SUST
Journal of Natural and Medical Sciences

Journal homepage: http://journals.sustech.edu/

Analysing Ordered Categorical Data: Various Methods to Attempt and Discrepancies in Interpretations to Accept

Amin I. Adam

Associate Professor in Applied Statistics,
Dept. of Statistics, Faculty of Economics and Political Sciences,
Omdurman Islamic University, Omdurman, P. O. Box 382, Sudan
Coressponding Authoer**:**

Article history: Recieved: October 2010
Accepted: Julay 2011

ABSTRACT

This study concerns with a number of methods that are used to analyze ordered categorical data, the data which are related to variables having a set of finite and distinctive classifications. This type of data is used extensively in all fields, especially in sociology, education, psychology, medicine, and biology. In most situations the categories of these data are natural and sometimes are ordered. A number of methods have been reviewed and applied in this context, ranging from the chi-squared test of independence to log-linear and logit models. In addition to compare the theoretical background of these techniques, the study compares the results given by these methods explaining areas where they lead to similar conclusions and where they lead to different interpretations. The study gives empirical estimates of the effects between the variables, explains estimates corresponding to the ordering nature of the variables, and, more importantly, takes into account whether one variable is to be treated as a response. The data for this research are collected from a random sample of 300 students at the Teachers College of Makkah and the concern is devoted to the academic performance of students and associated factors. The data of the study are analysed by SPSS (Statistical Package for the Social Sciences).

KEYWORDS: Categorical data, Chi-squared test, Log-linear model, Logit model.

المستخلص

هذا البحث يعنى بدراسة الطرق المختلفة لتحليل البيانات الفئوية المرتبة التي ترتبط بمتغيرات لها عدد محدود من الفئات المرتبة غير المتداخلة. وهذا النوع من البيانات يستعمل كثيراً في كل المجالات لا سيما الاجتماعية والتربوية والنفسية والطبية والبيولوجية. وفي كثير من الأحيان تكون فئات متغيرات هذه البيانات طبيعية وبعضها تكون مرتبة. تناول هذا البحث كيفية تحليل هذه البيانات باستخدام وتطبيق عدد من الطرق الإحصائية ابتداءً باختبار مربع كاي مروراً بالنموذج الخطي اللوغريثمي وانتهاءً بالنموذج اللوجستي. وقدم البحث تحليلاً وتفسيراً ومقارنة لنتائج هذه الطرق مبيناً أماكن اختلاف النتائج المستخلصة. وقد استخدم الباحث بيانات تم جمعها من طلاب الكلية الجامعية بجامعة أم القرى بمكة المكرمة، وتتعلق بمستوي التحصيل الأكاديمي وما يرتبط بذلك من

عوامل واستخدم في التحليل الحزمة الإحصائية للعلوم الاجتماعية SPSS.

INTRODUCTION

Ordinal categorical variables with levels range from high to low or the other way round are more common and well known, especially in social and biological sciences. Ordinal variables may result from grouping values of interval variables such as income or education, or they may arise when ordering, but not distance, between categories can be established. Although a category like "very satisfied" is more than the category "satisfied", the actual distance between the two response categories can't be determined.

Our objective is to illustrate a number of techniques that could be used for the analysis of this type of data, and to see whether similar or different results can be reached by these techniques. The data are collected from a random sample of 300 students at the Teachers College of Makkah and the variable concerned is their academic performance which is classified as poor, moderate, and high.

The Chi-squared test of independence:-

The ordinary chi-square test of independence, which used very often as a response of investigators faced with such data, is based on the product of the marginal probabilities.

The data in table(1) shows the cross-classification of the students in the sample according to their academic performance and the educational level of father. The classification of the academic performance is based on the cumulative rate, that is, a rate more than 3.75 is considered high, and a rate less than 2.50 is considered poor, whereas a rate between 2.50 and 3.75 is considered moderate. The categories of the education level of father, on the other hand, are: intermediate or less, secondary, university, and higher education. It clear that the categories of the academic performance are ordered and so the categories of the educational level of father.

Table (1): Cross-classification of academic performance and education of father

Count

		father education				
		intermediate education or <	secondary education	university education	higher education	Total
performance	poor	49	38	39	13	139
	moderate	21	37	32	12	102
	high	12	13	16	18	59
Total		82	88	87	43	300

The chi-squared statistics for testing independence of the academic performance and education of father are the Pearson's X^2=22.84 and the Likelihood ratio G^2=20.52, based on df=6. The corresponding p-values are 0.001 and 0.002, respectively. Hence the two variables are not independent, i.e., there is a statistical evidence for a relationship between the academic performance of a students and the educational level of his father.

