ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 286, 2013

Justyna Brzezińska[*]

ODDS RATIOS IN THE ANALYSIS OF CONTINGENCY TABLE

Abstract. Association and relationship is one of the most important tasks in statistical analysis. The main objective of the study is to examine odds ratios as a framework for understanding of contingency tables and log-linear models. Odds ratios are used to measure the association for contingency tables. They can be generalized to larger tables by local odds ratios or by the spanning cell approach. The properties of odds ratios and the relationship with log-linear analysis will be presented in the paper. An example is presented with the use of **R**.

Key words: odds-ratio, contingency table, log-linear analysis.

I. INTRODUCTION

Association is one of the most important tasks in statistical data analysis. The main purpose of most research is to assess relationship among set of variables. Choosing an appropriate technique depends on the type of variables. When the data are qualitative (nominal or ordinal), the common mode of analysis is tabular. A traditional way of examining the association in cross-table would be interpretation of differences in cell percentages, an appropriate measure of association (correlation-type coefficients) or log-linear modeling. In this paper odds-ratio (*cross-product ratio* [Mosteller 1968]) will be presented for the understanding of log-linear models and association. Surprisingly, this measure appears only rarely in social sciences research, although it is widely used in chemical, genetic, and medical contexts [Cornfield 1956].

II. ODDS-RATIO AND ITS FUNCTIONS AS A MEASURE OF ASSOCIATION

Odds-ratio is a measure of association among the variable forming the table [Rudas,1998] and is used to quantify the strength of association between nominal variables. The advantage of this measure is that it is appropriate for three types of sampling models for contingency table: Poisson, multinomial and product-multinomial. For 2×2 table the odds-ratio is defined as:

[*] M.Sc., Justyna Brzezińska, University of Economics in Katowice.

$$\theta = \frac{p_{11}p_{22}}{p_{12}p_{21}}, \tag{1}$$

where p_{hj} is the probability to fall to hj-cell.

The odds-ratio has several desirable properties [Fienberg 1980]:

1. θ is invariant under the change of rows and columns, although an interchange of only rows or only columns changes θ into $\frac{1}{\theta}$.

2. θ is invariant under row and column multiplications (if we multiply first row by c and row by d, odds-ratio is again equal to θ):

$$\frac{cp_{11}dp_{22}}{cp_{12}dp_{21}} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \theta. \tag{2}$$

3. θ has clear interpretation (if we think of row totals are fixed, then $\varpi_1 = \frac{p_{11}}{p_{12}}$ is the odds of being in the first column given that one is in the first row, and $\varpi_2 = \frac{p_{21}}{p_{22}}$ is the corresponding odds for the second row; the relative odds for the two rows, or the odds-ratio, is then: $\frac{\varpi_1}{\varpi_2} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \theta$).

4. θ can be used in $H \times J$ and multidimensional tables either through a series of 2×2 partitioning or by looking at several 2×2 subtables.

The quantity θ runs from 0 to ∞ and is symmetric in the sense that two values of odds-ratios θ_1 and θ_2 such that $\log(\theta_1) = -\log(\theta_2)$ represent the same degree of association, although in opposite direction. If $\theta = 1$, the variables corresponding to rows and columns are independent; if $\theta \neq 1$, they are dependent. Yule proposed the use of two functions based on odds-ratio for 2×2 tables. Measure of association is defined as [Yule 1900]:

$$Q = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{p_{11}p_{22} / p_{12}p_{21} - 1}{p_{11}p_{22} / p_{12}p_{21} + 1} = \frac{\theta - 1}{\theta + 1}. \tag{3}$$

Another is a measure of colligation [Yule 1912]:

$$Y = \frac{\sqrt{p_{11}p_{22}} - \sqrt{p_{12}p_{21}}}{\sqrt{p_{11}p_{22}} - \sqrt{p_{12}p_{21}}} = \frac{\sqrt{\theta} - 1}{\sqrt{\theta} + 1}. \tag{4}$$

Both Q and Y range from $[-1,1]$ taking the value 0 when row and column are independent and the value 1 or -1 when there is complete positive or negative association. As both are functions of θ they can take the value of 1 or -1 when only one cell probability is zero.

There is also correlation coefficient defined as [Bishop, Fienberg, Holland 1975]:

$$\rho = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}}, \tag{5}$$

which is invariant under interchange of both rows and columns and it ranges from -1 o 1, where 0 means that variables are independent.

