International Journal of Advanced Statistics and Probability, 9 (1) (2021) 18-21

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP

Research paper

Loglinear Modeling of Academic Performance Data

Bushirat Temilola Bolarinwa [1] ,* Ismaila Adewale Bolarinwa [1]

[1] *Department of Statistics, The Federal Polytechnic, P.M.B. 55, Bida, Niger State, Nigeria*
**Corresponding author E-mail: iabolarinwa@gmail.com*

Abstract

The focus of this article was to fit a hierarchical loglinear model to data on academic performance. Data on gender, university attended for B.Sc., B.Sc. and M.Sc. grades of 116 M.Sc. graduates were collected from Department of Statistics, University of Ilorin, Ilorin, Nigeria. Model estimation was carried out by iterative proportional fitting method. Likelihood ratio statistic was utilized for goodness of fit test. The final model generating class contained University, Gender, and B.Sc.*M.Sc., and in harmony with the principle of hierarchy, also contained B.Sc. and M.Sc. grades. Significant interaction was found between B.Sc. and M.Sc. grades only. All other 2-factor and all 3-factor interactions were found not to be significant. Thus, M.Sc. grade was neither associated with gender and university nor with their interaction. The likelihood ratio statistic with p-value of 0.722 suggested model adequacy. The study concluded that only B.Sc. grade was associated with M.Sc. grade obtained by students on graduation. The need to extend study to other departments in the University was recommended.

Keywords: *Academic performance; Association; Contingency table; Likelihood ratio; Loglinear model*

1. Introduction

The need to study association arises often in research endeavors, particularly in social and health sciences, but not limited to them. Contingency table is one of the tools for summarizing categorical data. Association in contingency tables is traditionally studied using chi-square statistic, but it can only handle two-way contingency table. However, over time, a new approach to analysis by Goodman [1], [2] capable of analyzing multi-way contingency tables evolved. This method is similar to ANOVA and it is called loglinear model. Associations in loglinear models are analogous to interactions in ANOVA models; in loglinear models, the main effects are usually not of interest and are in fact, described as nuisance parameters by Everitt and Dunn [3]. Loglinear models are commonly used to analyze multi-way contingency tables that involve more than two variables. They are a little different from other modeling methods in that they are applied to the natural logarithm of the expected frequencies.

The technique of loglinear analysis is similar to that of ANOVA for factor-response variables that have continuous distribution. The distinction between them is that in the latter, the response observations are assumed to be continuous normal while in the former, the response observations are counts with Poisson distributions (Lawal [4]). In loglinear analysis, all variables are treated as response variable and the interest is on statistical independence and dependence (Stokes, Davies & Koch [5]). The goal of loglinear modeling is to fit the best loglinear model that is parsimonious. The model does not only specify how expected frequencies depend on the levels of the categorical variables involved in the modeling exercise, it also describes association between the variables (Brzenzinka [6]). Brzenzinka further observes that treating ordered categorical variables as ordinal rather than nominal has a number of advantages.

Loglinear models have been applied in articles too numerous to mention. Mehdizadeh [7], Marascuilo and Busk [8] and Ting and Abella [9] applied loglinear model in measuring student course evaluations; Odetunmibi, Adejumo, and Sanni [10] and Odetunmibi, Adejumo and Anake [11] are applications in medical research.

The aim of this research is to fit hierarchical loglinear model to academic performance data. The article is organized as follows: Section 2 presents the Methodology, Section 3, the Results and Discussion while the last section concludes the article.

2. Methodology

This section presents data collection, model, model estimation, and goodness of fit tests.

2.1. Data

Data are gender, university attended for B.Sc., B.Sc. grade, and M.Sc. grade of 116 M.Sc. Statistics graduates of University of Ilorin, Nigeria.

Gender is classified as: Male and female. Male is coded 0 while Female is coded 1.

University attended is classified as follows: Group 1 for University of Ilorin, Group 2 for other universities. University of Ilorin is coded 1 while “other universities” is coded 0.

B.Sc. Grade is classified as: First Class, Second Class Upper, and Second Class Lower. Second Class Lower is coded 1, Second Class Upper is coded 2 while First Class is coded 3.

M.Sc. Grade is classified as: Terminal, M.Phil./Ph.D grade and Ph.D. grade. Terminal grade is coded 1, M.Phil./Ph.D. grade is coded 2 and Ph.D. grade is coded 3.