Scoring method:-

This technique deals with assigning (arbitrary) scores to the categories of each variable and then using normal regression techniques on these values. For our data, we chose this to be evenly spaced for the two variables, but they need not be so. Hence the scores for the academic performance are: -1, 0, and 1; whereas for the educational level of father they are chosen to be: -3, -1, 1, and 3. We then calculate the linear regression of the academic performance (Y) on the educational level of father (X), i.e. Y=α + β X, using the formula for the coefficient β given by

$$\beta = \frac{n\sum_i\sum_j n_{ij}x_i y_j - (\sum_i n_i x_i)(\sum_j n_j y_j)}{\sum_i n_i x_i^2 - (\sum_i n_i x_i)^2}$$

where n_{ij} is the observed frequency, $n_{i.}$and $n_{.j}$ are the marginal totals, and $n=\Sigma_i\Sigma_j n_{ij}$. The SPSS program we used shows that β = 0.0770 with a standard error of 0.0218. So the overall chi-squared statistic, which obtained before as 22.84, can be partitioned into component parts due to linear regression and other forms of relationship apart from linear. This is shown in table(2).

Table 2:Partitioning the Chi-squared Test

Source of Variation	df	X^2	p-value
Linear regression	1	13.10	0.000
Departure from regression	5	9.74	
Overall chi-squared	6	22.84	0.001

It is clear now from the p-values in the above table that the chi-squared statistic for the linear relationship between the two variables is statistically significant whereas the chi-squared statistic for the relationship which is not linear is statistically insignificant. So partitioning the overall value of the chi-squared has increased the sensitivity of the test. That is, it is not only that there is a relationship between the academic performance and the education of father, but there is an evidence now that this relationship is linear.

Ordinal log-linear model:-

This method is based on the usual log-linear model techniques which are, to some extent, similar to multiple regression models. In log-linear models, all variables that are used for classification are independent variables, and the dependent variable is the number of cases in a cell of the cross-tabulation. The techniques give a range of models: from the independent model which we saw before at the usual chi-squared test of independence to the saturated model where all the interactions are included. For the independent model, the expected frequencies and hence the chi-squared value are to be obtained in the usual way. For the other models, the expected frequencies may be obtained iteratively. Determination of which model is suitable could be done by fitting a saturated model and examining the standardised values for the parameter estimates. Effects with small standardised estimates can usually be deleted from a model. Another way to be used is testing the contribution to a model made by terms of a particular order. The difference in the chi-squared value between the two models is attributable to the term deleted.

For our ordinal variables situation, we use the above described techniques in the same way except the interactions term(s) which will be replaced by new coefficient (or some of the new coefficients in case of multidimensional tables) that taking into account the order of the variables.

The linear-by-linear association model for the academic performance (Y) and the education of father (X) can be expressed as

$$\ln F_{ij} = \delta+\alpha_i^x + \alpha_j^y + \beta\ (u_i-u)(v_j-v)$$

where$\delta$and the two alpha parameters are the usual log-linear terms for the overall mean and the main effects of the academic performance and the education of father. The coefficientβ is actually a regression coefficient that, for a particular cell, is multiplied by the scores assigned to that cell for the academic performance and the education of father.

Using SPSS statistical program, table (3) shows the result of fitting the linear-by-linear association model for the academic performance (PERFORM) and the education of father (FATHED). The model fits the data very well as the goodness-of-fit statistics are relatively low and hence their associated probabilities are very high. Inclusion of one additional parameter in the model has changed the observed significance level from 0.002(or 0.001) for the independence model to 0.17 (or 0.166) for the linear-by-linear interaction model. The model is also better than the one fitted by the scoring method (above) which has 0.000 significance level.