The standard error (SE) for the log odds-ratio in 2×2 tables is approximately:

$$SE(\log OR) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \tag{6}$$

and a 95% confidence interval for the log odds ratio is obtained as 1.96 standard errors on either side of the estimate.

For $H\times J$ table we have $(H-1)\times(J-1)$ 2×2 tables, where odds-ratio defined as:

$$\theta_{hj} = \frac{p_{hj}p_{(h+1)(j+1)}}{p_{h(j+1)}p_{(h+1)j}}, \tag{7}$$

where: $h = 1,\ldots,h+1,\ldots,H-1$, $j = 1,\ldots,j+1,\ldots,J-1$.

Each of these 2×2 tables are defined as the intersection of two adjacent rows and two adjacent columns [Knoke, Burke 1980]. For every 2×2 table, the odds-ratio can be computed, and the collection of these local odds-ratios can be used to describe the association structure of the $H\times J$ table [Rudas 1998]. There are also tables that have zero cells. Zero entries in contingency tables are of two types: *fixed* and *sampling* zeros. Fixed zeros occur when it is impossible to observe values for certain combinations of the variables and sampling zeros are due to sampling variation and the relatively small size of the sample when compared with large number of cells. They disappear when we increase the sample size sufficiently [Fienberg 1980]. When, in the data available, one or both of the frequencies (n_{12}, n_{21}) are 0, the odds-ratio cannot be computed. In such situations 0.5 can be added to such cells.

III. ODDS-RATIO IN LOG-LINEAR ANALYSIS

In log-linear analysis we construct a model such that cell frequencies in a contingency table are accounted by for by the minimum number of terms (parameters). This is done by backward elimination, where we start from the analysis including all possible variables (saturated model) and we remove the highest order interaction with the use of hierarchy principle. An example of log-linear model with the use of odds-ratio is presented on the example of homogeneous association model $[XY][XZ][YZ]$. Log of odds-ratio for two variables X and Y ($h=1,2,...,h^*,...,H$, $j=1,2,...,j^*,...,J$) for k-level of the third ($k=1,2,...,K$) is defined as:

$$\log\theta_{XY(k)} = \log\left(\frac{m_{hj}m_{h^*j^*}}{m_{h^*j}m_{hj^*}}\right) = \lambda_{hj}^{XY} + \lambda_{h^*j^*}^{XY} - \lambda_{h^*j}^{XY} - \lambda_{hj^*}^{XY}\,. \tag{8}$$

Log-odds-ratio for other two variables X and Z ($h=1,2,...,h^*,...,H$, $k=1,2,...,k^*,...,K$) for j-level of the third ($j=1,2,...,J$) is defined as:

$$\ln\theta_{XZ(j)} = \log\left(\frac{m_{hjk}m_{h^*jk^*}}{m_{h^*jk}m_{hjk^*}}\right) = \lambda_{hk}^{XZ} + \lambda_{h^*k^*}^{XZ} - \lambda_{h^*k}^{XZ} - \lambda_{hk^*}^{XZ}\,. \tag{9}$$

Log-odds-ratio for other two variables Y and Z ($j=1,2,...,j^*,...,J$, $k=1,2,...,k^*,...,K$) for h-level of the third ($h=1,2,...,H$) is defined as:

$$\ln\theta_{YZ(h)} = \log\left(\frac{m_{hjk}m_{hj^*k^*}}{m_{hj^*k}m_{hjk^*}}\right) = \lambda_{jk}^{YZ} + \lambda_{j^*k^*}^{YZ} - \lambda_{j^*k}^{YZ} - \lambda_{jk^*}^{YZ}\,. \tag{10}$$

Lambdas are parameters of log-linear model $[XY][XZ][YZ]$. Odds ratio is a very important tool in the interpretation of parameters of log-linear model for tables of any dimension and show that odds ratio depends on the magnitude and direction of the association between analyzed variables.

