Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Vol 1, No. 1, 2023, 1-7
ISSN XXXX-XXXX (Print)
ISSN XXXX-XXXX (Online)

Analysis of Linear Log Models on Covid-19 Data in Indonesia

Indah Suciati[1*], Warsono[2]

[1*,2] Department of Mathematics, Universitas Lampung, Lampung, Indonesia
[*]corresponding author: indahsuciati222@gmail.com

Received December 22, 2022; Received in revised form December 28, 2022; Accepted January 1, 2022

Abstract. Covid-19 is still a concern of the world, including Indonesia. The transmission of Covid-19 is very fast and has a wide impact on all people around the world, especially Indonesia. In everyday life, we find a lot of data that looks into a certain category. Categorical analysis of data can be done using the log linear model. The log linear model is used to analyze the relationship between categorical variables that form a contingency table of arbitrary dimensions. The analysis used in this study is to make descriptive statistics and three-way contingency tables, then perform the analysis with the help of SPSS 25.0 software where the goodness of fit test is used to see which models can be used or suitable. The purpose of this study is to analyze a log linear model, so that a log linear model is obtained that is suitable for Covid-19 data based on gender, province, and age group. The conclusion of this study is that of the 9 modexls used, the model (XY, XZ, YZ) is the most suitable model to be used, with a G^2 value of 18,885 and the equation of the log linear model is $log\, m_{ijk} = \mu + \lambda_i^{XY} + \lambda_j^{XZ} + \lambda_k^{YZ}$, which means that there is a relationship between the two factors for the variables gender and province (XY), gender and age group (XZ), and province and age group (YZ), in Covid-19 cases in Covid-19 in Indonesia by gender, province, and age group.
Keywords: Categorical Data Analysis, Log Linear Models, Covid-19

INTRODUCTION

Coronavirus Disease or Covid-19 is still a concern for the world, including Indonesia. Covid-19 first appeared in Wuhan, China in December 2019. Covid-19 is caused by a new strain of the Coronavirus, namely Novel Coronavirus 2019 (2019-nCoV) and officially named as Severe Acute Respiratory Syndrome-Coronavirus 2 (SARS-CoV- 2) (Bedford et al., 2020). Covid-19 is transmitted through droplets or splashes that come out when someone who is infected coughs, sneezes, or talks (Tian et al., 2020). The transmission of Covid-19 is very fast and has a broad impact on all people around the world, especially Indonesia.

In everyday life, there is a lot of data that is grouped into a certain category. Data that consists of several categories is called categorical data, for example the type of work that is divided into: civil servants and private employees (Lestyorini, 2010). Categorical data is sample observation data in a population that has similar conditions that are cross-grouped into several categorical variables (Fienberg, 2007). Analysis of categorical data is applied in a table that describes the frequency of observations that occur at the level of various

Vol 1, No. 1, 2023, 1-7
ISSN XXXX-XXXX (Print)
ISSN XXXX-XXXX (Online)

combinations of a variable. Tables that apply categorical data are called contingency tables. The contingency table method can answer the relationship between two, three or more research variables but not a causal relationship (Agresti, 2002). Categorical data analysis can be performed using a linear log model. The linear log model is used to analyze the relationship between the categorical variables that make up the contingency table of any dimension.

Various studies on the linear log model have been carried out, as was done by Sari et al. (2016) about the relationship between fuel type, vehicle type, engine compression ratio, and engine capacity, Maryana (2013) about the relationship between gender and education, Sihotang & Zuhri (2020) about the relationship between profession, gender, and type of reading, and others. As for research for the Covid-19 case study, namely research that has been conducted by Ai et al. (2020) to identify Covid-19 based on different age groups (<60 years and 60 years) and gender in China, Zhao et al. (2020) to investigate the relationship between CT scan findings and the clinical condition of Covid-19 pneumonia, Li et al. (2020) who showed a higher distribution of Covid-19 disease for male sex, Lippi & Henry (2020) about the relationship between smoking and the severity of Covid-19, Liu et al. (2020) regarding Covid-19 for elderly patients more likely to develop severe disease, Munayco et al. (2020) classified Covid-19 cases and the number of deaths by age and sex in Peru, and Altun (2021) on the relationship of sex, country, and age group. As seen from the literature review above, Covid-19 is highly dependent on variables such as age, gender, presence of chronic diseases, and country of residence.

