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

Vol. 1, No. 2, 2023, 56-64
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Log-Linear Model on Categorical Data of HIV Cases

Wardhani Utami Dewi[1*], Warsono[2]

[1*,2]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Indonesia
[*]corresponding author: dewiutamiwardhani@gmail.com

Received Mei 30, 2023; Received in revised form June ; Accepted July 6, 2023

Abstrak. *Data kategorik banyak digunakan pada penelitian sosial, kesehatan, pendidikan, dan psikologi. Tabel kontingensi adalah bentuk penyajian data tersebut. Salah satunya tentang kasus terinfeksi virus HIV. Model log-linear menjadi alternatif untuk menganalisis data ketegorik. Pada penelitian ini akan dianalisis menggunakan model log-linear kasus HIV yang dikelompokkan berdasarkan jenis kelamin, Age dan Province. Selain itu akan dibentuk beberapa model log-linear dan dipilih model terbaik berdasarkan uji statistik likelihood ratio (G^2). Menurut hasil analisis dan pertimbangan kompleksitas model, (**JK*P, JK*U, P*U**) merupakan model terbaik dan sesuai dengan data karena p-value =0.517 lebih besar daripada taraf nyata α=0.05. Artinya interaksi antara jenis kelamin, Age dan Province adalah signifikan. Studi dan penjelasan tentang virus HIV bahwa individu dengan Age antara 25-49 tahun lebih beresiko terinfeksi virus tersebut. Dikaji berdasarkan kelompok jenis kelamin, paling banyak terinfesi virus dialami oleh perempuan yaitu 513 orang. Selain itu juga, Papua Barat menjadi Province yang paling tinggi jumlah terinfeksi virus HIV dibandingkan dengan Maluku dan Maluku Utara.*

Kata kunci*:* data kategorik; model log-linear; virus HIV

Abstract. Categorical data is widely used in social, health, educational and psychological research. A contingency table is a form of presenting this data. One of them is about cases of being infected with the HIV virus. The log-linear model is an alternative for analyzing categorical data. In this study, HIV cases will be analyzed using a log-linear model grouped by gender, age and province. Apart from that, several log-linear models will be formed and the best model will be selected based on the likelihood ratio (G^2) statistical test. According to the results of the analysis and consideration of model complexity, (JK*P, JK*U, P*U) is the best model and fits the data because the p-value = 0.517 is greater than the real level α = 0.05. This means that the interaction between gender, age and province is significant. Studies and explanations about the HIV virus show that individuals between the ages of 25-49 years are more at risk of being infected with the virus. Examined by gender group, women were most infected with the virus, namely 513 people. Apart from that, West Papua is the province with the highest number of HIV infections compared to Maluku and North Maluku

Keywords: categorical data; log-linear model; HIV virus

Vol. 1, No. 2, 2023, 56-64
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

INTRODUCTION

Research in the fields of social, health, education and psychology often uses qualitative variables. The data collected is in the form of categorical data presented in the form of a contingency table. In statistics, you can use regression analysis or ANOVA to analyze relationships between variables, but in fact the regression principle is not capable of analyzing categorical data. In its development, the log-linear model emerged to overcome the limitations of the regression model (Carota et al., 2022; Mulugeta et al., 2022).

The concept of log-linear analysis in contingency tables is analogous to the concept of analysis of variance (ANOVA) for continuously distributed factor response variables (Altun, 2021a; Von Eye et al., 2012; Wang et al., 2022). Several researchers have applied the log-linear model to categorical data, including Altun (2021b) who conducted a multi-rater agreement analysis with a log-linear model. Carota et al. (2022) assessing semi-parameteric log-linear Bayesian models: application to the expression of risk estimates. Apart from that, Ali et al. (2021) researched investigating the interaction between age and gender on the influence of car users using a log-linear model: a Bayesian inference approach. Grover & Sharma (2018) also examined the effect of reducing predictors that influence the survival time of HIV/AIDS patients using log-linear. Some of these studies obtained the best model by conducting hypothesis tests based on goodness of fit values, namely likelihood ratio and chi-quare values.

