talenta PUBLISHER

Journal of Research in Mathematics Trends and Technology

Loglinear Model Formation using Hierachial Backward Method

Siti Fatimah Sihotang[1*], Zuhri[1]

[1]Universitas Potensi Utama, Medan, 59391, Indonesia

Abstract. The loglinear model is a special case of a general linear model for poisson distributed data. The loglinear model is also a number of models in statistics that are used to determine dependencies between several variables on a categorical scale. The number of variables discussed in this study were three variables. After the variables are investigated, the formation of the loglinear model becomes important because not all the model interaction factors that exist in the complete model become significant in the resulting model. The formation of the loglinear model in this study uses the Backward Hierarchical method. This research makes loglinear modeling to get the model using the Hierarchical Backward method to choose a good method in making models with existing examples. From the challenging examples that have been done, it is known that the Hierarchical Reverse method can model the third iteration or scroll. Then, also use better assessment methods about faster workmanship and computer-sponsored assessments that are used more efficiently through compatibility testing for each model made.

Keyword: Backward Hierarchical Method, Categorical Scale, Loglinear Model, Modeling, Poisson Distribution

Abstrak. *Model loglinear merupakan kasus khusus dari model linear umum untuk data berdistribusi Poisson. Model loglinear juga merupakan sejumlah model dalam statistik yang digunakan untuk menentukan ketergantungan antara beberapa variable pada suatu skala berkategori. Setelah variabel-variabel diselidiki, formasi dari model loglinear menjadi penting karena tidak semua faktor interaksi model yang ada pada model lengkap menjadi signifikan pada model hasil. Formasi model loglinear pada pembelajaran ini menggunakan metode Hierachical Backward. Penelitian ini menghasilkan pemodelan loglinear untuk memperoleh model menggunakan metode Hierachical Backward untuk memilih suatu metode yang bagus dalam pembuatan model dengan contoh yang sudah ada. Dari contoh-contoh yang menantang yang telah diselesaikan, diketahui bahwa metode Hierarchical Reverse dapat memodelkan iterasi atau putaran ketiga. Kemudian, juga menggunakan metode penilaian yang lebih baik melalui pengujian yang sesuai untuk setiap model yang dibuat.*

Kata Kunci:*Metode Backward Hierarchical, Skala Berkategori, Model Loglinear, Pemodelan, Distribusi Poisson*

Received 9 December 2019 | Revised 10 January 2020 | Accepted 8 February 2019

*Corresponding author at: Universitas Potensi Utama, Medan, 59391, Indonesia

E-mail address: siti.fatimah.sihotang@gmail.com

Copyright ©2020 Published by Talenta Publisher, e-ISSN: 2656-1514
Journal Homepage: http://talenta.usu.ac.id/jormtt

1. Introduction

The Loglinear Model is one of the special cases of the general linear model for Poisson distributed data. This distribution was first introduced by Siméon-Denis Poisson (1781-1840). Poisson opportunity distribution is believed to be one of the three most important distributions, the other two being binomial and normal distribution [1]. Poisson distribution is an opportunity distribution of the number of times random events occur [2].

The loglinear model is an extension where the natural logarithm of the frequencies for each cell is equal to the mean (constant, mu) plus the lambda parameter to estimate the first independent effect plus the lambda for each other independent, as well as the lambda added for all the interaction effects be it the 2-factor interaction effect , 3-factor and interaction effects for higher orders according to independent quantities. Thus, a model of this type is also called a complete model [3]. The loglinear model is also referred to as a statistical model that is useful for determining dependencies (trends) between several variables on a categorical scale. The logline model is very dependent on the number of variables to be analyzed. The use of the variables discussed in this study are grouped into two types namely the dependent variable and the independent variable. In the loglinear model, there is an assumption that all the variables investigated have the same status as the dependent variable. In other words, no distinction is made between the dependent variable and the independent variable, this is due to the loglinear model which shows the dependencies (tendencies) between the variables.

Loglinear model analysis depends on the number of dependent variables contained in it. In this study, we will describe the loglinear model which discusses the analysis of the relationship between the three variables called trivariate analysis and the model is termed a three-factor logline model. The three-factor logline model contains all possible parameters and cannot be entered by other parameters. Models of this type are also called complete models. In general, the three-factor logline model can be written as follows [4]:

$$\log \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \quad (1)$$

With:

$\log \hat{m}_{ijk}$ = Logarithm of ijk cell frequency

μ = Logarithmic constant or average of all ijk cells

λ_i^X = The parameter of the effect of the first i-th variable on the model

λ_j^Y = The parameter of the influence of the second variable j to the model

λ_k^Z = The parameter of the influence of the third variable k to the model

λ_{ij}^{XY} = The parameters of the influence of the interaction of the first variable i to the second variable j to the model