Based on the description above, the authors are interested in conducting research on the analysis of linear log models on Covid-19 data in Indonesia based on gender, province, and age group, with the parameters used to evaluate the model, namely the G^2 value.

METHOD

The data used in this study is Covid-19 data for 2020 based on gender, province, and age group in Indonesia, sourced from the 2020 Indonesian Health Profile Catalog Book published by the Ministry of Health of the Republic of Indonesia in 2021 (Indonesia, 2021). The gender variable consists of male and female gender, the provincial variable consists of the Provinces of Lampung, South Sumatra and West Sumatra, and the age group variable consists of the age group 0-15 years, 16-30 years, 31-45 years, 46-60 years, and >60 years. These data were then cross-classified in three directions for further analysis using a linear log model.

The linear log model is a model for obtaining a statistical model which states the relationship between variables and qualitative data (nominal or ordinal scale) [17]. A contingency table or what is often called a cross tabulation (cross tabulation or cross classification) is a table that contains data on the number or frequency or several classifications (categories). The contingency table method can answer the relationship of two or more research variables but not a causal relationship, the more the number of variables tabulated, the better the interpretation (Wulandari et al., 2009). To perform a linear log model analysis, this study will use a three-dimensional contingency table.

A three-dimensional table that has $(i \times j \times k)$ cells, consisting of i rows, j columns, and k layers, which is then referred to as an i×j×k contingency table [17]. The $i \times j \times k$ contingency table is presented in Table 1 as follows.

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Vol 1, No. 1, 2023, 1-7
ISSN XXXX-XXXX (Print)
ISSN XXXX-XXXX (Online)

Table 1. The $i \times j \times k$ Contingency Table

Variable 1 (X)	Variable 2 (Y)	Variable 3 (Z)				Total
		Z_1	Z_2	$\cdots$	Z_k	
X_1	Y_1	n_{111}	n_{112}	$\cdots$	n_{11k}	n_{11+}
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	Y_j	n_{1j1}	n_{1j2}	$\cdots$	n_{1jk}	n_{1j+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_i	Y_1	n_{i11}	n_{i21}	$\cdots$	n_{i1k}	n_{i1+}
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	Y_j	n_{ij1}	n_{ij2}	$\cdots$	n_{ijk}	n_{ij+}
Total		n_{++1}	n_{++2}	$\cdots$	n_{++k}	n_{+++}

The independent log linear model for the three variables is as follows [17]:

$$log\, m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \quad (1)$$

with:

μ : the logarithm of the sum of the expected values or the average of all the logarithms of the expected values.

$$\mu = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\log m_{ijk}}{IJK}$$

λ_i^X : the effect of variable 1 on the model.

$$\lambda_i^X = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}\log m_{ijk}}{JK} - \mu$$

λ_j^Y : the effect of variable 2 on the model.

$$\lambda_i^Y = \frac{\sum_{j=1}^{I}\sum_{k=1}^{K}\log m_{ijk}}{IK} - \mu$$

λ_i^Z :the effect of variable 3 on the model.

$$\lambda_i^Z = \frac{\sum_{j=1}^{I}\sum_{k=1}^{J}\log m_{ijk}}{IJ} - \mu$$

and the degrees of freedom are as follows.

$$\begin{aligned} db &= db\left(\log m_{ijk}\right) - \left(db(\mu) + db\left(\lambda_i^X\right) + db\left(\lambda_i^Y\right) + db\left(\lambda_i^Z\right)\right) \\ &= IJK - ((1) + (I-1) + (J-1) + (K-1)) \\ &= IJK - I - J - K + 2 \end{aligned}$$

