u_{134(ikl)} + u_{1234(ijkl)} \qquad (2.1)\end{aligned}$$

Where

$$\sum u_{1(i)} = \sum u_{2(j)} = \sum u_{3(k)} = \sum u_{4(l)} = 0$$

$$\sum_i u_{12(ij)} = \sum_j u_{12(ij)} = \sum_i u_{13(ik)} = \sum_k u_{13} = \sum_i u_{14(il)} = \sum_l u_{14(il)} = 0$$

$$\sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = \sum_i u_{124(ijl)} = \sum_j u_{124(ijl)} = \sum_l u_{124(ijl)} = 0$$

$$\sum_j u_{234(jkl)} = \sum_k u_{234(jkl)} = \sum_l u_{234(jkl)} = \sum_i u_{134(ikl)} = \sum_k u_{134(ikl)} = \sum_l u_{134Iikl)} = 0$$

$$\sum_i u_{1234(ijkl)} = \sum_j u_{1234(ijkl)} = \sum_k u_{1234(ijkl)} = \sum_l u_{1234(ijkl)} = 0$$

and

µ: overall mean

$u_{1(i)}$: ith level of gender

$u_{2(j)}$: jth level of school attended for ND

$u_{3(k)}$: kth level of ND grade

$u_{4(l)}$: lth level of HND grade

$u_{12(ij)}$: interaction between ith level of gender and jth level of school attended for ND

$u_{123(ijk)}$: interaction between ith level of gender, jth level of school attended for ND and kth level of ND grade.

Other interactions are similarly defined.

The “sum to zero” constraints on the parameters are to ensure that the model contains as many parameters as the number of cells in the table. Such model is called saturated model. Equation (2.1) is therefore a saturated formulation for a four dimensional table.

We shall entertain only hierarchical models. The hierarchical principle emphasizes that whenever a higher order effect is included in a model, all the lower order effects composed from variables in the higher effect are also included (Everitt, 1977). Non-hierarchical models should not be entertained because non-hierarchical modeling does not provide statistical procedure for choosing among potential models (Jeansonne, 2002).

Parameter estimation

The iterative proportional fitting (IPF) algorithm due to Deming and Stephan (1940) is used to estimate model parameters. This is to ensure that expected values are obtained iteratively for model whose expected values are not directly obtainable (from marginal totals of observed values).

Consider the 4-factor model without 4-factor interaction given as:

$$\begin{aligned}\log_{e} m_{ijkl} &= \mu + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{12(ij)} + u_{13(ik)} + u_{14(il)} + u_{23(jk)} + u_{24(jl)} \\ &\quad + u_{34(kl)} + u_{123(ijk)} + u_{124(ijl)} + u_{234(jkl)} + u_{134(ikl)} \qquad (2.2)\end{aligned}$$

Equation (2.2) is an example of a model without direct estimates. This model is hereby used to illustrate the IPF algorithm for estimating expected frequencies (m_{ijkl}), which are inputs to parameter estimation in the fitted model. Estimates of parameters are functions of the logarithms of m_{ijkl} (Everitt, 1977).

Totals $\hat{m}_{ijk.}, \hat{\mathrm{m}}_{\mathrm{ij.l}}, \hat{\mathrm{m}}_{\mathrm{i.kl}}, \text{and}\, \hat{\mathrm{m}}_{\mathrm{.jkl}}$ are constrained to be equal to the corresponding observed marginal totals. To start the IPF procedure, we set initial values $\hat{m}_{ijkl}(0) = 1$ and proceed by adjusting these proportionally to satisfy the first marginal constraint ($\hat{m}_{ijk.} = n_{ijk.}$), calculated from:

$$\hat{m}_{ijkl}(1) = \frac{\hat{m}_{ijkl}(0) n_{ijk.}}{\hat{m}_{ijk.}(0)} \qquad (2.3)$$

Revise expected values $\hat{m}_{ijkl}(1)$ to satisfy the second marginal constraint ($\hat{m}_{ij.l} = n_{ij.l}$) using:

$$\hat{m}_{ijkl}(2) = \frac{\hat{m}_{ijkl}(1) n_{ij.l}}{\hat{m}_{ij.l}(1)} \qquad (2.4)$$

