DE GRUYTER
OPEN

Folia Oeconomica Stetinensia
DOI: 10.1515/foli-2016-0017

WYDZIAŁ NAUK EKONOMICZNYCH
I ZARZĄDZANIA
UNIWERSYTET SZCZECIŃSKI

ORDINAL LOG-LINEAR MODELS FOR CONTINGENCY TABLES

Justyna Brzezińska, Ph.D.

University of Economics in Katowice
Faculty of Finance and Insurance
Department of Economic and Financial Analysis
1 Maja 50, 40-287 Katowice, Poland, Poland
e-mail: justyna.brzezinska@ue.katowice.pl

Received 23 October 2015, Accepted 27 June 2016

Abstract

A log-linear analysis is a method providing a comprehensive scheme to describe the association for categorical variables in a contingency table. The log-linear model specifies how the expected counts depend on the levels of the categorical variables for these cells and provide detailed information on the associations. The aim of this paper is to present theoretical, as well as empirical, aspects of ordinal log-linear models used for contingency tables with ordinal variables. We introduce log-linear models for ordinal variables: linear-by-linear association, row effect model, column effect model and RC Goodman`s model. Algorithm, advantages and disadvantages will be discussed in the paper. An empirical analysis will be conducted with the use of R.

Keywords: association models, ordinal variables, contingency table, log-linear models

JEL classification: C35, C52

Introduction

There has been a tremendous increase in the publication of research on analyzing categorical data measured on an ordinal scale. Since about 1980, there has been an increasing emphasis on having data analyses distinguish between an ordered and unordered scale for the categories. Many advantages can be gained from treating an ordered categorical variable as ordinal rather than nominal. An ordinal analysis can provide a greater variety of models, and those models are more parsimonious and have simpler interpretations than the standard models for nominal variables. Also, they have greater power for detecting relevant trend or location alternatives to the null hypothesis of no effect of an explanatory variable on the response one. Ordinal variables can include an interesting model that for standard nominal models are trivial or have too many parameters to be tested for goodness of fit (Agresti, 2010).

In the past years, scientists have become more increasingly familiar with log-linear models for nominal categorical variables. For joint distribution of categorical response variables in a multi-way table, log-linear models describe the dependence structure. They can analyze whether the association between a pair of variables is homogenous across the categories of other variables, and if so, whether those variables are conditionally independent. Those models examine the relationship among categorical variables by analyzing observed data. For nominal log-linear models, no assumptions are made about the order of the measurement of the variables. Because nominal models are insensitive to the ranking of these ordinal variables, they ignore important information when at least one variable is ordinal.

Ordinal log-linear models can be treated as an extension of nominal log-linear models. When nominal variables X and Y are examined in a log-linear analysis, the saturated model includes the interaction term between X and Y. However, this model has no degrees of freedom, and is always of little importance since we are interested in testing a more parsimonious model. For a two-way table, the next model is an independence model, however, this model is unrealistic and the fit of it is usually poor. For nominal variables, there are no other models between independence and saturated models. The ordinal approach provides a model between these two. With ordinal models we can test a greater variety of substantively important models.

So far, in the context of classical log-linear models, there are just two options for modeling two-way contingency tables: the parsimonious but restrictive model of independence, and the saturated model. Association models fill the gap between these two extreme cases by imposing a special structure on the association and reducing the number of interaction parameters, providing thus intermediate models for independence. For better understanding and also for the purpose of interpretation, it is convenient to think in terms of the local association in the table and first to define the models on the local odds ratios.

The aim of this paper is to present theoretical, as well as empirical, aspects of ordinal log-linear models used for contingency tables with ordinal variables. We introduce log-linear models for ordinal variables: linear-by-linear association, row effect model, column effect model and RC Goodman`s model. This study focuses on the methods of analysis of categorical data having ordered categories for multi-way tables. This paper discusses some of the specialized models which use the information on the ordering, unlike standard methods for the data measured on a nominal scale. Several log-linear models for ordinal variables will be presented in the paper based on the report *Social diagnosis 2013. Conditions and quality of life by Poles* with the use of R software.

1. Ordinal log-linear models as association models

Log-linear models are a standard tool to analyze the structures of dependency in multi-way contingency tables. Standard log-linear models treat all classification variables as nominal, unordered factors. The criteria to be analyzed are the expected cell frequencies as a function of all the variables measured on a nominal scale. A saturated model for a two-way table $I \times J$ ($i = 1, 2, ..., I, j = 1, 2, ..., J$) includes all the possible effects [Bishop, Fienberg, Holland 1975, Knoke, Burke 1980, Ishii-Kuntz 1994, Christensen 1997, Agresti 2002]:

$$\log\left({}_{ij} \right) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \qquad (1)$$

where: λ represents an overall effect or a constant, λ_i^X and λ_j^Y represent the effect of the row and column variable, λ_{ij}^{XY} represents the interaction between two variables.