To determine whether the model is appropriate or not, the goodness of fit test will be used. The goodness of fit test can use two test statistics, namely the Chi-Square statistic or the Likelihood Ratio Square. Chi-Square statistics are also used to determine whether or not there is a significant relationship between the variables being measured. The Chi-Square statistic is used to test the hypothesis that the expected population frequency meets a certain model, namely by using the Likelihood Ratio Test (G^2) or the Pearson Chi-Square test (χ^2) (Agresti, 1990). The test steps are as follows:

a. Hypothesis
   H_0 : The modelcorresponds to the actual situation
   H_1: The model does not match the actual situation
b. Significance level
   $\alpha = 0.05$
c. Test Statistics
   There are two test statistics that can be used, namely as follows:

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Vol 1, No. 1, 2023, 1-7
ISSN XXXX-XXXX (Print)
ISSN XXXX-XXXX (Online)

1) Likelihood Ratio Test (G^2) : $G^2 = \sum_{ijk} n_{ijk}\, log\left(\frac{n_{ijk}}{m_{ijk}}\right)$
2) Pearson Chi-Square (χ^2) : $\chi^2 = \sum_{ijk}\left[\frac{(n_{ijk}-m_{ijk})^2}{m_{ijk}}\right]$

d. Rejection Criteria
1) Reject H_0 if $G^2_{hit} > \chi^2_{(\alpha,df)}$ or $p - value < \alpha$
2) Reject H_0 if $\chi^2_{hit} > \chi^2_{(\alpha,df)}$ or $p - value < \alpha$

In determining the best model in the analysis of linear log models, the model must meet the following criteria [3]:
a. Meet the goodness of fit test.
b. Easy to interpret or translate.
c. Simplest model possible

The steps for the analysis of the linear log model used in this study are as follows:
1) Make descriptive statistics with the help of SPSS 25.0 software to find out the characteristics of Covid-19 in Indonesia.
2) Create a three-dimensional contingency table with the help of SPSS 25.0 software to measure the relationship (association) between the three variables studied.
3) Perform linear log analysis with the help of SPSS 25.0 software to form the best three-dimensional linear log model.
   a) Describes the models that may be formed with three variables, model building starts from the simplest model to the most complete model.
   b) Conduct goodness of fit tests or model significance tests to see which models can be used or which models are suitable.

RESULT AND DISCUSSION

In this study, a linear log model analysis will be carried out for Covid-19 data for 2020 based on gender, province, and age group in Indonesia. Then the linear log model will be evaluated using the G^2 value. The data used in this study are presented in Table 2 as follows.

Table 2. Covid 19 Data Based on Gender, Province, and Age Group

gender (X)	province (Y)	age group (Z)					Total
		0-15 years	16-30 years	31-45 years	46-60 years	>60 years	
M	Lampung	205	275	807	766	402	2455
	Sumatera Selatan	570	552	1781	1340	836	5079
	Sumatera Barat	1054	1235	2981	2293	1177	8740
F	Lampung	211	257	1002	752	318	2540
	Sumatera Selatan	539	554	1673	1193	658	4617
	Sumatera Barat	1241	1531	3491	2736	1155	10154
	Total	3820	4404	11735	9080	4546	33585

Descriptive Analysis

Based on the output obtained, out of 33,584 people infected with Covid-19, the male sex in Lampung and West Sumatra Provinces has the same percentage order for each age group, with the percentage order of the age group infected with Covid-19 from largest to lowest.

Vol 1, No. 1, 2023, 1-7
ISSN XXXX-XXXX (Print)
ISSN XXXX-XXXX (Online)

the smallest, namely the age group 31-45 years, 46-60 years, >60 years, 16-30 years, and 0-15 years. However, in the province of South Sumatra the age group 0-15 years has a higher percentage compared to the age group 16-30 years. From this we can conclude that the age group of 31 years and over with male sex in the Provinces of Lampung, South Sumatra and West Sumatra are more at risk of being infected with the Covid-19 virus.