One case that uses qualitative variables and categorical data is the HIV case. Human Immunodeficiency Virus (HIV). HIV is a virus that attacks and destroys the body's CD4 cells that fight infection in the immune system (Du et al., 2022; Grover & Sharma, 2018). The number of people infected with the HIV virus in Indonesia in 2020 was 543,100 people and 30,137 people were recorded as having died due to the virus (Kementerian Kesehatan Republik Indonesia, 2021). Overall, there are fewer HIV cases in Indonesia than men, with an age range of 25-49 years.

METHOD

The secondary data used is the number of HIV cases based on gender and age group from three provinces, namely Maluku, North Maluku and West Papua in 2020. The data source is from the publication of the Ministry of Health of the Republic of Indonesia in 2020. The analysis method used is a log-linear regression analysis, with research numerical results obtained using IBM SPSS 25.0 software. Three-dimensional cross table obtained according to gender (male=1, female=2), Province (Maluku=1, North Maluku=2, West Papua=3) and age (≤ 4 years = 1, 5-14 years= 2, 15-19 years=3, 20-24 years=4, 25-49 years=5, and ≥50 years=6). The best model was determined using the backward elimination method. After estimating the best model parameters.

When the case is more than two categorical variables, the use of the chi-square test of independence in determining the relationship between the variables in a contingency table becomes difficult or sometimes impossible. In this case, the log-linear model, which allows testing a much larger number of hypotheses compared to chi-square, which does not impose restrictions on the number of rows and columns in both two-dimensional tables where chi-square can be applied, and three-dimensional tables where chi-square is insufficient, it is preferred. Multidimensional contingency table in the log-linear model, a model is formed to

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

Vol. 1, No. 2, 2023, 56-64
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

investigate the relationship between variables, the parameters in the model are estimated and the significance of this model is tested (Abdallah, 2022; Fujisawa & Tahata, 2022). Overall model fit was assessed by comparing the expected frequencies with the observed cell frequencies for each model (Alzahrani, 2022; Asmare & Agmas, 2022).

Pearson's chi-square statistic or likelihood ratio can be used to test model fit. The likelihood ratio is more commonly used because it is a statistic that is minimized in maximum likelihood estimation (Agresti, 2003; Jamaludin et al., 2022). The test statistical forms of likelihood ratio and chi-square are respectively written as follows:

$$G^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{K} n_{ijk} \log\left(\frac{n_{ijk}}{E_{ijk}}\right)$$

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{K} \frac{(n_{ijk} - E_{ijk})^2}{E_{ijk}}$$

Information:
G^2 = likelihood comparison test statistic
χ^2 = chi-square test statistic
n_{ijk} = number of observations
E_{ijk} = expected value

If the model has a large number of observations then G^2 and χ^2 approach the chi-square distribution, where the number of columns and rows minus the number of free parameters that fit in the model will be equal to the degrees of freedom (Aliverti & Dunson, 2022). The criterion for testing the hypothesis is that if the calculated χ^2 is greater than the table or the p-value is smaller than the significance level (α = 0.05), then Ho is rejected (Maryana, 2013).

RESULT AND DISCUSSION

Below is a three-dimensional contingency table for the number of HIV cases based on gender, age and province.

Table 1. Number of HIV cases by gender, age and Province

Gender	Province	Age						Total
		≤ 4	5-14	15-19	20-24	25-49	≥50	
Male	Maluku	4	2	7	41	154	9	217
	Maluku Utara	9	2	3	23	78	10	125
	Papua Barat	3	2	7	20	113	13	158
	Total	16	6	17	84	345	32	500
Female	Maluku	5	2	5	44	100	5	161
	Maluku Utara	6	1	9	24	59	6	105
	Papua Barat	6	4	25	69	137	6	247
	Total	17	7	39	137	296	17	513

Source: Kementerian Kesehatan Republik Indonesia (2021)

When Table 1. is examined in terms of gender, it can be concluded that this disease occurs more often in women. This is explained further with the following diagram illustration.

