GENDER AND AGE RELATIONSHIP ON COVID-19 IN INDONESIA BY GLM MODEL

A H Ulya [1], E S Nugraha [2]
School of Business, Study Program of Actuarial Science, President University, Cikarang, Bekasi
17550, Indonesia
[1]*aulia.ulya@student.president.ac.id*
[2] *edwin.nugraha@president.ac.id*

ABSTRACT

Coronaviruses are a family of viruses that can infect humans and animals. COVID-19 is caused by a newly identified coronavirus. COVID-19 was declared as a pandemic by WHO on March 11, 2020, since it is emerging in many countries around the world including Indonesia. Several forms and policies have been taken by the government to prevent the spread of COVID-19. By using the GLM model, we try to see the relationship between gender and age on COVID-19 cases and mortality that occur in Indonesia until April 12, 2021. The results show that the dependent variables have no relationship with gender, but do have a relationship with age. Inverse Gaussian with the link function 1/mu^2 is the best model for describing the relationship for the independent variable age. The results of the study serve as an evaluation for the government to obtain improve quality in healthcare quality, especially in developing strategies for dealing with COVID-19 cases, and mortality.

Keywords: *COVID-19, Statistical Modelling, Categorical Data Analysis, GLM.*

1. Introduction

Based on government data seen by the South China Morning Post (SCMP), the first case of COVID-19 caused by the Novel Coronavirus occurred on November 17, 2019, which infected a person from Hubei province. COVID-19 virus has spread rapidly around the world, started with people who exposed to the virus traveling and transmitting it while traveling until it was finally declared a pandemic by WHO on March 11, 2020. Government take several forms and policies in dealing with and preventing the increase in COVID-19 including implementing social distancing and large-scale social restrictions (LSSR).

COVID-19 has been discussed in many studies about the relations with gender and age. Gebhard et al. (2020), for example, examined the effects of gender lifestyle, behaviors, and psychological factors on COVID-19. By gathering the most recent epidemiological data cases in Germany, Spain, China, Switzerland, Italy, and France as of April. 1st. They found that females and males have different responses to infections, so it makes variations gender in disease. Klein et al. (2020) analyzed male-biased extreme COVID-19 outcomes, and how differences in gender can affect male-biased COVID-19 outcomes. From data per 100,000 people in NYC by biological factors that affect gender disparities in extreme COVID-19 outcomes. They found that no consideration has been given to consider how gender contribute in the increase of COVID-19 outcomes. Ahrenfeldt et al. (2020) investigate the extent of gender diversity in COVID-19 survivorship in different age groups and European area based on 10 regions. They discovered that gender diversity increased up to ages 60-69 years in most regions, but then decreased with the smallest gender diversity occurring at the age of above 80 years. Singh R and Adhikari R (2020) analyzed the role of age and social contract structures in determining the impact of applying social distancing among Chinese, Indians and Italians. The authors conclude that the three-week lockdown would be inadequate based on the finding of the SIR model. According to the model, prolonged periods of lockdown followed by intermittent relaxation would reduce the number of cases to the point where individualized social interaction tracing and quarantine may be possible. Davies (2020) discussed the age impact on COVID-19 epidemic transmission and control in 13 Chinese provinces, 12 Italian areas, Ontario, Canada, South Korea, Japan, and Singapore. The author’s model, together with preliminary evidence16, suggests that vulnerability to COVID-19 infection are age-dependent.

There are several reports in the studies that discuss the distribution of COVID-19 in terms of gender and age. The majority of these studies have data limitations that are not yet accessible in all countries since data on the distribution of COVID-19 by gender and age is typically not evident in public datasets. Indonesia is one of the countries that provide data on the distribution of COVID-19 cases according to age and gender that can be accessed by the public. Based on data from the 2020 Population Census, Kompas.com (2021), provide percentage of Indonesian population based on gender and age. In Indonesia, the male population is more than the female population, with a male ratio of 50.58 percent or around 136 million people, while women are 49.42 percent or around 133.54 million people. Indonesia has a higher proportion of people in the working age group (15-65 years) than people in the non-working age population, which is 70.72 percent. A large number of productive ages means that many Indonesians who have outdoor activities are at risk of contracting the virus. As a result, this paper will use the Generalized Linear Model (GLM) with the help of R software to investigate the relationship between gender and age in the case of COVID-19, especially in Indonesia. We hope that by understanding the factors that affect COVID-19 cases and death, the government and health institutions able to predict how best to deal with this pandemic in the future.

2. Methodology

The aim of the Generalized Linear Model (GLM) is to find the cause–effect relationship of the independent variable on the dependent variable. On GLM, the dependent variable is not normally distributed, but rather belongs to the exponential family of distributions, such as Binomial, Poisson, Binomial Negative, Normal or Gaussian, Gamma, and Inverse Gaussian. To encompass non-normal response distributions and possibly nonlinear functions of the mean, GLM extends standard linear regression models. GLM has three components, which is random component, linear predictor, and link function according to Agresti A (2015). Random component consist of a response variable y with independent observation $(\mathrm{y}_1,\ldots,\mathrm{y}_\mathrm{n})$ having probability density or mass function for a distribution in the exponential family. Linear predictor for a parameter vector $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_\mathrm{p})^\mathrm{T}$ and a $n\times p$ model matrix $\boldsymbol{X}$ that contains values of p explanatory variables for the n observations, the linear predictor is $\boldsymbol{X\beta}$. Link function connects the random component with the linear predictor. This is a function g applied to each component of E(y) that relates it to the linear predictor $\mathrm{g}\left[E(\mathrm{y})\right]=\boldsymbol{X\beta}$.