Log-linear models that use the ordered nature of the factors offer several advantages (Brzezińska, 2015). Because they are more focused, the tests that use the ordinal structure of the table variables are more powerful when the association varies systematically with the ordered values of a factor. Because they consume fewer degrees of freedom, we can fit an unsaturated model where the corresponding model for nominal factors would be saturated. In a two-way table, for example, a variety of models for ordinal factors may be proposed that are intermediate between the independence model and the saturated model. Another advantage is that the parameter estimates from these models are fewer in numbers, are easier to interpret, and quantify the nature of effects better than the corresponding quantities in the model for nominal factors. Estimating fewer parameters typically gives smaller standard errors.

Ordinal log-linear models are used when one or more variables are ordinally measured, that is, if one of the variables is ordinal, a nonsaturated model can be used to measure the association. In an ordinal log-linear analysis, we can distinguish a row effect model, column effect model, uniform association model and RC Goodman's model. The association parameters in log-linear models describe the ordinal characteristics of the data. The descriptive statement made with these methods is always more informative than the one with nominal. There is also greater potential for detecting certain forms, and the greater variety of ways of describing the association.

Another type of an association model for ordinal variables is a row effects model and a column effect model. When for a two-way table one variable is treated as ordinal and the other as nominal, we have a row effects model or a column effect model. In a row effects model, a row variable is nominal and a column variable is ordinal. We can construct this model by replacing λ_{ij}^{XY} in (1) with the association term $\tau_h(v_j - \overline{v})$ (Ishii-Kuntz, 1994):

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \tau_h(v_j - \overline{v}) \tag{2}$$

where the τ_h s are row effects parameters and the v_j ($v_1 < v_2 < ... < v_J$) are scores assigned to the columns in the contingency table, and $\tau_h(v_j - \overline{v})$ is an association parameter in which $v_j - \overline{v}$ is the ordinal of categories, replacing interaction parameter λ_{ij}^{XY} in (1). In addition, the zero-sum constraints $\sum_{i=1}^{I}\lambda_i^X = \sum_{j=1}^{J}\lambda_j^Y = \sum_{i=1}^{I}\tau_i = 0$ are imposed in order to identify the model`s parameters. Model (2) has $(I-1)(J-1)$ degrees of freedom. The row effects model (2) is more parsimonious than the saturated model (1). For the arbitrary pair of rows h and i, and the adjacent columns j and j + 1, the log odds ratio is:

$$\begin{aligned}\log\left(\frac{m_{hj}m_{i,j+1}}{m_{h,j+1}m_{ij}}\right) &= [\mu + \lambda_h^X + \lambda_j^Y + \tau_h\left(v_j - \overline{v}\right) + \mu + \lambda_i^X + \lambda_{j+1}^Y + \tau_i\left(v_{j+1} - \overline{v}\right)] - \\ &- [\mu + \lambda_h^X + \lambda_{j+1}^Y + \tau_h\left(v_{j+1} - \overline{v}\right) + \mu + \lambda_i^X + \lambda_j^Y + \tau_i\left(v_j - \overline{v}\right)] = \\ &= \tau_h\left(v_j - \overline{v}\right) + \tau_i\left(v_{j+1} - \overline{v}\right) - \tau_h\left(v_{j+1} - \overline{v}\right) - \tau_i\left(v_j - \overline{v}\right) = \\ &= \tau_h\left(v_j - v_{j+1}\right) + \tau_i\left(v_{j+1} - v_j\right) = \tau_i - \tau\end{aligned} \tag{3}$$

because $\left(v_j - v_{j+1}\right) = -1$ and $\left(v_{j+1} - v_j\right) = 1$. Thus, the odds ratio is: $\theta = e^{\tau_i - \tau_h}$. A column effect model is a simple variation of the row effects model, with the difference that the row variable is ordinal, and the column variable is nominal.