λ_{ik}^{XZ} = The parameter of the influence of the interaction of the first i-th variable with the third variable k to the model

λ_{jk}^{YZ} = The parameters of the interaction effects of the second variable jth and third variable k to the model

λ_{ijk}^{XYZ} = The parameters of the interaction effects of the first variable to-i, the second variable to-j and the third variable to-k towards the model

With the loglinear approach, numbers in cells can be arranged in contingency tables. Friendly [5] states a contingency table is used when there is more than one categorical variable, which is usually the data presented in a list of rows and columns. This form of presentation in rows and columns is usually called a contingency list. Contingency table analysis is a data preparation technique that is simple enough to see the relationships between variables in one table. To interpret the data in contingency tables, one of the statistical tests that can be used is the Chi-Square test.

Chi-Square Test is symbolized by, which is used to find out whether there is a relationship between variables that are measured significantly or not. In this case the analysis of the measured variables is as many as three variables or what is referred to as a trivariat analysis. The hypothesis applies to the three independent variables assuming there is no interaction between variables, namely:

$$H_0: P_{ijk} = P_{i..}P_{.j.}P_{..k} \quad (2)$$

$$H_1: P_{ijk} \neq P_{i..}P_{.j.}P_{..k} \quad (3)$$

Chi-Square test statistics used to test the relationship between variables can be formulated as follows:

$$\chi^2 = \sum_i \sum_j \sum_k \frac{\left(Y_{ijk} - \hat{m}_{ijk}\right)^2}{\hat{m}_{ijk}} \quad (4)$$

While the Likelihood Ratio test for the independent model can also be formulated as follows:

$$G^2 = 2 \sum_i \sum_j \sum_k Y_{ijk} \ln\left(\frac{Y_{ijk}}{\hat{m}_{ijk}}\right) \quad (5)$$

With:

Y_{ijk} = Observations on variables i, j, and k

$\hat{m}_{ijk}$ = Expected frequency for Y_{ijk}

Where the degree of freedom is (I-1) (J-1) (K-1) and taken $\alpha = 0.05$

Test Criteria:

Reject H_0 if χ^2 or G^2 count $\geq \chi^2_{(df;\alpha)}$ and accept H_0 if χ^2 count $< \chi^2_{(df;\alpha)}$ in other words the model $\log \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$ is accepted.

Surely it will be a problem if all interactions are included simultaneously in the model without first knowing the suitability of the interaction of existing factor models before forming a model that is indeed appropriate and significant after passing the model compatibility test. In this case loglinear modeling techniques will be carried out in shaping the problem model. The type of procedure that the author uses in completing the loglinear modeling is a procedure that uses the Backward Hierarchical method. For the purposes of analyzing data that contains the relationship of three variables using a categorical scale focused on how the determination of the model by testing the interaction of both 2-factor and 3-factor model factors using the goodness of fit test of the model to test the suitability of the model [8].

The problem that will be discussed in this research is to carry out loglinear modeling procedures using the Backward Hierarchical method, in terms of obtaining a final model that describes the suitability of the relationships between the variables investigated. The purpose of this study is to carry out the procedure of loglinear modeling using the Backward Hierarchical method in forming the model, then find out the advantages and disadvantages of this method based on existing case examples. The benefits of this research are expected to be able to apply contingency table analysis and existing loglinear modeling methods for solving problems related to various fields of life so that they are more easily processed, so that the objectives are clearly in accordance with the scope of the problem to classify the data to be investigated in deciding an issue related to interaction between factors of categorical variables.

Simulation is a numerical technique to do an experiment in computer involving a certain mathematical and logical model displaying business characteristic and economic system in a long time period [5]. Taylor (2001) [8] simulation is a tool to analyze the inventory system where a demand is a random variable reflecting an uncertain demand. A survey conducted in 1978 by The Institute Management Science (TIMS) and The Operations Research Society of America (ORSA) in America informs that a simulation on the third rank after analysis of economics and statistics.

This information states that a simulation is a tool or method that can be used to solve the problem and give a solution. The reliability of simulation can face the complex problem, measure a performance from a various data and give the alternative solution quickly using computer program. One of simulation model is Monte Carlo simulation. Djati (2007) [2] stating that a Monte Carlo simulation will give an indication of several inventory that must be at store and when the order conducted.

2. Research Methodology

The method used in this study uses the Backward Hierarchical method for loglinear modeling. The search for solutions to loglinear modeling problems can be done by determining a model flexibly and deeply and choosing the independent variables appropriately inclusively. This makes it possible to find the best independent variables that can be used as well as looking at the suitability of the model that considers the presence or absence of interaction between variables.