As for the female sex, the Provinces of Lampung and South Sumatra have the same percentage order for each age group, with the percentage order of the age group infected with Covid-19 from largest to smallest, namely the age group 31-45 years, 46-60 years, > 60 years, 16-30 years, and 0-15 years. However, in the province of West Sumatra the age group >60 years has a smaller percentage compared to the age groups 0-15 years and 16-30 years. From this we can conclude that for ages 31-60 years with female sex in the Provinces of Lampung, South Sumatra and West Sumatra are more at risk of being infected with the Covid-19 virus.

Besides that, based on the Chi-Square test we get that the p-value = .000. Because the p-value is smaller than the significant level $\alpha = 5\%$ (0.005), we reject H_0. So we can conclude that at the real level $\alpha = 5\%$there is a relationship between gender, province, and age group.

Goodness of Fit Test

The goodness of fit test was carried out to determine whether the model was significant or not. In this study 9 models will be used, then the best model will be selected to be used. The statistical test used is the likelihood ratio test (G^2). It is known that the results of the goodness of fit test with a significance level of $\alpha = 0.05$ are as follows.

Table 3. Comparison of G^2, db, and p-values for Each Model

Model	G^2 value	db	$P-value$
X, Y, Z	371.807	22	.000
X, YZ	160.085	14	.000
Y, XZ	322.310	18	.000
Z, XY	274.526	20	.000
XZ, YZ	110.588	10	.000
XY, YZ	62.804	12	.000
XY, XZ	225.028	16	.000
XY, XZ, YZ	18.885	8	.015
XYZ	.000	0	.000

Based on Table 2 above, we get that the G^2 and p-values for the model (XY, XZ, YZ) are 18,885 and .015. Because the value of G^2 is relatively small and the p-value is greater than the significance level $\alpha = 5\%$ (0.005), we do not reject H_0. So we can conclude that the model (XY, XZ, YZ) is a model that fits the model equation (XY, XZ, YZ), which is as follows.

$$log\, m_{ijk} = \mu + \lambda_i^{XY} + \lambda_j^{XZ} + \lambda_k^{YZ}$$

From this model, it can be interpreted that there is an interaction between the two factors for the variables gender and province (XY), gender and age group (XZ), and province and age group (YZ) in the case of Covid-19 in Indonesia based on gender, province, and age group.

Vol 1, No. 1, 2023, 1-7
ISSN XXXX-XXXX (Print)
ISSN XXXX-XXXX (Online)

CONCLUSION

Based on the results and discussion in the previous chapter, it was concluded that of the 9 models used, the model $(XY, XZ, YZ$ was the most suitable model to use, with a G^2 value of 18,885 and the log linear model equation was obtained, namely $log\ m_{ijk} = \mu + \lambda_i^{XY} + \lambda_j^{XZ} + \lambda_k^{YZ}$, which means that there is a relationship between the two factors for the variables gender and province (XY), gender and age group (XZ), as well as province and age group (YZ) in cases of Covid-19 in Indonesia by gender, province and age group.

REFERENCE

Agresti, A. (1990). *Categorical data analysis*. John Wiley & Sons.