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

Vol. 1, No. 2, 2023, 56-64
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

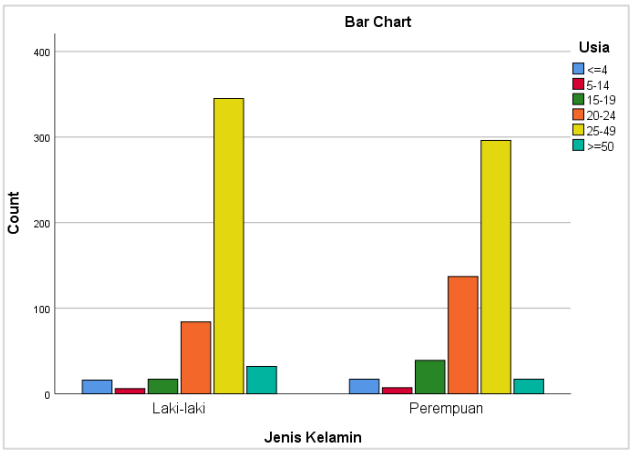

Figure 1. Number of HIV cases by gender and age

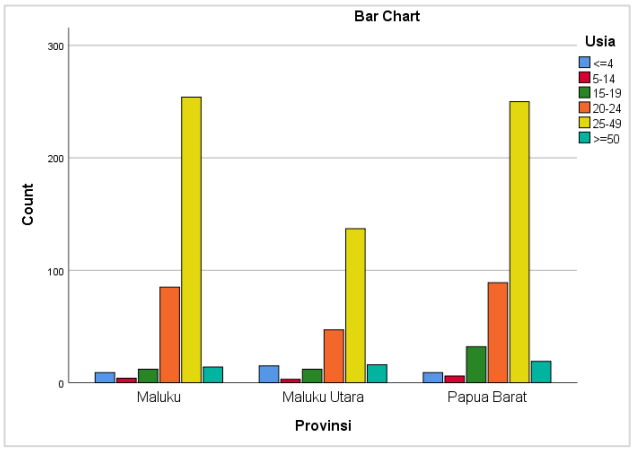

Figure 2. Number of HIV cases by gender and age

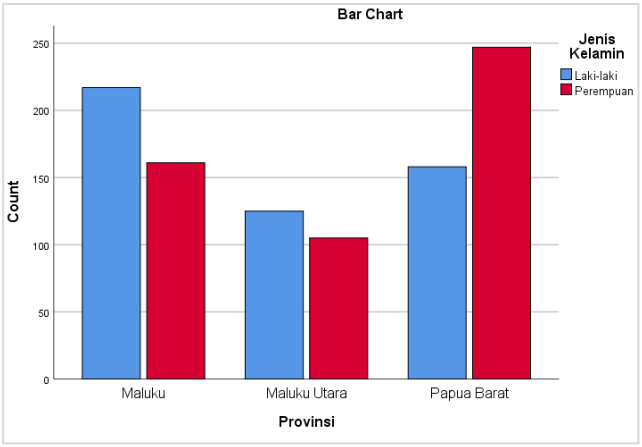

Figure 3. Number of HIV cases by gender and province

Figure 1 shows that the number of HIV cases for both men and women is quite high, but women dominate. When examining the age of individuals infected with the HIV virus, it appears that individuals aged between 25-49 years are more at risk of developing the disease. It might be thought that the reason for this is that individuals are in the productive age range, where individuals tend to engage in promiscuous sex, sharing needles, contagion during childbirth, and blood transfusions. This is more dominantly experienced by men.