Model

The loglinear model under consideration is one with 4 variables and it is of form

$$\begin{aligned}\ln(\hat{m}_{ijkl}) = \mu + \lambda_i^G + \lambda_j^U + \lambda_k^B + \lambda_l^M + \lambda_{ij}^{GU} + \lambda_{ik}^{GB} + \lambda_{il}^{GM} + \lambda_{jk}^{UB} \\ + \lambda_{jl}^{UM} + \lambda_{kl}^{BM} + \lambda_{ijk}^{GUB} + \lambda_{ijl}^{GUM} + \lambda_{jkl}^{UBM} + \lambda_{ikl}^{GBM} + \lambda_{ijkl}^{GUBM}\end{aligned}$$

where

$$\sum \lambda_i^G = \sum \lambda_j^U = \lambda_k^B = \lambda_l^M = 0$$

$$\sum_i \lambda_{ij}^{GU} = \sum_j \lambda_{ij}^{GU} = \sum_i \lambda_{ik}^{GB} = \sum_k \lambda_{ik}^{GB} = \sum_i \lambda_{il}^{GM} = \sum_l \lambda_{il}^{GM} = 0$$

$$\sum_j \lambda_{jk}^{UB} = \sum_k \lambda_{jk}^{UB} = \sum_j \lambda_{jl}^{UM} = \sum_l \lambda_{jl}^{UM} = \sum_k \lambda_{kl}^{BM} = \sum_l \lambda_{kl}^{BM} = 0$$

$$\sum_i \lambda_{ijk}^{GUB} = \sum_j \lambda_{ijk}^{GUB} = \sum_k \lambda_{ijk}^{GUB} = \sum_i \lambda_{ijl}^{GUM} = \sum_j \lambda_{ijl}^{GUM} = \sum_l \lambda_{ijl}^{GUM} = 0$$

$$\sum_j \lambda_{jkl}^{UBM} = \sum_k \lambda_{jkl}^{UBM} = \sum_l \lambda_{jkl}^{UBM} = \sum_i \lambda_{ikl}^{GBM} \sum_k \lambda_{ikl}^{GBM} \sum_l \lambda_{ikl}^{GBM} = 0$$

$$\sum_i \lambda_{ijkl}^{GUBM} = \sum_j \lambda_{ijkl}^{GUBM} = \sum_k \lambda_{ijkl}^{GUBM} = \sum_l \lambda_{ijkl}^{GUBM} = 0$$

and

μ : *overall mean*

λ_i^G : *ith level of gender*

λ_j^U : *jth level of* university attended for B.Sc.

λ_k^B : *kth level of* B.Sc. Grade

λ_l^M : *lth level of* M.Sc. Grade

λ_{ij}^{GU} : *Interaction* between *ith level of gender* and jth level of university attended for B.Sc.

λ_{ijk}^{GUB} : *Interaction* between *ith level of gender* and jth level of university attended for B.Sc. and kth level of B.Sc. Grade

Other interactions are similarly defined.

2.2. Model estimation

Mode was estimated by maximum likelihood estimation method. To ensure that some expected values that were not directly obtainable from marginal totals of observed values (that did not have closed forms) could be obtained, iterative proportional fitting (IPF) method due Deming and Stephan [12] was used.
To illustrate the IPF principle, consider a 4-factor model without 4-factor interaction (an example of a model without direct estimates) given below

$$\ln(\hat{m}_{ijkl}) = \mu + \lambda_i^G + \lambda_j^U + \lambda_k^B + \lambda_l^M + \lambda_{ij}^{GU} + \lambda_{ik}^{GB} + \lambda_{il}^{GM} + \lambda_{jk}^{UB}$$
$$+ \lambda_{jl}^{UM} + \lambda_{kl}^{BM} + \lambda_{ijk}^{GUB} + \lambda_{ijl}^{GUM} + \lambda_{jkl}^{UBM} + \lambda_{ikl}^{GBM}$$

The estimates of expected frequencies (m_{ijkl}) which are inputs to parameter estimation in fitted model are iteratively obtained as follows:
Totals $\hat{m}_{ijk.}$, $\hat{m}_{ij.l}$, $\hat{m}_{i.kl}$, *and* $\hat{m}_{.jkl}$ are constrained to equal corresponding observed marginal totals. The IPF procedure begins by setting 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:

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

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

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

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

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

The cycle is completed by revising $\hat{m}_{ijkl}(3)$ to satisfy the fourth marginal constraint, $\hat{m}_{.jkl} = n_{.jkl}$, using

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

The four-step cycle is repeated until convergence to the desired accuracy, say, 0.01 or 0.001 is attained.

2.3. Goodness of fit tests

Methods that exist for assessing goodness of fit of models include but not limited to the following: Pearson Chi-square statistic by to Pearson [13], likelihood ratio statistic (G^2) by Wilks [14], Neyman modified Chi-square (NM^2) due to Neyman [15], and Freeman Tukey ((T^2) due to Freeman and Tukey [16]. According to [14], studies have suggested preference for G^2 statistic over others.
The G^2 statistic is

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

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 number of cells in the table less number of independent parameters estimated. That is, d.f is number of parameters set equal to zero for identifiability purpose.
If more than one model provides good fit to the data, it is logical to pick the more parsimonious one. The goodness of fit of two competing models say A and B (where A is nested in 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 G^2 values for Models A and B respectively. Given that the d.f for Models A and B are respectively, $d.f_A$ and $d.f_B$, then, $G^2(B, A)$ is Chi-square distributed with d.f. ($d.f_A$ - $d.f_B$). $G^2(B, A)$, not being significant implies that Model A is not significantly worse than Model B and therefore, the more parsimonious Model A is chosen.

