International Journal of Statistics and Applications 2012, 2(5): 56-59
DOI: 10.5923/j.statistics.20120205.02

Analysis of Categorical Panel Data

A. O. Adejumo*, O. O. M. Sanni, E. T. Jolayemi, R. O. Ogedengbe

Department of Statistics University of Ilorin, Ilorin, Nigeria

Abstracts In some categorical tables, one of the classifying variables may be at least ordinal (ranked) arising from a follow-up or any similar study. The other classifying variable(s) may be that which separates the population into groups using variables such as gender, race or location, or a combination of some of them. The counts obtained this way are analyzed recognizing that one of the variables is nearly metric and must be used and interpretation becomes easier when appropriate model is fitted to the arising product multinomial. An example of such an approach is provided using the data from Tuberculosis Management in a Teaching Hospital. We observed that the recovery rate of females was faster than their males counterpart on the assumption that those discharged through management system follows an exponential distribution.

Keywords Panel data, Categorical data, Multinomial

1. Introduction

Categorical data are obtained when the variables which are discrete in nature are cross-classified and subjects having the same levels of the cross-classification are aggregated to form counts. Clearly such variables are at most ordinal in nature. Variables that are purely metric are reduced appropriately for categorical data analysis to be effected. In a follow-up (longitudinal) study the progression of positive outcome is critical and should be examined.

Cross-classified data can have any of full-multinomial, hypergeometric, independent Poisson or product multinomial distributions, Bishop, Feinberg and Holland[1], Agresti[2], Sanni and Jolayemi[3], Adejumo[4] among many authors. All these distributions have fixed, but unknown, parameters. Each underlying distribution is dictated by the sampling scheme, even though the parameter estimates within each are identical as demonstrated by Birch[5], see also Jolayemi et al[6]. It is possible, however, that the parameters involved in the categorical data, have a specific pattern, especially when one or more of the categorical variables are metric but of constant interval. A statistical analysis approach for such data may be appropriate to use some models for probability outcomes. The model used, if appropriate can then be used to determine termination of management. This approach is in focus in this work.

In this research, the main objective is to examine a model fitting-algorithm for a longitudinal categorical data.

The follow-up data of this form becomes a panel data if the period for reassessment is constant.

* Corresponding author:
aodejumo@unilorin.edu.ng (A. O. Adejumo)
Published online at http://journal.sapub.org/statistics

2. Methodology

Consider an r x c contingency table. The row (r) categories are the sup-populations to be compared and column (c) equals the number of possible follow-ups. Let the matrix of observation be represented by

$$n_{r\times c} = (n_1, n_2, ..., n_r)^1,$$

where

$$\underset{\sim}{n_i}^1 = (n_{i1}, n_{i2}, ..., n_{ic}) \tag{1}$$

Within the foregoing, assume the product multinomial distribution for $n_{r\times c}$. Thus

$$n_{r\times c} \sim \prod_{i=1}^{r} M_n(n_i, P_i) \tag{2}$$

where n_i is as represented in 1.1,

$$n_{i.} = \sum_{j=1}^{c} n_{ij}, \qquad n_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$$

and

$$P = P_{rxc} = (P_1, P_2, \ldots, P_r)^{|}$$

such that

$$P_{i.} = (P_{i1}, P_{i2}, \ldots P_{ic})$$

and

$$\sum_{j=1}^{c} P_{ij} = 1 \qquad \forall i \tag{3}$$

Furthermore, assume that for each i, the vector P_i has a known or suspected pattern $f_i(\theta_i)$. The mixture model is with a compelling assumption if each $f_i(\theta_i)$ is unique, see

for example Brooks et al.[7], when the variable characterizing the column is ordinal.

The main aim of this study is to test some hypothesis regarding $f_i(\theta_i)$. In particular, we assume that $f_i(\theta_i)$ is exponential in this research paper with parameters

$$\theta_i^1 = (\alpha_i, \beta_i)$$

In this formulation,

$$P_i \sim f_i(\theta_i) = e^{\alpha_i + j\beta_i} \qquad i = 1,2,....,r \qquad (4)$$

where *j=1,2, ..., c*; indicating the outcome of the column variable. If $\beta_i < 0$, the probability reduces over j (usually indexing time) or over jth follow-up time of constant period. What may be of interest here are various hypotheses regarding θ_i. Some of these include.

(i) $\alpha_i = \alpha \quad \forall i$ which represents all r rows are identical before follow-up

(ii) $\beta_i = \beta \quad \forall i$ which can be interpreted to be identical reactions of the r subpopulation for the intervention of the follow-up.

(iii) $\theta_i = \theta \quad \forall i$ is the combination of (i) and (ii) above.