Meanwhile, Figure 2 shows that in the same age range, namely the 25-49 year age range, the number of HIV cases is quite high for the provinces of Maluku, North Maluku and West Papua. However, there is a slight difference between the three provinces, with those aged 20-24 years having the lowest number of HIV cases in North Maluku. It can be concluded that the age range that is least affected is individuals in the 5-14 year age group. The reason for this can be predicted as the fact that there is a small possibility that these people do things that can cause infection with the HIV virus.

When viewed based on gender and province, West Papua is inversely proportional to Maluku and North Maluku. Women are more likely to be infected with the HIV virus than men. Apart from that, among the three provinces, the number of people infected with the HIV virus is dominated by women in West Papua.

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

Vol. 1, No. 2, 2023, 56-64
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Based on this HIV case, the hypothesis proposed is:
Ho: There is no relationship between gender and age
H1: There is a relationship between gender and age

Ho: There is no relationship between Province and age
H1: There is a relationship between Province and age

Ho: There is no relationship between gender and Province
H1: There is a relationship between gender and province

The relationship between variables will be analyzed using the chi-square test statistic. The following are the results of data processing using SPSS.

Table 2. Test statistics for analysis of relationships between variables

Variable	χ^2	df	p-value
Gender-Age	29.636	5	0.000
Province-Age	22.987	10	0.011
Gender-Province	29.431	2	0.000

The table above explains that with a value of χ^2= 29.636 and p-value = 0.000, statistically reject Ho because the p-value is less than the significance level α=0.05, which means that in the case of the number of HIV there is a relationship between gender and age. Then with 10 degrees of freedom, χ^2= 22.987 produces a p-value of 0.011 so reject Ho, meaning there is a relationship between Province and age. Likewise, gender and Province have a χ^2 of 29.431 and p-value = 0.000, reject Ho, so there is a relationship between gender and Province. This fact shows a tendency for a relationship between gender, province and age in HIV cases. Contingency tables with three dimensions can be created into 7 different log-linear models.

Table 3. Log-linear model with main effects and interactions between variables

No	Model	db	G^2	p-Value
1	JK, P, U	27	88.372	0.000
2	JK, P, U, JK*P	25	58.752	0.000
3	JK, P, U, JK*U	22	58.300	0.000
4	JK*P, JK*U, P*U	10	9.165	0.517
5	JK, P, U, P*U	17	66.799	0.000
6	JK, P, U, JK*P, JK*U	20	28.680	0.094
7	JK, P, U, P*U, JK*P	15	37.179	0.001

* JK = Gender, P= Province, dan U= Age.

The hypotheses built in the log-linear model include:
Ho: Appropriate model
Ha: The model is not suitable

We can find out the complexity of the model by examining the model based on the Goodness-of-Fit test. The models (JK, P, U), (JK, P, U, JK*P), and (JK, P, U, JK*U) are not suitable for describing the data because they have a p-value that is smaller than the level real α=0.05, and the value of G^2 is very large, so it is not significant. This means that the alternative hypothesis is accepted which states that the model is not appropriate. Furthermore, for the model (JK*P, JK*U, P*U),

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

Vol. 1, No. 2, 2023, 56-64
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

each variable has a two-way relationship with the smallest G^2 value, namely 9.165 and p-value = 0.517. Based on the test criteria, it is clear that there is a tendency not to reject Ho, which is significant at the real level α=0.05 because it is smaller than the p-value of the model. So, the model is a suitable model to describe the data.

Meanwhile, the models (JK, P, U, P*U), (JK, P, U, P*U, JK*P) have very large G^2 values, namely 66,799 and 37,179. The hypothesis test value for both models is rejecting Ho because the p-value for both is less than α=0.05, so the model is declared inappropriate. This is different from the model (JK, P, U, JK*P, JK*U) with G^2= 28,680 and p-value = 0.094, which can be considered because it does not reject Ho, meaning the model is appropriate. However, due to consideration of the complexity of the model, this model was not chosen as the best model. Based on the Goodnes-of-Fit hypothesis test analysis, the model (JK*P, JK*U, P*U) is the best and most appropriate model for describing the data. Below is a table showing the goodness of fit results for the best model using multinomial.