Table3: Fit of the Linear-by-linear Model

```
   Goodness-of-Fit test statistics
   Likelihood Ratio Chi Square      =    431.06328       DF = 17        P =  .000
   Pearson Chi Square               =    347.87697       DF = 17        P =  .000
   Estimates for Parameters
   PERFORM
Parameter          Coeff.           Std.Err.    Z-Value    Lower 95CI    Upper 95CI
         1       1.013548626       .19766     5.12782       .62614        1.40096
         2       .1039497150       .08504     1.22235      -.06273          .27063
   FATHED
Parameter           Coeff.            Std.Err.     Z-Value     Lower 95CI     Upper
95CI
        3        .8166513546     .21834      3.74020       .38870         1.24461
        4        .4561275849     .12549      3.63463       .21016           .70210
        5         -.026510565    .11221       -.23625     -.24645           .19343
   B
Parameter    Coeff.            Std.Err.      Z-Value    Lower 95CI    Upper 95CI
        6        .2682041440     .07679      3.49285       .11770           .41871
```

The above table also shows the estimates of the parameters in the linear-by-linear model and what concerned us among these estimates is the estimate of the coefficientβ. It is positive (0.2682) and large when compared to its standard error (0.07679), indicating that there is a positive association between the academ ic performance and the education of father. That is, as the educational level of a father becomes higher the academic performance of the student becomes

higher and vice versa. The estimated uniform odds ratio for adjacent rows and adjacent columns is exp (0.2682)=1.31, indicates that the chance of observing more academic performance of a student is higher as the educational level of his father gets higher.

The significance of coefficient β can be assessed by testing H_o: $\beta = 0$ which gives a conditional test of independence in table(4), under the assumption that linear-by-linear model holds.

Table4:Conditional Test of Independence

Model	G^2	df	p-value
Independence	20.5200	6	0.002
Linear-by-linear	7.7613	5	0.170
Independence given linear-by-linear	12.7590	1	0.000

The value of the this conditional test is 12.759 with 1 degree of freedom. Hence, there is stronger evidence of an association between the academic performance and the education of father than was by comparing the independence model with a saturated model. In other words, if a positive association really exists between the academic performance and the education of father we are in a better position to detect it using this ordinal approach.

Another type of models which belong to this family of log-linear techniques are the models of mixed ordinal and nominal linear categories. The trend coefficient here is multiplied by the scores of the ordinal variable. The magnitude and sign of the coefficient indicates whether cases are more or less likely to fall in an ordinal variable with a high or low scores, as compared to the independence model.

Consider the situation where we included a third variable to represent the specialization of the student (SPECIAL), labeled as social and scientific. The chi-squared test of independence can be extended for the three variables. The result in table(5) shows that the chi-squared statistics are very high and hence the hypothesis that the three variables are independent is to by rejected. That means the independence model for the academic performance and the education of father and specialization of student does not fit well and so we have to search for a better model to represent the relationship between the three variables.

Table 5: Fit of the general Independence Model

```
 Goodness-of-Fit test statistics
Likelihood Ratio Chi Square      =    431.06328        DF = 17        P =  .000
Pearson Chi Square               =    347.87697        DF = 17        P =  .000
Estimates for Parameters
 FATHED
 Parameter   Coeff.          Std.Err.     Z-Value    Lower 95CI    Upper 95CI
      1      .1289281732    .09870        1.30625      -.06453         .32238
      2       .1995457404    .09657       2.06629       .01026         .38883
      3       .1881170446    .09691       1.94116      -.00183         .37806
 PERFORM
 Parameter   Coeff.          Std.Err.     Z-Value    Lower 95CI    Upper 95CI
      4        .3888125364    .07855      4.94991       .23486         .54277
      5        .0793114165    .08390        .94526      -.08514         .24376
 SPECIAL
 Parameter   Coeff.          Std.Err.     Z-Value    Lower 95CI    Upper 95CI
      6       .1341319933    .05826      2.30249       .01995         .24831
```

Treating specialization of student as a nominal variable, and apart from the saturated model, a general model can be written as

$$\text{Ln } F_{ijk} = \delta+\alpha_i^x + \alpha_j^y + \alpha_k^z + \beta^{xy}(u_i-u)\,(v_j-v) + \tau_k^{xz}(u_i-u)+ \tau_k^{yz}(v_j-v)$$

where z denotes the effect of specialization of student. Table(6) gives a number of possible models formed from the general model by deleting some interaction terms.