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

$$\hat{m}_{ijkl}(3) = \frac{\hat{m}_{ijkl}(2) n_{i.kl}}{\hat{m}_{i.kl}(2)} \qquad (2.5)$$

Complete the cycle by adjusting $\hat{m}_{ijkl}(3)$ to satisfy the fourth marginal constraint ($\hat{m}_{.jkl} = n_{.jkl}$) using:

$$\hat{m}_{ijkl}(4) = \frac{\hat{m}_{ijkl}(3)n_{.jkl}}{\hat{m}_{.jkl}(3)} \tag{2.6}$$

This four-step cycle is repeated until convergence to the desired accuracy is attained. That is, the process is continued until differences between expected values differ by less than some small amount say 0.01 or 0.0001.

Goodness of Fit Tests

After fitting the model, it becomes imperative to assess the goodness of its fit. This is done by comparing the expected frequencies to the observed cell frequencies for the model. This can be done with a number of statistics. The statistics include Pearson Chi-square (χ^2) due to Pearson (1900); likelihood ratio statistic (G^2) due to Wilks (1938); Neyman modified Chi-square (NM^2) due to Neyman (1949); Freeman Tukey (T^2) due to Freeman and Tukey (1950); modified log likelihood ratio (GM^2) due to Kullback (1959); modified Freeman Tukey (FT) due to Bishop, Fienberg, and Holland (1975).

Comparative studies have suggested preference for use of G^2 statistic over others owing to decomposability into small components and simplicity when comparing two competing models (Lawal, 2003).

The G^2 statistic is given as:

$$G^2 = 2\sum_i n_i \log\left(\frac{n_i}{m_i}\right) \tag{2.7}$$

Where

n_i is the observed frequency and m_i is the expected frequency.

G^2 is Chi-square distributed with degree of freedom (d.f) equal to:

d.f = number of cells in the table – number of independent parameters estimated. That is, d.f is the number of parameters set equal to zero for the purpose of identifiability.

It is possible that more than one model (often the case) is providing a good fit to the data. When that happens, the goodness of fit of any two competing models (A and B, where A is nested within B) can be compared using the quantity:

$$G^2\,(B, A) = G^2(A)-G^2(B)$$

where $G^2(A)$ and $G^2(B)$ are the G^2 values for model A and model B respectively.

If their respective d.f are d.f_A and d.f_B, then G^2 (B, A) is Chi-square distributed with (d.f_A - d.f_B) d.f. If G^2 (B, A) is not significant, then model A is not significantly worse than model B and hence, we would choose the more parsimonious model A.

Model Selection

Many techniques exist in the literature for selecting models. These include: forward selection; backward selection; stepwise procedure; selection based on saturated parameters; selection based on marginal and partial association due to Brown (1976) and Aitkin (1979) method. The backward selection method is however used in this work.

In backward selection, we usually start with the most complex model. Terms are then sequentially deleted from the model. G^2 is computed for each of the current and the reduced model (model resulting from deletion) and using a cut off of predetermined α, say 0.05, we delete the term for which p-value is least significant (term with highest p-value). The process continues until further deletion would lead to a significantly poorer fit.

3. RESULTS AND DISCUSSION

The marginal and partial association tests have suggested the significance of two-factor and three-factor interactions. The partial association tests indicated significance of all two-factor interactions except School*HND (U_{24}) with chi-square value of 3.051 and p-value of .384. It also indicated significance of all three-factor interactions but Gender*ND*HND (U_{134}) with chi-square value of 7.669 and p-value of .568.

The final model obtained through backward elimination procedure has the generating class: Gender*School*ND (U_{123}), Gender*School*HND (U_{124}), School*ND*HND (U_{234}). The goodness of fit statistics are:

Likelihood ratio (G^2) = 8.376; d.f =18; p-value = .972

Pearson Chi-square (χ^2) = 8.157; d.f =18; p-value = .976

Both the G^2 and χ^2 have suggested model adequacy.