Table 3. Goodness-of-Fit

	Value	df	Sig.
Likelihood Ratio	9.165	10	.517
Pearson Chi-Square	9.090	10	.524
a. Model: Multinomial			
b. Design: Constant + Gender * Province + Gender * Age + Province * Age			

The best model parameter estimation results are given in table 4.

Tabel 4. Parameter Estimates model (**JK*P, JK*U, P*U**)

Parameter	Estimate	Std. Error	z-value	p-value	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	2.124[a]					
[Gender = 1] * [Province = 1]	0.195	.436	.447	.655	-.659	1.049
[Gender = 1] * [Province = 2]	0.292	.429	.681	.496	-.549	1.134
[Gender = 1] * [Province = 3]	0.241	.314	.767	.443	-.374	.855
[Gender = 2] * [Province = 1]	-0.779	.369	-2.110	.035	-1.502	-.055
[Gender = 2] * [Province = 2]	-0.555	.359	-1.546	.122	-1.259	.149
[Gender = 2] * [Province = 3]	0[b]	.	.	.	.	.
[Gender = 1] * [Age = 1]	-1.166	.482	-2.421	.015	-2.110	-.222
[Gender = 1] * [Age = 2]	-1.562	.603	-2.592	.010	-2.743	-.381
[Gender = 1] * [Age = 3]	-.315	.389	-.810	.418	-1.079	.448
[Gender = 1] * [Age = 4]	.872	.304	2.866	.004	.276	1.469
[Gender = 1] * [Age = 5]	2.323	.277	8.383	.000	1.780	2.867
[Gender = 1] * [Age = 6]	0[b]	.	.	.	.	.
[Gender = 2] * [Age = 1]	-.386	.458	-.843	.399	-1.283	.511

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

Vol. 1, No. 2, 2023, 56-64
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

[Gender = 2] * [Age = 2]	-.797	.541	-1.473	.141	-1.857	.263
[Gender = 2] * [Age = 3]	1.064	.344	3.089	.002	.389	1.739
[Gender = 2] * [Age = 4]	2.028	.308	6.585	.000	1.424	2.632
[Gender = 2] * [Age = 5]	2.828	.295	9.589	.000	2.250	3.406
[Gender = 2] * [Age = 6]	0[b]	.	.	.	.	.
[Province = 1] * [Age = 1]	.444	.595	.746	.456	-.722	1.609
[Province = 1] * [Age = 2]	.035	.745	.048	.962	-1.424	1.495
[Province = 1] * [Age = 3]	-.435	.496	-.878	.380	-1.406	.536
[Province = 1] * [Age = 4]	.463	.390	1.188	.235	-.301	1.228
[Province = 1] * [Age = 5]	.410	.368	1.114	.265	-.311	1.130
[Province = 1] * [Age = 6]	0[b]	.	.	.	.	.
[Province = 2] * [Age = 1]	.798	.547	1.460	.144	-.273	1.869
[Province = 2] * [Age = 2]	-.408	.791	-.516	.606	-1.957	1.141
[Province = 2] * [Age = 3]	-.610	.486	-1.256	.209	-1.562	.342
[Province = 2] * [Age = 4]	-.297	.390	-.762	.446	-1.061	.467
[Province = 2] * [Age = 5]	-.356	.359	-.991	.322	-1.059	.348
[Province = 2] * [Age = 6]	0[b]	.	.	.	.	.
[Province = 3] * [Age = 1]	0[b]	.	.	.	.	.
[Province = 3] * [Age = 2]	0[b]	.	.	.	.	.
[Province = 3] * [Age = 3]	0[b]	.	.	.	.	.
[Province = 3] * [Age = 4]	0[b]	.	.	.	.	.
[Province = 3] * [Age = 5]	0[b]	.	.	.	.	.
[Province = 3] * [Age = 6]	0[b]	.	.	.	.	.