3. Analysis and Discussion

The following is a three-factor contingency table that aims to find a model that can state the relationship in the data set precisely, which is as follows:

Table 1 Hypothesis Table Data for Multifactorial Frequency Analysis

Profesision	Gender	Type of reading		Total
		Scientific Fiction	Novel	
Politician	Male	38	25	63
	Female	20	15	35
Dancer	Male	12	27	39
	Female	18	30	48
Total		88	97	185

Solution :

The initial step taken before modeling is to divide the three existing categorical variables according to their respective scale types, namely:

a. The first variable (I) is Profession. The profession is measured through a nominal measurement scale with a categories: 0 (if the profession is a politician) and 1 (if the profession is as a dancer).
b. The second variable (J) is the Reading Type. The type of reading is measured through a nominal measurement scale with categories: 0 (if the type of reading read is science fiction), and 1 (if the type of reading read is novel).
c. The third variable (K) is Gender. Gender is measured using a nominal measurement scale with categories: 0 (if the gender is male), and 1 (if the gender is female).

Furthermore, based on the contingency table above, as an initial step an independence test is performed to determine whether there is a significant relationship between the profession, type of reading read and gender with the following hypotheses:

H_0 : There is no relationship between profession, type of reading read and gender

$$\left(P_{ijk} = P_{i..}P_{.j.}P_{..k}\right)$$

H_1 : There is a relationship between professions, types of readings read and gender

$$\left(P_{ijk} \neq P_{i..}P_{.j.}P_{..k}\right)$$

From the calculation results that have been obtained through SPSS software version 22, score χ^2 or G^2 count of 18,852 or 19,024 compared to $\chi^2_{(1;0,05)} = 3{,}841$. Because of value χ^2 or G^2 count is greater that $\chi^2_{(1;0,05)} = 3{,}841$, H_0 rejected so that it can be said there is a relationship between professions, types of readings read, and gender. After it is known that there is a relationship between the three variables investigated, loglinear modeling procedure will be carried out to see further the other components of the model that will be significant with the model that will be formed because the test results obtained above have not been sufficient to provide information about it.

The information that is not yet known is about whether there is an interaction between professions, types of reading, and gender. Therefore, loglinear modeling used in this study to form the model is using the Backward Hierarchical method. In this modeling, a complete loglinear model must be formed in advance for three factors which in this model the expected cell frequency will always be the same value as the observed frequency, without the remaining degree of freedom [7].

3.1. Loglinear Modeling uses the Backward Hierarchical Method

Loglinear modeling with the Backward Hierarchical method will form a hierarchical model that states the relationships in the data set appropriately. This is done by selecting the complete model and starting to remove the higher interactions until the fit test of the model becomes unacceptable based on the probability standard or p-value adopted by the investigator. The complete model includes all possible interaction effects both 2-factor and 3-factor interaction effects according to the number of variables investigated in this study. Where every time a variable is removed a statistical test is performed to determine the accuracy of its predictions by comparing it to the likelihood ratio test [8].

Table 2 Table Process of Elimination Hierarchical Backward

Phase	Model	G^2	df
1	$\log \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$	0,000	0
2	$\log \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	0,475	1
3	$\log \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$	**0,527**	**2**

From the above results several steps were taken to form a hierarchical model, namely:

Step 1: For example the **model (0)** = $\log \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$ is the best model.

Step 2: By issuing the interaction of three factor from the model, obtained a model (1),

Model (1) = $\log \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$.

Step 3: Conduct a Statistical Test with the following hypothesis:

H_0 : model (1) = the best model

H_1 : model (0) = the best model

The conditional test statistics are: $G^2_{(1-0)} = G^2_{(1)} - G^2_{(0)}$

Where : $G^2_{(1)}$ = likelihood statistics for the model (1)

$G^2_{(0)}$ = likelihood statistics for the model (0)

It is known from Table 2 values $G^2_{(1)} = 0{,}475$ dan $G^2_{(0)} = 0{,}000$,

then obtained: $G^2_{(1-0)} = G^2_{(1)} - G^2_{(0)} = 0{,}475 - 0{,}000 = 0{,}475$.

Degree of freedom = degree of freedom of the model (1) - degree of freedom of the model (0) = 1 - 0 = 1. Score $\chi^2_{(1;0{,}05)} = 3{,}841$.

Step 4: From the calculation results that have been obtained, score $G^2_{(1-0)}$ compared with $\chi^2_{(1;0{,}05)} = 3{,}841$.

Because $0{,}475 < 3{,}841$, then H_0 is accepted which states **model (1) as the best model.**

Step 5: Form model (2) the model obtained from model (1) if one of from the interaction two factors ejected from model.