Note that other forms of $f_i(\theta_i)$ are possible. Such other forms includes $f_i(\theta_i) = e^{\alpha_i + j\beta_{1i} + j^2\beta_{2i}}$ which is essentially used when the response is quadratic. It is also used for studying medical intervention.

Let $L(n_{rxc}, P_{r\times c})$ be the likelihood function for $n_{r\times c}$.

Then,

$$L(n_{rxc}, P_{r\times c}) = \prod_{i=1}^{r} M_r(n_{i.}, P_i)$$

$$= \prod_{i=1}^{r}\prod_{j=1}^{c}\binom{n_{i.}}{n_{ij}}\left(e^{\alpha_i + j\beta_j}\right)^{n_{ij}}$$

so that the log likelihood L under the constraint in equation (1) is given by

$$L = \sum_{i=1}^{r}\sum_{j=1}^{c}\log\binom{n_{i.}}{n_{ij}} + \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}(\alpha_i + j\beta_i) + \lambda\left(\sum_{j=1}^{c} e^{\alpha_i + j\beta_i} - 1\right) \qquad (5)$$

where λ is the Langrange multiplier (indicating the boundary limit). Clearly the log likelihood of equation (5) does not give normal equations which are linear in the parameters, see McCulagh and Nelder[8], Jolayemi and Okoro[9] for example.

Let $L_{H_o} = L(n_{rxc}, P_{r\times c} \mid H_o)$ be the likelihood function using estimation $P_{r\times c}$ under the null hypothesis and $L_\Omega = L(n_{rxc}, P_{r\times c} \mid \Omega)$ be the similar likelihood under the parameter space.

The Likelihood ratio test statistic can be obtained from

$$\Delta = L_{H_o} \big/ L_\Omega$$

Under some regularity conditions, see Bickel et al.[10] and Adejumo[4],

$$-2\log\Delta$$

has the chi-square distribution with (k-m) degrees of freedom, where k and m are the number of parameters estimated under Ω and Ho respectively.

2.1. Estimation of Parameters

First consider the log likelihood function of equation (5) and let the null hypothesis H_o be given by

$$H_o : \theta_i = \theta$$

This is equivalent to

$$H_o : \theta_i = \theta = \begin{cases} \alpha_1 = \alpha_2 = ... = \alpha_r = \alpha \\ \beta_1 = \beta_2 = ... = \beta_r = \beta \end{cases}$$

which represents gender insensitivity. Other forms of H_o can be used.

The likelihood function L_{Ho} is given by

$$L_{H_o} = \sum_{i=1}^{r}\sum_{j=1}^{c}\log\binom{n_{i.}}{n_{ij}} + \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}(\alpha + j\beta) + \lambda\left(\sum_{j=1}^{c} e^{\alpha + j\beta} - 1\right) \qquad (6)$$

The normal equations from equation (6) are obtained as follows:

$$\frac{\delta L_{H_o}}{\delta\alpha} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} + \lambda\sum_{j=1}^{c} e^{\alpha + j\beta} = 0 \qquad (7)$$

$$\frac{\delta L_{H_o}}{\delta\beta} = \sum_{i=1}^{r}\sum_{j=1}^{c} jn_{ij} + \lambda\sum_{j=1}^{c} je^{\alpha + j\beta} = 0 \qquad (8)$$

$$\frac{\delta L_{H_o}}{\delta\lambda} = \sum_{j=1}^{c} e^{\alpha + j\beta} - 1 = 0 \qquad (9)$$

From equation (7) it is clear that $\lambda = -n_{..}$

Thus equations (7) and (8) become respectively

$$\frac{\delta L_{H_o}}{\delta\alpha} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} - n_{..}\sum_{j=1}^{c} e^{\alpha + j\beta} = 0 \qquad (10)$$

$$\frac{\delta L_{H_o}}{\delta\beta} = \sum_{i=1}^{r}\sum_{j=1}^{c} jn_{ij} - n_{..}\sum_{j=1}^{c} je^{\alpha + j\beta} = 0 \qquad (11)$$

Let S_{2x1} represent the vector of normal equations.