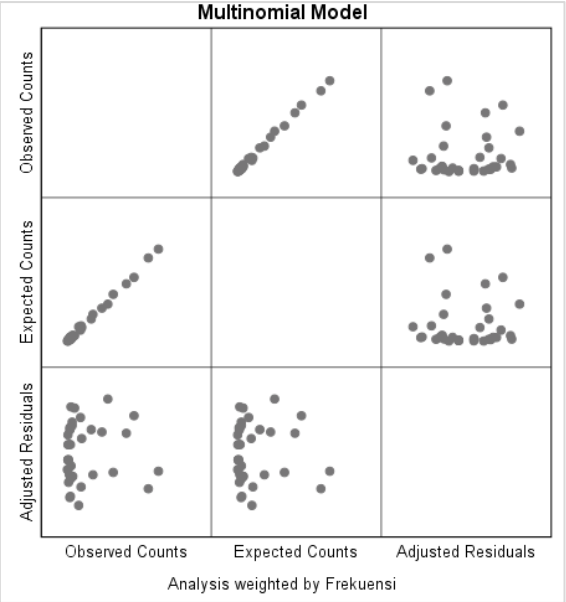

Figure 4. Distribution of the Multinomial Model (**JK*P, JK*U, P*U**)

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

Vol. 1, No. 2, 2023, 56-64
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

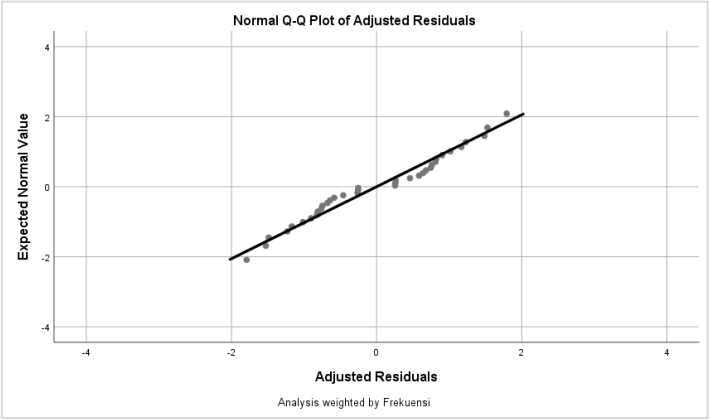

Figure 5. Distribution Residual Curve (**JK*P, JK*U, P*U**)

As can be seen in Figures 4 and 5, the distribution between the observed and expected data is very similar, meaning that there is a good match between the estimation results using the log-linear model and the HIV case data. This is reinforced by the residuals which spread normally around the estimator line, so that the model is very appropriate for analyzing this data.

It has been determined that the JK*P, JK*U, P*U models are the best models, so it can be concluded that the interactions between gender and province, gender and age, and province and age are statistically significant. The risk of being infected with the HIV virus in this study was shown to be highest between the ages of 25-49 years. If we monitor by gender, women are more at risk of being infected with the HIV virus than men with a total of 513 people from three provinces. Meanwhile, based on province, the one with the highest number of cases infected with the HIV virus is West Papua, namely 405 people.

CONCLUSION

In the case of HIV, 1013 patients infected with the HIV virus were grouped according to gender, age and province. The data used is categorical data, analyzed using a log-linear model. According to the results of the analysis and consideration of model complexity, (JK*P, JK*U, P*U) is the best model and fits the data because the p-value = 0.517 is greater than the real level $\alpha = 0.05$. This means that the interaction between gender, age and province is significant. Studies and explanations about the HIV virus show that individuals between the ages of 25-49 years are more at risk of being infected with the virus. Examined by gender group, women were most infected with the virus, namely 513 people. Apart from that, West Papua is the province with the highest number of HIV infections compared to Maluku and North Maluku.