Based on the results of the interaction associations that exist in the model, the interaction with values G^2 the smallest is the addition of interactions λ^{YZ}_{jk} which produces a value of 0.052 (the acquisition of values obtained from the SPSS output version 22), then the model with the addition of interaction λ^{YZ}_{jk} must be removed first [4] so that a model (2) will be formed, namely:

model (2) = $\log \hat{m}_{ijk} = \mu + \lambda^X_i + \lambda^Y_j + \lambda^Z_k + \lambda^{XY}_{ij} + \lambda^{XZ}_{ik}$.

In the same way as above, a Conditional Test Statistic will be conducted with the following hypothesis:

H_0 : model(2) = the best model

H_1 : model(1) = the best model

The conditional test statistics are : $G^2_{(2-1)} = G^2_{(2)} - G^2_{(1)}$

With : $G^2_{(2)}$ = Statistics likelihood G^2 for model (2)

$G^2_{(1)}$ = Statistics likelihood G^2 for model (1)

It is known from table 2 above the value, $G^2_{(2)} = 0{,}527$ and $G^2_{(1)} = 0{,}475$;

So obtained : $G^2_{(2-1)} = G^2_{(2)} - G^2_{(1)} = 0{,}527 - 0{,}475 = 0{,}052.$

Degree of freedom = degree of freedom of the model (2) - degree of freedom of the model (1) = 2 - 1 = 1. Value of $\chi^2_{(1;0{,}05)} = 3{,}841.$

It turned out that it was obtained $G^2_{(2-1)} < \chi^2_{(1;0{,}05)} = 0{,}052 < 3{,}841$, so H_0 is accepted which states model (2) as the best model. **model (2)** = $\log \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$**.** The model formed states that there is a dependency between the first variable with the second variable and the first variable with the third variable so that by inserting the elements of the variable it can be said that this model states a dependency between the profession with reading type and the profession with gender.

Based on the results of modeling with the Backward Hierarchical method that has been done for the example above turns out to provide results to shape the problem model, namely:

$$\log \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} \quad (6)$$

With :

$\log \hat{m}_{ijk}$ = Logarithm of ijk cell frequency

μ = Logarithm average of all ijk cells

λ_i^X = The parameter of the influence of the i-th Professional (i = 1 (politician) and i = 2 (dancer)) on the model

λ_j^Y = The parameters of the influence of the type of reading j (j = 1 (science fiction) and j = 2 (novel) on the model

λ_k^Z = The parameters of the influence of the sex k (k = 1 (male) and k = 2 (female)) on the model

λ_{ij}^{XY} = The parameter of the influence of the i-th Professional interaction with the j-th Read Type on the model

λ_{ik}^{XZ} = The parameters of the influence of the i-th Professional interaction with the Third Gender on the model

In other words, this model states that there are significant dependencies between professions with reading types and professions with gender.

4. Conclusion

Based on the results of research and discussion in previous chapters, the authors try to draw conclusions as follows:

Loglinear modeling can be done using the Backward Hierarchical method. The advantage of the Backward Hierarchical method is that it is faster to eliminate the components of the model because the process of elimination starts from the model that has the highest association known as the complete model to the model that contains the lower association. While the lack of a Backward Hierarchical method is that the selection process will be difficult if an increasing number of variables will be investigated because there will be an increasingly rapid increase in identifying associations and interactions that exist in each model that is formed.

REFERENCES

[1] M. F. Al-Saleh, "Teachers' corner A rich learning lesson using the Poisson distribution," *Stat. Methodol.*, vol. 4, pp. 504–507, 2007, doi: 10.1016/j.stamet.2007.01.005.

[2] V. Buonaccorsi and A. Skibiel, "A 'Striking' Demonstration of the Poisson Distribution," *Teach. Stat.*, vol. 27, no. 1, pp. 8–10, 2005.

[3] J. K. Vermunt, *LEM: Log-Linear Modelling, User's manual*, Department of Methodology and Statistics, Netherlands: Tilburg University Press, 2005.

[4] A. Agresti, *Categorical Data Analysis*, Canada: John Wiley & Sons, Inc, 1990.

[5] M. Friendly, *Visualizing Categorical Data*, SAS Press, Cary, NC, 2000.

[6] C. M. Friel, *Hierarchical Loglinear Analysis: A Statistical Technique for the Analysis of Frequency Data in Multiway Cross-Tabulation Cross*, USA: Criminal Justice Center, Sam Houston State University, 2005.

[7] A. Jeansonne, *A History Loglinier Models*, Introductory Topics, New York, USA, 2002.

[8] J. H. Maindonald, *Statistical Computation*, New Zealand: John Wiley & Sons, Inc, 2001.