Then S is given by

$$S = \begin{pmatrix} \dfrac{\delta L_{H_0}}{\delta\alpha} \\ \dfrac{\delta L_{H_0}}{\delta\beta} \end{pmatrix}$$

So that the Hessian matrix, Morisson[11] is given by

$$\frac{\delta S}{\delta \theta} = \begin{pmatrix} \frac{\delta^2 L}{\delta \alpha^2} & \frac{\delta^2 L}{\delta \alpha \delta \beta} \\ \frac{\delta^2 L}{\delta \alpha \delta \beta} & \frac{\delta^2 L}{\delta \beta^2} \end{pmatrix} \quad (12)$$

Let $E(\underset{\sim}{\theta}) = \theta_o$, then by mid-value theorem: Mood et al.[12]

$$S(\underset{\sim}{\theta}) = S(\theta_o) + \frac{\delta S(\theta_o)}{\delta \theta}(\underset{\sim}{\theta} - \theta_o) = 0 \quad (13)$$

so that the θ can be obtained as

$$\underset{\sim}{\theta} = \theta_o - \left(\frac{\delta S(\theta_o)}{\delta \theta}\right)^{-1} S(\theta_o)$$

An iterative procedure is then used to obtain an estimate $(\hat{\underset{\sim}{\theta}})$ *for* $\underset{\sim}{\theta}$ using an initial vector $\theta_o = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and tolerance

$$\underset{\theta}{Max}\left|\underset{\sim}{\theta}^{k-1} - \underset{\sim}{\theta}^{k}\right| < \delta \quad (14)$$

It is easy to note that the vector $S(\theta)$ is given by

$$S(\underset{\sim}{\theta}) = \begin{bmatrix} \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} & -n_{i.}\sum_{j=1}^{c} e^{\alpha + j\beta} \\ \sum_{i=1}^{r}\sum_{j=1}^{c} j n_{ij} & -n_{i.}\sum_{j=1}^{c} j e^{\alpha + j\beta} \end{bmatrix}$$

And the cell values of the Hessian matrix is given by

$$\frac{\delta^2 L}{\delta \alpha^2} = -n_{i.}\sum_{j=1}^{c} e^{\alpha + j\beta}$$

$$\frac{\delta^2 L}{\delta \alpha \delta \beta} = -n_{i.}\sum_{j=1}^{c} j e^{\alpha + j\beta}$$

$$= \frac{\delta^2 L}{\delta \beta \delta \alpha}$$

$$\frac{\delta^2 L}{\delta \beta^2} = -n_{i.}\sum_{j=1}^{c} j^2 e^{\alpha + j\beta}$$

Under Ω, the above procedure is obtained for each i. Thus $S_i(\theta_i)$

$$S_i(\theta_i) = \begin{bmatrix} \sum_{j=1}^{c} n_{ij} & -n_{i.}\sum_{j=1}^{c} e^{\alpha_i + j\beta_i} \\ \sum_{j=1}^{c} j n_{ij} & -n_{i.}\sum_{j=1}^{c} j e^{\alpha_i + j\beta_i} \end{bmatrix}$$

$$\frac{\delta^2 L}{\delta \alpha^2} = -n_{i.}\sum_{j=1}^{c} e^{\alpha_i + j\beta_i}$$

$$\frac{\delta^2 L}{\delta \alpha_i \delta \beta_i} = -n_{i.}\sum_{j=1}^{c} j e^{\alpha_i + j\beta_i}$$

$$\frac{\delta^2 L}{\delta \beta_i^2} = -n_{i.}\sum_{j=1}^{c} j^2 e^{\alpha_i + j\beta_i}$$

Finally, the estimate $\hat{\theta}_i$ *of* θ_i is obtained as explained earlier. The test statistic in this case which is the likelihood ratio test statistic is given by

$$\Delta = \frac{L(n_{r\times c}; \hat{P}_{r\times c} \mid H_o)}{L(n_{r\times c}; \hat{P}_{r\times c} \mid \Omega)}$$

So that $-2\log\Delta \sim \chi_d^2$ where d = (k - m) degrees of freedom. (k - m) = (2r - 2)=2(r - 1).

If $\hat{\underset{\sim}{\theta}}$ under H_o is given as

$$\hat{\theta}_{H_o} = \begin{pmatrix} \hat{\alpha}_o \\ \hat{\beta}_o \end{pmatrix}$$

And θ under Ω is given as

$$\hat{\theta}_{\Omega} = \begin{pmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{pmatrix}$$

Then

$$\Delta = \frac{\prod_{i=1}^{r}\prod_{j=1}^{c}\binom{n_{i.}}{n_{ij}}\left(e^{\hat{\alpha}_o + j\hat{\beta}_o}\right)^{n_{ij}}}{\prod_{i=1}^{r}\prod_{j=1}^{c}\binom{n_{i.}}{n_{ij}}\left(e^{\hat{\alpha}_i + j\hat{\beta}_i}\right)^{n_{ij}}}$$

$$= \prod_{i=1}^{r}\prod_{j=1}^{c}\left(e^{(\hat{\alpha}_o - \hat{\alpha}_i) + j(\hat{\beta}_o - \hat{\beta}_i)}\right)^{n_{ij}}$$

$$\equiv \prod_{i=1}^{r}\prod_{j=1}^{c}\left\{\exp[(\hat{\alpha}_o - \hat{\alpha}_i) + j(\hat{\beta}_o - \hat{\beta}_i)]\right\}^{n_{ij}}$$

And $-2log\Delta$ is given as

$$-2\log\Delta = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\left\{(\hat{\alpha}_i - \hat{\alpha}_o) + j(\hat{\beta}_i - \hat{\beta}_o)\right\}$$

The above is a demonstration of how to produce software to perform the process for execution.