A uniform association model treats the levels of both a row and column variable as ordinal. We use a set of integer scores for both a row and column variable to reflect the ordering of these variables: $\{u_i\}$ for a row variable ($u_1 < u_2 < ... < u_I$) and $\{v_j\}$ for a column variable ($v_1 < v_2 < ... < v_J$). The choice of scores will reflect the assumed distances between the midpoints of categories for an underlying interval scale. Equally spaced scores result in the simplest interpretation for the model. In practice, the integer scores $\{u_i = i\}$ and $\{v_j = j\}$ are most commonly used, and this approach will be explained in the empirical part of the paper.

The goal is to pose a model more complex than the independence model, but not saturated. This is done by including an association term reflecting the relationship between two ordinal variables (Ishii-Kuntz, 1994):

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \beta(u_i - \overline{u})(v_j - \overline{v}) \tag{4}$$

To identify this model, the zero-sum constraint is imposed and $\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = 0$ and the transformation of scores by $(u_i - \overline{u})(v_j - \overline{v})$ corresponds to the order $(-1, 0, 1)$. Since $\{u_i\}$ and $\{v_j\}$ are fixed, the uniform association model has only one more parameter (β) than the independence model. Thus, there are $IJ - [1 + (I - 1) + (J - 1)] - 1$ degrees of freedom for the uniform association model. The log odds ratio is given as:

$$\begin{aligned}\log\left(\frac{m_{ac}m_{bd}}{m_{ad}m_{bc}}\right) &= \left[\mu + \lambda_a + \lambda_c + \beta\left(u_a - \overline{u}\right)\left(v_c - \overline{v}\right) + \mu + \lambda_b + \lambda_d + \beta\left(u_b - \overline{u}\right)\left(v_d - \overline{v}\right)\right] - \\ &- \left[\mu + \lambda_a + \lambda_d + \beta\left(u_a - \overline{u}\right)\left(v_d - \overline{v}\right) + \mu + \lambda_b + \lambda_c + \beta\left(u_b - \overline{u}\right)\left(v_c - \overline{v}\right)\right] = \\ &= \left[\beta\left(u_a - \overline{u}\right)\left(v_c - \overline{v}\right) + \beta\left(u_b - \overline{u}\right)\left(v_d - \overline{v}\right)\right] - \\ &- \left[\beta\left(u_a - \overline{u}\right)\left(v_d - \overline{v}\right) + \beta\left(u_b - \overline{u}\right)\left(v_c - \overline{v}\right)\right] = \\ &= \beta\left(u_a v_c + u_b v_d\right) - \beta\left(u_a v_d + u_b v_c\right) = \beta\left(u_b - u_a\right)\left(v_d - v_c\right)\end{aligned} \tag{5}$$

The independence model is a special case of (3) when $\beta = 0$.

As the backward elimination procedure is to be used in selecting the most representative model by utilizing the natural ordering of the variables, we start the model fitting procedure by taking the selected standard model into account. Ordinal log-linear models can be used for two- and multi-way contingency tables. The overall goodness-of-fit of a model is assessed by comparing the expected frequencies to the observed cell frequencies for each model. It is necessary to assess the goodness-of-fit of the model. The following hypothesis are used: H_0: the model represents association well enough *vs.* H_1: the model does not represent association well enough. The goodness-of-fit of a log-linear model is usually tested using either the Pearson chi-square statistic test or the likelihood ratio (Knoke, Burke, 1980; Christensen, 1997):

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \log\left(\frac{n_{ij}}{m_{ij}}\right) \tag{6}$$

Therefore, the larger G^2 values indicate that the model does not fit the data well, and thus the model should be rejected. This strategy is the opposite of the usual chi-square test of independence, where we seek to reject the null hypothesis of no association. However, in trying to find the best fitting log-linear model to describe a cross-table, we hope to accept the hypothesized model, hence we want to find a low G^2 value relative to *df* (Knoke, Burke, 1980). The likelihood ratio can also be used to compare an overall model within a smaller, nested model (i.e. a saturated model with one interaction or main effect dropped to assess the importance of that term). The equation is $\Delta G^2 = G_2^2 - G_1^2$ with: $\Delta df = df_2 - df_1$, where 2 is a nested model, 1 is the higher parameterized model, df_1 and df_2 are degrees of freedom for model 1 and 2.
Also, information criteria can be used to test the goodness-of-fit: AIC (Akaike, 1973) and
BIC (Raftery, 1986). Akaike information criterion (Akaike, 1973) refers to the information contained in a statistical model according to equation:

$$AIC = G^2 - 2df \tag{7}$$

The model that minimizes AIC will be chosen.