REFERENCE

Abdallah, F. D. M. (2022). Log-linear Model for Describing the Relationship between Chromosomal Aberrations and Infertility Problems in Holstein-Friesian Cows. *Zagazig Veterinary Journal*, *50*(2), 177–185.

Agresti, A. (2003). *Categorical Data Analysis*. John Wiley & Sons, New Jersey.

Ali, F., Ali, S. A. S., Rahayu, S. B., Kamarudin, N. D., & Rahman, A. S. A. (2021). Investigation of interaction between age and gender effects of car users by using log-linear model: A bayesian inference approach. *Environment and Ecology Research*, *9*(4), 145–151.

Vol. 1, No. 2, 2023, 56-64
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

https://doi.org/10.13189/eer.2021.090401

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

Altun, G. (2021b). Analysis of the Multiraters Agreement with Log-Linear Models. *Bilge International Journal of Science and Technology Research*, *1*(1969), 2–5. https://doi.org/10.30516/bilgesci.950797

Alzahrani, S. M. (2022). A log linear Poisson autoregressive model to understand COVID-19 dynamics in Saudi Arabia. *Beni-Suef University Journal of Basic and Applied Sciences*, *11*(1), 1–6.

Asmare, A. A., & Agmas, Y. A. (2022). *Exploring the association of undernourishment indicators for under-five children in Sub-Saharan Africa: an application of log-linear model for three-way table*.

Carota, C., Filippone, M., & Polettini, S. (2022). Assessing Bayesian Semi-Parametric Log-Linear Models: An Application to Disclosure Risk Estimation. *International Statistical Review*, *90*(1), 165–183.

Du, M., Yuan, J., Jing, W., Liu, M., & Liu, J. (2022). The Effect of International Travel Arrivals on the New HIV Infections in 15–49 Years Aged Group Among 109 Countries or Territories From 2000 to 2018. *Frontiers in Public Health*, *10*(February), 1–9. https://doi.org/10.3389/fpubh.2022.833551

Fujisawa, K., & Tahata, K. (2022). Quasi Association Models for Square Contingency Tables with Ordinal Categories. *Symmetry*, *14*(4), 805.

Grover, G., & Sharma, A. (2018). Examining the effect of reduction of predictors affecting the survival time of HIV/ AIDS patients using a multiple correlation/ association technique. *Journal of Communicable Diseases*, *50*(3), 15–21. https://doi.org/10.24321/0019.5138.201815

Jamaludin, S. Z. M., Romli, N. A., Kasihmuddin, M. S. M., Baharum, A., Mansor, M. A., & Marsani, M. F. (2022). Novel logic mining incorporating log linear approach. *Journal of King Saud University-Computer and Information Sciences*.

Kementerian Kesehatan Republik Indonesia. (2021). *Profil Kesehatan Indonesia 2020*. https://doi.org/10.1524/itit.2006.48.1.6

Maryana. (2013). *Model Log Linier yang Terbaik untuk Analisis Data Kualitatif pada Tabel Kontingensi Tiga Arah*. *2*(2), 32–37.

Mulugeta, S. S., Muluneh, M. W., Belay, A. T., Moyehodie, Y. A., Agegn, S. B., Masresha, B. M., & Wassihun, S. G. (2022). Multilevel log linear model to estimate the risk factors associated with infant mortality in Ethiopia: further analysis of 2016 EDHS. *BMC Pregnancy and Childbirth*, *22*(1), 1–11.

Von Eye, A., Mun, E., & Mair, P. (2012). Log-linear modeling. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 218–223.

Wang, N., Massam, H., & Li, Q. (2022). Confidence intervals for discrete log-linear models when MLE doesn't exist. *Statistics & Probability Letters*, 109532.