Table6:Set of Models Fitted for the academic performance and the education of father and the specialization of student

Model with deleted term(s)	G^2	df	p-value
All interactions	431.06	17	0.000
$\beta^{xy}(u_i-u)\,(v_j-v)$	14.119	16	0.590
$\tau_k^{xz}(u_i-u)$	7.7613	16	0.956
$\tau_k^{yz}(v_j-v)$	46.413	15	0.000
$\beta^{xy}(u_i-u)\,(v_j-v),\ \tau_k^{xz}(u_i-u)$	20.524	17	0.248
$\beta^{xy}(u_i-u)\,(v_j-v),\ \tau_k^{yz}(v_j-v)$	418.25	16	0.000
$\tau_k^{xz}(u_i-u),\ \tau_k^{yz}(v_j-v)$	88.370	16	0.000
None	6.0997	15	0.978

It is clear now from the p-values in the above table that some models fit the data quite well and these are:

$$\text{Ln } F_{ijk} = \delta+\alpha_i^x + \alpha_j^y + \alpha_k^z + \beta^{xy}(u_i-u)\,(v_j-v) + \tau_k^{xz}(u_i-u)+ \tau_k^{yz}(v_j-v)$$

$$\text{Ln } F_{ijk} = \delta+\alpha_i^x + \alpha_j^y + \alpha_k^z + \beta^{xy}(u_i-u)\,(v_j-v) + \tau_k^{yz}(v_j-v)$$

$$\text{Ln } F_{ijk} = \delta+\alpha_i^x + \alpha_j^y + \alpha_k^z + \tau_k^{xz}(u_i-u)+ \tau_k^{yz}(v_j-v)$$

$$\text{Ln } F_{ijk} = \delta+\alpha_i^x + \alpha_j^y + \alpha_k^z + \tau_k^{yz}(v_j-v)$$

and the model

$$\text{Ln } F_{ijk} = \delta+\alpha_i^x + \alpha_j^y + \alpha_k^z + \beta^{xy}(u_i-u)\,(v_j-v) + \tau_k^{yz}(v_j-v)$$

is best one among them since it has a high significance (i.e., better fit) and it is relatively simple (i.e., parsimonious in parameters).

Ordinal logit model:-

When one variable is thought to depend on the others, a special class of models, called logit models, can be used to examine the relationship between variable, such as the academic performance and one independent variable the education of father or more. Since our response variable is ordinal, so it makes sense to use form logits in a way that takes the category order into account. We are using the cumulative logit in which the original response classification is collapsed into two categories and hence similar to the basic logit model for dichotomous response variable. So, for the fixed point j of the ordinal response variable the jth cumulative logit in level i of the explanatory variable is then

$$L_{j(i)} = \alpha_j + \beta_j(u_i-u)$$

where α_j represents the logit of the response variable for point j. If the model is true and $\beta_j = 0$, then the two variables are independent.

For our data, the academic performance is regarded as the response variable and the education of father is the explanatory variable with scores assigned to it. Since there are three categories of the response variable, so we have two cumulative logits: the case where 'high' and 'moderate' performances are pooled together in a single category versus 'poor' performance and the case of 'high' versus 'moderate' and 'poor' together.

The SPSS output in table (7) shows for the first logit the G^2 goodness-of-fit is 2.130 while for the second logit it is 4.077. The corresponding p-values are 0.345 and 0.130, respectively. Accordingly thelogit model fits quite well for both cumulative logits. The association parameter β is positive for the two logits, indicating that the established logit for student academic performance increases as the education of father gets higher.

Table7: Fits of Two Logit Models after category j when collapsing the categories of the academic performance with the education of father.

	j=1	j =2
α_j	0.638	2.512
β_j	0.171	0.229
s.e.(β_j)	0.058	0.073
exp(β_j)	1.186	1.257
G^2: Logit Model	2.130	4.077
df	2	2
p-value	0.345	0.130
G^2: Independence Model	11.01	13.96
df	3	3
p-value	0.012	0.003
G^2: β_j=0	8.879	9.886
df	1	1
p-value	0.000	0.000

For the first logit, the common odds ratio for adjacent rows, that is, exp(β_1), is 1.186. This means that the odds for moderate and high academic performances instead of poor performance is estimated to be 1.186 times higher for educational level i+1 of father than for level i, i=1,2,3. Similarly, for the second logit, the common odds ratio for adjacent rows is 1.257, indicating that the odds ratio for high academic performance instead of moderate and poor performances is estimated to be 1.257 higher for educational level i+1 of father than for level i, i=1,2,3. For both models, therefore, it clear that the odds for the academic performance to be above a certain point rather than below it is more likely to be higher when the educational level of father gets higher.