2. Application in R

First, ordinal log-linear models will be presented with the use of the row effects model based on the data on influenza in 2013 from the National Institute on Public Health. The sample size was 213,906. A two-way table was build for the nominal and ordinal variables: *voivodeship* (16 provinces of Poland) and *age* (0–4, 5–14, 15–64 and 65+). As *age* is measured ordinally, a new variable is included in the row effects model: *c.age* with scores (1, 2, 3, 4), which is the main effect confounded with *age*. Thus, its coefficient is not estimable.

To obtain the row effects model in R, we use function: `glm(formula=count~Age +Voivodeship*c.Age,family=poisson)`. The goodness-of-fit of the independence model and of the row effects model is summarized in Table 1.

Table 1. Goodness-of-fit statistics for saturated, independence and row effects model

Model	G^2	df	ΔG^2	Δdf	AIC
[VA]	0	0	1,201.6	0	731.42
Row effects model	1,201.6	30	1,005.2	5	1,873.1
[V][A]	2,206.8	45	---	---	2,848.2

Source: own calculations in R.

The saturated model [VA] fits the data perfectly with 0 degrees of freedom. The likelihood ratio statistic for the row effects model is 1,201.6 with 30 degrees of freedom, and the AIC criterion is 1,873.1. The independence model [V][A] has likelihood ratio statistic 2,206.8 with 45 degrees of freedom, and the AIC criterion is 2,848.2. The ΔG^2 tests whether the corresponding model results in a significant reduction in the residual G^2 compared to the independence model. We can see, that the best fit occurs for the row effects model. For this model also the minimum value of AIC criterion is obtained (except the saturated model). We obtain the row effects model with the use of `coef` function, and the expected cell frequencies which are fitted values with the use of fitted function.

A uniform association model was built on the data from Social Diagnosis 2013 – "Objective and Subjective Quality of Life in Poland." The sample size was 26,307 respondents by *age* (0–24, 25–34, 35–44, 45–59, 60–64 and 65+) and *time spent on watching TV* (0–1, 1–3 and 3+ hours). The independence, saturated and uniform association model was built.

The model including the row and column scores is defined as: `glm(count~Age*Time+c.Age*c.Time,data=data,family=poisson).` The goodness-of-fit statistics for the independence model and the uniform association model are summarized in Table 2.

Table 2. Goodness-of-fit statistics for saturated, independence and uniform association model

Model	G^2	df	ΔG^2	Δdf	AIC
[VA]	0	0	0	0	196.93
Uniform association model	0	0	1,606.6	10	196.93
[V][A]	1,606.6	10	---	---	1,783.5

Source: own calculations in R.

The analysis of Table 2 shows that the uniform association model fits the data very well with 10 degrees of freedom. In comparison to nominal models (the saturated and complete independence model), the uniform association model is chosen as the best fitting.

The analysis of the ordinal log-linear models proves that there is a large variety of models that can be used to analyze ordinal categorical data with the use of model-based methods. These methods are always more informative than those based on non-model based methods that ignore the ordinal nature of the variables.

Another dataset presented in this paper is based on the report *Social diagnosis 2013. Conditions and quality of life by Poles* on time spent on watching television. The sample size was 26,307 adult respondents. The survey was based on two questions: *time spent on watching TV* (0–1, 1–3, 3 and more hours) and *age* (18–24, 25–34, 35–44, 45–59, 60–64, and 65 and more). First of all, a correspondence analysis was applied to measure the association between the variables (treated as nominal variables). The value of total inertia is $\lambda = 0.0614$ which means that there is very a week association between the variables. The first dimension is explained by 91.3% of the total inertia, and the two dimensions explain 100% of the total inertia. The independence between the variables can be also seen on the perception map (Figure 1).