The only three-factor interaction (Gender*ND*HND (U_{134})), declared insignificant by the association test would not appear in the final model. The same does not however apply to the insignificant two-factor interaction (School*HND (U_{24})).U_{24} has to be included in the final model in harmony with the hierarchy principle.

The final model is hence:

$$\begin{aligned}\log_e m_{ijkl} &= \mu + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{12(ij)} + u_{13(ik)} + u_{14(il)} + u_{23(jk)} + u_{24(jl)} \\ &\quad + u_{34(kl)} + u_{123(ijk)} + u_{124(ijl)} + u_{234(jkl)} \qquad (2.8)\end{aligned}$$

Although ND grade is associated with HND grade, the association does not depend on gender. Association between gender and ND grade does not depend on HND grade just as association between gender and HND grade does not depend on ND grade. There is gender imbalance in both ND and HND grades. HND grade is found not to be associated with school attended for ND. It is however associated with ND grades. This is indicative of the fact that the HND tutors are not partial in the HND grading. Preference has not been given to HND students who bagged the ND at the Federal Polytechnic, Bida. The major determinant of HND grade is the ND grade, regardless of school attended for the ND. This is an indication that standard does not significantly vary from school to school.

4. CONCLUSION

The research has performed log linear modeling on four factors: Gender, school attended for ND, ND grade and HND grade. All two-factor interactions but School*HND are significant. HND grade is hence not influenced by school attended for the ND but rather by the ND grade - an indication that standard does not significantly vary across schools. All three-factor interactions but Gender*ND*HND are significant. Gender imbalance exists in both ND and HND grades - an issue that deserves attention of the society.

REFERENCES

Aitkin, M. (1979). A simultaneous test procedure for contingency tables. Applied Statistics, 28, 233-242.

Agresti, A. (1996). An introduction to categorical data analysis. New York: John Wiley.

Agresti, A. (2002). Categorical data analysis (2nd ed.). New York: John Wiley.

Bishop, Y.M.M., Fienberg, S.E., & Holland, F.W. (1975). Discrete multivariate analysis: Cambridge: MIT Press.

Brown, M.B. (1976). Screening effects in multi dimensional contingency tables. Applied Statistics, 25, 37-46.

Deming, W.E. & Stephan, F.F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal totals are known. Annals of Mathematical Statistics, 11, 427-444.

Everitt, B.S. (1977). The analysis of contingency tables. London: Chapman and Hall.

Everitt, B.S. & Dunn, G. (1991). Applied multivariate analysis. London: Edward Arnold.

Freeman, M.F. & Tukey, J.W. (1950). Transformation related to the angular and the square root. Annals of Mathematical Statistics, 27, 601-611.

Goodman, L.A. (1964). Interaction in multi-dimensional contingency tables. Annals of Mathematical Statistics, 35, 632-646.

Goodman, L.A. (1968). The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or Without missing entries. Journal of American Statistical Association, 63, 1091-1131.

Goodman, L.A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. Journal of American Statistical Association, 65. 226-256.

Goodman, L.A. (1971). The analysis of multi-dimensional tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics, 13, 33-61.

Haberman, S.J. (1978). Analysis of qualitative data. New York: Academic Press.

Jeansonne, A. (2002). Log linear models. Retrieved October 25, 2009 from http://userwww.sfsu.edu/~efc/classes/bio1710/loglinear/Log%20Linear%20Models.htm

Jibasen, D. (2004). Application of log linear model to prison data. Journal of Nigerian Statistical Association, 17, 49-58.

Kullback, S. (1959). Information theory and statistics. New York: Wiley.

Lawal, H.B. (2003). Categorical data analysis with SAS and SPSS applications. New Jersey: Lawrence Erlbaum.

Neyman, J. (1949). Contribution to the theory of $\chi 2$ test. Proceedings of the First Berkeley Symposium on Mathematical statistics and Probability, 239-273.

Pearson, K. (1900). On the criterion that a given system of deviations in the case of a correlated system of variables is such that it can be reasonably supposed to have to have arisen from random sampling. Philo. Mai. Series, 50, 157-175.

Wilks, S.S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. Annals of Mathematical Statistics, 9, 60-62.