IV. APPLICATION IN R

Odds-ratio can be computed in **R** software with the use of functions: `oddsratio {vcd}` for $2\times2\times...\times2$ tables or `loddsratio {vcdExtra}` for multi-way tables. For an $H\times J$ table, odds ratios are formed for the set of

$(H-1)\times(J-1)$ 2×2 tables, corresponding to some set of contrasts among the row and column variables. Data on accidents in Polish mines from 1st January - 22nd October 2012 come from State Mining Authority (http://www.wug.gov.pl/). Three variables were considered: “mine” (hard coal mining, copper ore mining, surface mining, borehole mining, others), “crew” (own crew, contractors), “accident” (fatal accident, serious accident, others). The sample size was 2951. To create $2\times 2\times 5$ table fatal and serious accidents were summed to one category; in case of zero entries, 0.5 was added to the table. For the association between “Crew” and “Accident”, log-odds-ratios, standard errors, the Wald test statistic z -value and the associated p -value were obtained.

```
Log Odds Ratio Std. Error z value  Pr(>|z|)
Hard_coal       -1.71269    0.32111 -5.3337 4.812e-08 ***
Copper           1.77784    1.08062  1.6452  0.049965 *
Surface         -3.04452    1.04951 -2.9009  0.001860 **
Borehole        -0.22314    1.55050 -0.1439  0.442783
Others          -0.91629    1.26555 -0.7240  0.234525
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The summary includes log-odds-ratios, standard errors, Wald test statistic z and the associated p-value. If log-odds-ratio is 0, there is said to be no association between variables. The farther from 0 the value is, the stronger association. Log-odds-ratios that are positive show indicate direct covariation between variables, while smaller than 0 – an inverse relationship. It is easy to obtain observed odds-ratios (OR) from log-odds-ratio (LOR) using `exp()` function, as well as other measures based on odds-ratio (Tab. 1).

Table 1. Log-odds-ratios, odds-ratios and Q-Yule, Y-Yule, ρ-correlation coefficients

Mine	LOR	OR	Q-Yule	Y-Yule	Correlation r
Hard coal mine	–1.713	0.180	–0.694	–0.404	–0.118
Copper mine	1.778	5.917	0.711	0.418	0.075
Surface mine	–3.045	0.048	–0.909	–0.642	–0.409
Borehole mine	–0.223	0.800	–0.111	–0.056	–0.034
Other mine	–0.916	0.400	–0.375	–0.225	–0.076

Source: own calculations in **R**.

The greatest value of log-odds-ratio is for copper mine ($\log(\theta) = 1.778$, $\theta = \exp(1.778) = 5.917$) shows that 5.917 times more likely fatal and serious accidents will appear (as opposed to other accidents) for own crew miners as compared with contractors. Other odds-ratios in hard coal, surface and other mines are smaller than 1 and show the inverse relation. The analysis of correlation measurements (Q-Yule, Y-Yule, ρ-correlation coefficient) in different mines types show, that the strongest association between accident and crew type is in surface mine. For this mine all the correlation coefficients are maximum. We can also plot log-odds-ratios to see the relationship between these values. (Fig. 1), where red line for $\log(\theta) = 0$ shows no association between variables.

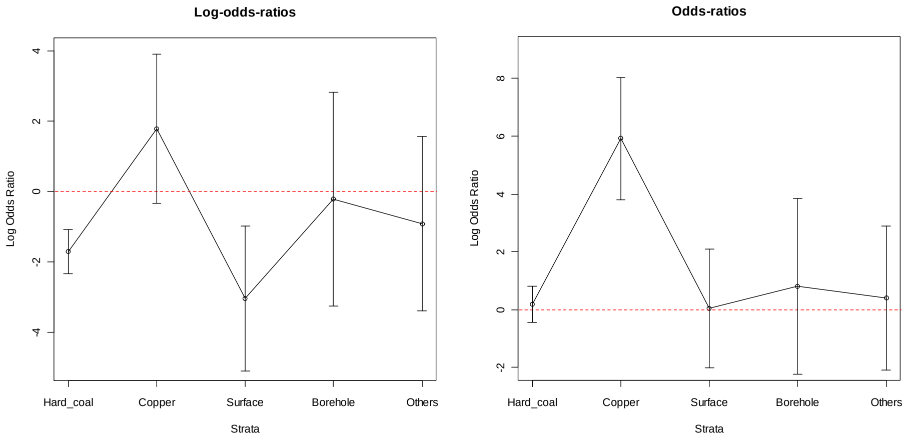

Fig. 1. Plot of log-odds-ratios and odds-ratios for all mine types
Source: own calculations in **R**.

Vertical bars in Figure 1 give 95% confidence interval. Independence with the use of odds-ratio can be also visualized by fourfold display (`fourfold` function in `vcd` package).

Log-linear analysis was conducted for three-way table and goodness of fit statistics (likelihood ratio statistic G^2 and Akaike Information Criterion AIC [Akaike 1973] were computed for choosing the best fitting model.

Table 2. Goodness of fit statistics for three-way table

Model	G^2	*df*	*p-value*	*AIC*
$[M][C][A]$	100.2386	13	1.44329e-15	74.2386
$[C][AM]$	92.16674	9	5.551115e-16	74.1667
$[A][MC]$	42.18322	9	3.042537e-06	24.1832
$[M][CA]$	80.35544	12	3.53062e-12	56.3554
$[MA][CA]$	72.28361	8	1.722733e-12	56.2836
$[MC][CA]$	22.30009	8	4.389264e-03	6.3001
$[MC][MA]$	34.11139	5	2.262467e-06	24.1114
$[MC][MA][CA]$	13.13516	4	0.01063391	5.1352
$[MCA]$	0.0000	1	1	–2.0000

Source: own calculations in **R**.

With the use of information criterion and ANOVA the best fitting model is model of conditional independence $[MC][CA]$. The parameter estimates for this model are:

```
$`(Intercept)`
[1] 1.942496
$Crew
   Own_crew Contractors
  0.4229722  -0.4229722
$Accident
Fatal_serious        Others
    -1.867422      1.867422
$Mine
Hard_coal     Copper    Surface   Borehole     Others
2.9878332  1.2037099 -0.9467088 -2.2342087 -1.0106255
$Crew.Accident
             Accident
Crew          Fatal_serious     Others
  Own_crew       -0.3296981  0.3296981
  Contractors     0.3296981 -0.3296981
$Crew.Mine
            Mine
Crew            Hard_coal     Copper    Surface   Borehole
Others
Own_crew     0.07405832 -0.3713987  0.5462355 -0.6359038
0.3870087
Contractors -0.07405832  0.3713987 -0.5462355  0.6359038 -
0.3870087
```

The lambda effects in additive model greater than 0 show, that there will be bigger than the average number of cases expected in the cell, while lambda less than 0, that there will be smaller than the average number of cases expected in that cell. With the use of `exp()` function we obtain thus, parameters of multiplicative model.

V. CONCLUSIONS

Contingency tables are an important source of categorical data in economic research. The main task of a researcher is to measure the association (relationship) between variables. Odds-ratio is surprisingly little known in economic sciences, where categorical data are used. It has several desirable properties: easy interpretation, it can be extended to multidimensional tables and it provides a helpful heuristic device for understanding log-linear analysis, where odds-ratios are functions of model parameters. In this paper odds-ratio was used to measure association in three-way table. Data on accidents in Polish mines coming from the National Mining Authority were presented in the paper. Three-way table was analyzed with the use of odds-ratio. It shows that for best fitting model we can use odds ratio, as well as to interpret the parameters of the model.

REFERENCES

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

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

Cornfield J. (1956). A statistical problem arising from retrospective studies*. in: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* ed. Neyman J.. vol. 4. 135-148. Berkeley. University of California.

Fienberg S. (1980). *The Analysis of Cross-classified Categorical Data. 2nd Edition*. M.I.T. Press. Cambridge.

Knoke D.. Burke P. J. (1980). *Log–linear models.* Quantitative Applications in the Social Science". nr 20. Sage University Papers. Sage Publications. Newbury Park. London. New Delhi.

Le C. T. (2010). *Applied categorical data analysis and translational research*. Wiley.

Mosteller F. (1968). *Association and estimation in contingency tables*. J. Amer. Assoc. 63. 1-28.

Rudas T. (1998). Odds ratios in the analysis of contingency tables. Sage Publications. no. 119.

Yule G. U. (1900). On *the association of attributes in statistics.* Phil. Trans. Ser. Q 194. 257-319.

Yule G. U. (1912). *On the methods of measuring association between two attributes*. J. Roy. Statist. Soc. 75. 579-642.

Justyna Brzezińska

ILORAZ SZANS W ANALIZIE TABLIC KONTYNGENCJI

Analiza zależności w statystyce stanowi jeden z podstawowych tematów badawczych. Celem artykułu jest zaprezentowanie ilorazu szans jako narzędzia opisu tablic kontyngencji, a także parametrów modelu logarytmiczno-liniowego. Iloraz szans jest miernikiem wykorzystywanym do badania związku w tablicach kontyngencji. Miernik ten można także uogólnić do tablic wielodzielczych poprzez użycie lokalnych ilorazów szans oraz podejścia opartego na krzyżowaniu komórek. Zaprezentowane zostaną własności ilorazu szans, a także ich związek z parametrami interakcji w analizie logarytmiczno-liniowej. Prezentacja ilorazu szans w części empirycznej zostanie zaprezentowana przy użyciu programu **R**.