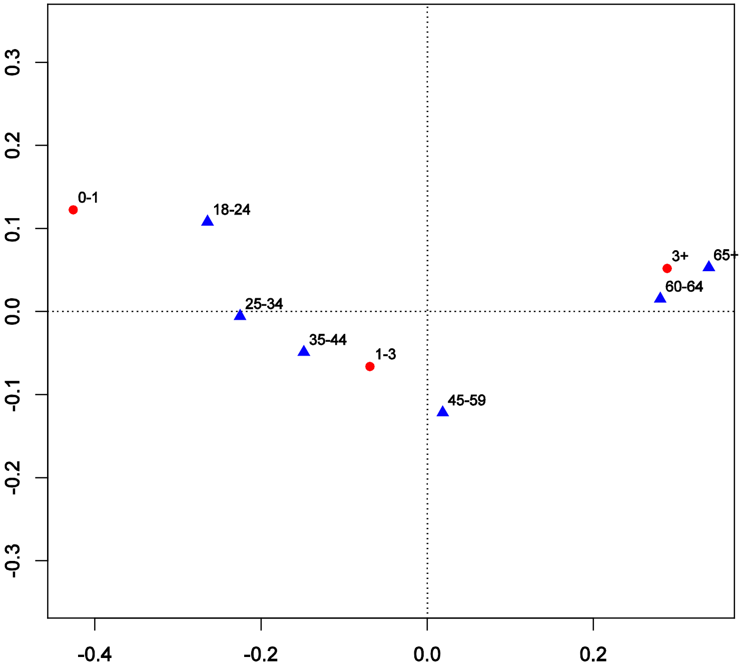

Figure 1. Perception map in a correspondence analysis

Source: own calculations in R.

To check whether the independence appears also in a log-linear model, we built a uniform association model, row effects model, and column effects log-linear model.

Table 3. Akaike Information Criteria (AIC) for ordinal log-linear models

Model	G^2	df	AIC
Saturated	0	0	196.93
Uniform association	190.10	9	369.03
Row effects	182.64	8	363.58
Column effects	85.52	5	272.46
Independence	1,606.60	10	1,783.50
RC Goodmans`	79.15	0	79.15

Source: own calculations in R.

From all the models analyzed we can see that the model that fits best is the column effects model defined with the equation $\log(m_{ij}) = \lambda + \lambda_i^T + \lambda_j^A + \tau_i(v_j - \overline{v})$ with the likelihood ratio value 85.52 and the degrees of freedom equal to 5. Also, the information criteria for this model indicate the best fit and the smallest value. From the detailed analysis of the parameters values we can see that the parameters for the interaction of age, and the ranges connected with the column variable are increasing (0.0595, 0.1763, 0.4348, 0.8631, 0.9656). The parameters with a positive sign for the column analyzed mean that more observations appear in the columns with the higher values of the ordinal variable, and less in the columns with the smaller values of the variable compared to independence.

Here, with the use of ordinal log-linear models, we can obtain a more detailed analysis of association compared to the classical methods of association. It means that log-linear models provide unique information on the path of association that can be rarely found with the use of other methods. However, log-linear models are not free of disadvantages; the choice of scores may highly depend on the data and the context of the problem that is analyzed. Therefore, there are other ways of using and modeling ordinality, e.g. cumlative logit models.

Conclusions

Log-linear models are a powerful statistical tool for analyzing cross tables with nominal and ordinal variables. There are many advantages in using ordinal instead of nominal log-linear models. The main advantage of ordinal log-linear models is that they have structured association and interaction terms that contain fewer parameters and retain more residual degrees of freedom than the nominal models. In comparison to nominal models, we can only choose between a saturated model and model of independence. Ordinal models provide a greater variety of models, including the ones that exist between a saturated and independence model. We can test several models that exist between an independence and saturated model, which is impossible with the use of the well known methods of association.

The main objective of the paper was to discuss the advantages of using orderings of ordinal categorical variables. The models for ordinal variables provide easier quantification of association in terms of odds ratios, and have more power to detect interactions when compared to the models for nominal variables. Log-linear models are a powerful statistical tool for analyzing cross tables with nominal and ordinal variables. There are many advantages in using ordinal instead of nominal log-linear models. First, where nominal models are saturated, there are unsaturated ordinal log-linear models. Ordinal models have structured association and interaction terms that contain fewer parameters and retain more residual degrees of freedom than nominal models.

Refferences

Agresti, A. (2002). *Categorical data analysis.* New Jersey: Wiley & Sons, Hoboken.

Agresti, A. (2010). *Analysis of ordinal categorical data.* New Jersey: Wiley & Sons, Inc. Publication.

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.* In: B.N. Petrow, F. Czaki (eds.), *Proceedings of the 2nd International Symposium on Information*. Budapest: Akademiai Kiado.

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

Brzezińska, J. (2015). *Analiza logarytmiczno-liniowa. Teoria i zastosowania z wykorzystaniem programu R*. Warszawa: C.H. Beck.

Christensen, R. (1997). *Log-linear models and logistic regression*. New York: Springer-Verlag.

Ishii-Kuntz, M. (1994). Ordinal log-linear models. *Quantitative Applications in the Social Science*, *97*.

Knoke, D., Burke, P.J. (1980). Log-linear models. *Quantitative Applications in the Social Science*, *20*.

Raftery, A.E. (1986). Choosing models for cross-classification. *Amer. Sociol. Rev.*, *51*, 145–146.