Comparing the two values of G^2 with the corresponding two values in the independence models, we can arrive at a test for β_j=0 which is shown in table(7).So, the test for

$$H_o : \beta_j=0$$

results in a value of 8.879 for the first logit and a value of 9.886 for the second logit. Since the significance value is very small in both cases, we reject the null hypothesis that $\beta_j \neq 0$ (j=1,2). Hence, there is a strong evidence of an association between the academic performance and the education of father.

CONCLUSION

In dealing with categorical data with ordinal categories there are a number of techniques that could be used. The usual chi-squared test of independence is the easiest and quickest method to apply, but, as we have seen, it doesn't take into consideration the monotonic nature of the ordinal variables. In addition, it doesn't provide estimates of the effects of the variables on each other.

The scoring method is an alterative technique to use. The method is very easy to use by a simple calculator in the absence of any statistical package and, as we have seen, has gone further than the previous chi-squared test of independence in assessing the pattern of the relation. However, the method based on some sort of approximations and, in addition, different scorings might lead to different results and hence different unsystematic evaluations. The last defect, however, is shared with ordinal log-linear models and ordinal logit models as well. But the superiority of ordinal log-linear models and the ordinal logit models is very eminent, since they give empirical estimates of the effects between the variables, estimates corresponding to the ordering nature of the variables, and, more important, take into account whether one variable is to be treated as a response (the ordinal logit models). These later techniques fit the data we used quite well and reached a conclusion, though similar, but very précised and meaningful from what is reached by the usual chi-squared test of independence and the scoring method.

REFERENCES

1. Adam, Amin. I. (1993). Analysing Categorical data: Various Solutions and Different Conclusions. Young Statistical Meeting, Liverpool.
2. Adam, Amin. I. (1996). Analysis of Categorical Data from a Case Study of Child Safety. Unpublished PhD thesis: University of Keele, Dept. of Mathematics, England.
3. Adam, Amin. I. (2010). Concepts on the Chi-square Test of Independence for Analyzing Categorical data. Journal of the Faulty of Economics & Political Science, Omdurman I. University, Sudan, Vol.4:131-143.
4. Agresti, A. (2002). Categorical Data Analysis, 2nd ed. Wiley, New York.
5. Agresti, A. (2007). An Introduction to Categorical Data Analysis. Wiley, New York.
6. Agresti, A. (2010). Analysis of Ordinal Categorical Data. Wiley, New York.
7. Baglivo, J., Oliver, D. & Pagano, M. (1992). Methods for Exact Goodness-of-Fit Tests. Journal of the American Statistical Association 87:464-469.
8. Becker, M. P. & Clogg, C. C. (1989). Analysis of Sets of Two-Way Contingency Tables Using Association Models. Journal of the American Statistical Association 84:142-151.
9. Bilder, C. & Loughin, T. M. (2007). Modeling Association Between Two or More Categorical Variables that Allow for Multiple Categorical Choices. Communications in Statistics 36:433-451.
10. Everitt, B. S. (1977). The Analysis of Contingency Tables. Chapman & Hall, London.
11. Eye, A. V. & Bogat, G. A. (2009). Analysis of Intensive Categorical

Longitudinal Data. Springer, New York.

12. Fan, Y.. (2008). Strategic Groups and cluster Analysis. Henry Stewart, London.
13. Fienberg, S. E. (2007). The Analysis of Cross-classified Categorical Data. Springer, New York.
14. Freeman, D.H. (1987). Applied Categorical Data Analysis. Marcel Dekker, New York.
15. Greenland, S. (1991). On the Logical Justification of Conditional Tests for Two-by-Two Contingency Tables. American Statistician 45:248-251.
16. Imrey, P. B. & Koch, G. G. (2005). Categorical Data Analysis. Wiley, New York.
17. Liu, I. & Agresti, A. (2005). The Analysis of Ordinal Categorical Data: An Overview and a Survey of Recent Development. Sociedad de Estadistica e Investigacion OperativaTe Vol.14 No. 1:1-73.
18. McCullagh, P. (1980). Regression Models for Ordinal Data. J. Roy. Statist. Soc. B 42:109-142.
19. Ott, R. L. & Longnecker M. (2008). An Introduction to Statistical Methods and Data Analysis, 6th ed. Brooks/Cole, Bolmont, U.S.A.
20. Powers, D. A. (2008). Statistical Methods for Categorical Data Analysis, 2nd ed. Emerald, Bingley, U.K.
21. Simono, J.(2003). Analyzing Categorical Data. Springer, New York.
22. Upton, G. J. G. (1978). The Analysis of Cross-tabulated Data. Wiley, New York.