3. Results and Discussion

The final model contains University, Gender, and B.Sc.*M.Sc., and in harmony with the principle of hierarchy, also contains B.Sc. and M.Sc. grades. The final model is hence,

$$\ln(\hat{m}_{ijkl}) = \mu + \lambda_i^G + \lambda_j^U + \lambda_k^B + \lambda_l^M + \lambda_{kl}^{BM}$$

Where

$$\sum \lambda_i^G = \sum \lambda_j^U = \sum \lambda_k^B = \sum \lambda_l^M = 0$$

$$\sum_k \lambda_{kl}^{BM} = \sum_l \lambda_{kl}^{BM} = 0$$

The only interaction in the final model is B.Sc.*M.Sc., this implies that B.Sc. grade is associated with M.Sc. grade obtained, regardless of the university from which the B.Sc. was obtained and gender. This gives credibility to the results obtained at the M.Sc. level in the Department of Statistics, University of Ilorin, Nigeria, as intellectual ability of the M.Sc. student represented by B.Sc. grade has proved to be the main determinant of the outcome in the M.Sc. Progamme.

It also indicates that M.Sc. students that bagged their B.Sc. from the University of Ilorin have not been enjoying special favor since there is no interaction between M.Sc. grade and university attended. It may also imply that quality of B.Sc. grade does not significantly vary from one university to another. All other 2-factor and all 3-factor interactions are found not to be significant. Thus, M.Sc. grade is neither associated with gender and university nor with their interaction.

The likelihood ratio statistic with p-value of 0.722 suggests adequacy of the fitted loglinear model.

4. Conclusion

This article has modeled academic performance data of M.Sc. graduates of the Department of Statistics, University of Ilorin, Ilorin, Nigeria using the loglinear approach. Significant interaction is found between B.Sc. and M.Sc. grades only. Significant interaction was found between B.Sc. and M.Sc. grades only. The need to extend study to other departments in the University is recommended.

References

[1] L.A. Goodman, The multivariate analysis of qualitative data interactions among multiple classifications, Journal of American Statistical Association 65(1970) 226-256. https://doi.org/10.1080/01621459.1970.10481076.

[2] L.A. Goodman, 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(1974) 1091-1131. https://doi.org/10.1080/01621459.1968.10480916.

[3] B.S. Everitt, G. Dunn, Applied Multivariate Analysis, Edward Arnold, London, 1991.

[4] B. Lawal, Categorical Data Analysis with SAS & SPSS Applications, Lawrence Erlbaum New Jersey, 2003. https://doi.org/10.4324/9781410609168.

[5] M. Stokes, C.S. David, G.G. Koch, Categorical Data Analysis, 2nd ed., SAS Institute and Wiley, North Carolina, 2003.

[6] J. Brzezińska, Ordinal log-linear models for contingency tables, Folia Oeconomica, (2016) https://doi.org/10.1515/foli-2016-0017.

[7] M. Mehdizadeh, Loglinear models and student course evaluations, Research in Economic Education 21(1) (1990) 7-21. https://doi.org/10.1080/00220485.1990.10844649.

[8] L.A. Marascuilo, P.L. Busk, Loglinear models: A way to study main effects and interactions for multidimensional contingency tables with categorical data, Journal of Counseling Psychology *34*(4) (1987) 443–455. https://doi.org/10.1037/0022-0167.34.4.443.

[9] D.H. Ting, M.S. Abella, Measuring student course evaluations: The use of a loglinear model, International Education Journal 8(1) (2007) 194-204.

[10] O.A. Odetunmibi, A.O. Adejumo, O.O.M. Sanni, Loglinear modelling of cancer patients cases in Nigeria: An exploratory study approach, Open Science Journal of Statistics and Application 1(1) (2013) 1–7.

[11] O.A. Odetunmibi, A.O. Adejumo, T.A. Anake, Log-Linear modelling of effect of age and gender on the spread of Hepatitis B virus infection in Lagos State, Nigeria, Open Access Maced J Med Sci. 7(13) (2019) 2204–2207. https://doi.org/10.3889/oamjms.2019.573.

[12] W.E. Deming, F.F. Stephan, On a least squares adjustment of a sample frequency table when the expected marginal totals are known, Annals of Mathematical Statistics 11 (1940) 427-444. https://doi.org/10.1214/aoms/1177731829.

[13] K. Pearson, On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philo. Mag. Series 5(50) (1900) 157-175. https://doi.org/10.1080/14786440009463897.

[14] S.S. Wilks,. The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Statist. 9 (1938) 60-62. https://doi.org/10.1214/aoms/1177732360.

[15] J. Neyman, Contribution to the theory of the χ^2 test, Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability (1949) 239-273.

[16] F. Freeman, J.W. Tukey, Significance levels for a k-sample slippage test, Annals of Mathematical Statistics 21(4) (1950) 607-611. https://doi.org/10.1214/aoms/1177729756.