Agresti, A. (2002). Categorical Data Analysis. John Wiley & Sons. *Inc., Publication*, *15*, 24. https://doi.org/10.1002/0471249688

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., & Xia, L. (2020). Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology*. https://doi.org/10.1148/radiol.2020200642

Altun, G. (2021). A Study On Covid-19 Data With Log-Linear Model Approach. *Mugla Journal of Science and Technology*, *7*(1), 52–58. https://doi.org/10.22531/muglajsci.835562

Bedford, J., Enria, D., Giesecke, J., Heymann, D. L., Ihekweazu, C., Kobinger, G., Lane, H. C., Memish, Z., Oh, M., & Schuchat, A. (2020). COVID-19: towards controlling of a pandemic. *The Lancet*, *395*(10229), 1015–1018. https://doi.org/10.1016/S0140-6736(20)30673-5

Fienberg, S. E. (2007). *The analysis of cross-classified categorical data*. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-72825-4

Indonesia, K. K. R. (2021). *Profil Kesehatan Indonesia 2020, Kementrian Kesehatan Republik Indonesia*. Availableat: Ttps://Pusdatin. Kemkes. Go. Id/Resources/Download/Pusdatin

Lestyorini, M. (2010). *Model Log Linear Multivariat Empat Dimensi (Studi Kasus : Akses Internet Mahasiswa Jurusan Pendidikan Matematika di Universitas Negeri Yogyakarta)*. Fakultas MIPA UNY, Yogyakarta.

Li, L., Huang, T., Wang, Y., Wang, Z., Liang, Y., Huang, T., Zhang, H., Sun, W., & Wang, Y. (2020). COVID-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis. *Journal of Medical Virology*, *92*(6), 577–583. https://doi.org/10.1002/jmv.25757

Lippi, G., & Henry, B. M. (2020). Active smoking is not associated with severity of coronavirus disease 2019 (COVID-19). *European Journal of Internal Medicine*, *75*, 107–108. https://doi.org/10.1016/j.ejim.2020.03.014

Liu, D., Li, L., Wu, X., Zheng, D., Wang, J., & Yang, L. (2020). Pregnancy and perinatal outcomes of women with coronavirus disease (COVID-19) pneumonia: a preliminary analysis. AJR. 2020; 1–6. *American Journal of Roentgenology*, 1–6. https://doi.org/10.2214/AJR.20.23072

Maryana, M. (2013). Model Log Linier yang Terbaik untuk Analisis Data Kualitatif pada Tabel Kontingensi Tiga Arah. *Industrial Engineering Journal*, *2*(2).

Munayco, C., Chowell, G., Tariq, A., Undurraga, E. A., & Mizumoto, K. (2020). Risk of death by age and gender from CoVID-19 in Peru, March-May, 2020. *Aging (Albany NY)*, *12*(14), 13869. https://doi.org/10.18632/aging.103687

Sari, J. S., Wilandari, Y., & Hoyyi, A. (2016). Pembentukan Model Log Linier Empat Dimensi (Studi Kasus: Rata-rata Pengguna Jenis Bahan Bakar Minyak berdasarkan Jenis

Vol 1, No. 1, 2023, 1-7
ISSN XXXX-XXXX (Print)
ISSN XXXX-XXXX (Online)

Kendaraan, Rasio Kompresi dan Kapasitas Mesin). *Jurnal Gaussian*, *5*(3), 437–446.

Sihotang, S. F., & Zuhri, Z. (2020). Analisis Model Log Linier Tiga Dimensi Untuk Data Kualitatif Dengan Metode Forward. *MES: Journal of Mathematics Education and Science*, *6*(1), 62–69.

Tian, H., Liu, Y., Li, Y., Wu, C.-H., Chen, B., Kraemer, M. U. G., Li, B., Cai, J., Xu, B., & Yang, Q. (2020). An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science*, *368*(6491), 638–642. https://doi.org/10.1126/science.abb6105

Wulandari, S. P., Salamah, M., & Susilaningrum, D. (2009). *Diktat pengajaran analisis data kualitatif*. Jurusan Statistika Institut Teknologi Sepuluh Nopember Surabaya, Surabaya.

Zhao, W., Zhong, Z., Xie, X., Yu, Q., & Liu, J. (2020). Relation between chest CT findings and clinical conditions of coronavirus disease (COVID-19) pneumonia: a multicenter study. *AJR Am J Roentgenol*, *214*(5), 1072–1077. https://doi.org/10.2214/AJR.20.22976