3. Empirical Results, Discussions and Conclusions

The method of application of mixture models for the 2-dimensional categorical data is demonstrated using a data set from a disease management from a hospital, the Univer-

sity of Ilorin Teaching Hospital (UITH), Nigeria, spanning the period between 1996 and 1998 on the management of Tuberculosis patients. The data excluded those who were lost to follow-up, so that, those who were successfully discharged were considered in Table 1 using approximated periods.

Table 1. 109 Tuberculosis patients classified by length of treatment and gender using one month follow up interval

Duration (in month)	1	2	3	4	5	6	Total
No. of Male	44	17	6	3	3	1	74
No. of Female	22	8	4	1	0	0	35
Total	66	25	10	4	3	1	109

The analysis of the data followed equation (2) and the imposed models in equation (4). Using the tolerance limit $\delta = 0.001$ for maximum difference in the parameter estimates as dictated by

$$\underset{\theta}{Max}\left|\underset{\sim}{\theta}^{k-1} - \underset{\sim}{\theta}^{k}\right| < \delta$$

of equation (13), the following estimates were obtained:

$$\hat{\theta}_o = \begin{pmatrix}\hat{\alpha}_o \\ \hat{\beta}_o\end{pmatrix} = \begin{pmatrix}-0.3526 \\ -0.8756\end{pmatrix}$$

$$\hat{\theta}_1 = \begin{pmatrix}\hat{\alpha}_1 \\ \hat{\beta}_1\end{pmatrix} = \begin{pmatrix}-0.1478 \\ -0.8167\end{pmatrix}$$

$$\hat{\theta}_2 = \begin{pmatrix}\hat{\alpha}_2 \\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix}-0.5566 \\ -1.0235\end{pmatrix}$$

The likelihood ratio test statistic for $\theta_o \forall i$ of $G^2 = -2log\Delta = 15.24$ with d=2 degrees of freedom with p-value of 0.001 provided a bad fit

This implies that a uniform distribution cannot be used for both males and females. Consequently, different models existed for males and females which were θ_1 and θ_2. This showed that the period of treatment was gender sensitive. While males would be treated for seven months the female counterpart would be treated for 4 months.

REFERENCES

[1] Bishop, Y. M. M, Feinberg, S. E. Holland, P. W. (1975). Discrete Multivariate Analysis. Cambridge MA; MIT Press.

[2] Agresti, A. (2002). Categorical Data Analysis. John Wiley and Sons. 2nd Edition. New York.

[3] Sanni, O. O. M. and Jolayemi, E. T. (1998). On the use of some Categorical Test Statistic on Sparse Contingency Table. Journal of Pure and Applied Science. 509 – 514.

[4] Adejumo, A. O (2005). Modelling Generalized Linear (Log-linear) Models for Raters Agreement measures; Peter Lang, Frankfurt am Main. (http://www.peterlang.de)

[5] Birch, M. W. (1966). Maximum Likelihood in Three Way contingency Table. J. Royal Statistics Society, Series B25, 220 – 233.

[6] Jolayemi, E. T. and Brown, M. B. (1984). The Choice of a log-linear model using Cp-type Statistics. Computational Statistics and Data Analysis.

[7] Brooks, S. P, Morgan, B. J. T, Riobut, M. S, and Peak, S.C. (1997). Finite Mixture Models for Proportions. Biometric, 53; 1097 – 1115.

[8] McCullagh, P. and Nelder, J. A. (1989). Generalised Linear Models. Chapman and Hall. New York.

[9] Jolayemi, E. T. and Okoro, E. O (1995). On the estimation of mean IC50. Biosciences Research Communication, 7, 175 – 178.

[10] Bickel, P. J. and Doksum, J. A. (1973). Mathematical Statistics. Holden Day, San Francisco.

[11] Morrison, D. (1976). Multivariate Statistics Methods. McGraw Hill, New York.

[12] Mood, A. M., Graybill, F. A., and Boes, D. C. (1963). Introduction to the Theory of Statistics. McGraw Hill, New York.