Table 1. Exponential family link function on R (Kabacoff R I, 2017)

Family	Default Link Function
Binomial	(link=”logit”)
Gaussian	(link=”identity”)
Gamma	(link=”inverse”)
Inverse Gaussian	(link=”1/mu^2”)
Poisson	(link=”log”)

The Generalized Linear Model will be used to assess the significant influencing factors for total cases and total mortality of COVID-19 in Indonesia. According to Zahro et al. (2018) when constructing a GLM model, there are several steps that must be done, as shown in Figure 1.

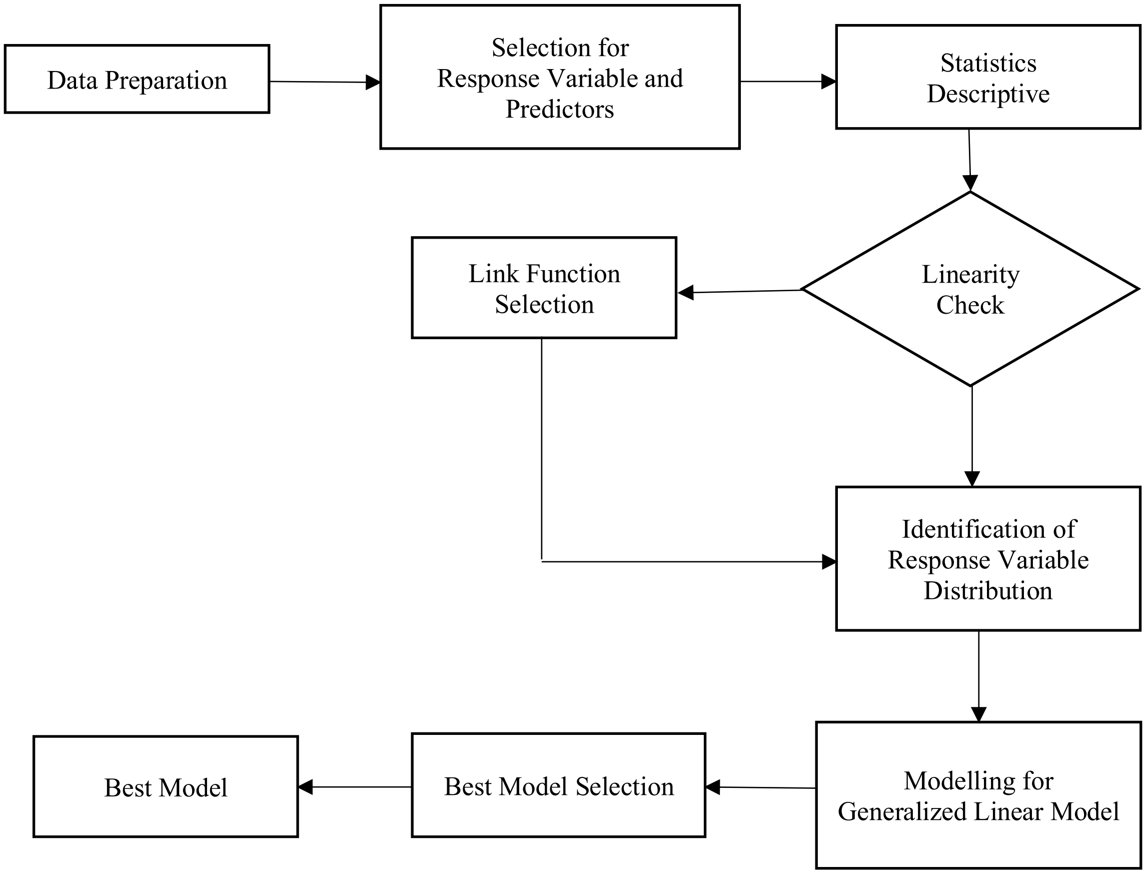

Data Preparation
Selection for Response Variable and Predictors
Statistics Descriptive
Linearity Check
Link Function Selection
Identification of Response Variable Distribution
Modelling for Generalized Linear Model
Best Model Selection
Best Model

Figure 1. Creating GLM model (Zahro J, Caraka R E, amd Herliansyah R, 2018)

2.1. Data Preparation

In data preparation, prepare the data that will be used to build the GLM model. The first step in data processing is to gather information from reliable sources. Then, to avoid missing data and other problems, do data profiling and cleansing. The final step is to structure, transform, and validate the data so that it will fit the format that we can use in data modelling.

2.2. Determine Dependent Variable and Independent Variable

The data gathered will be divided into two groups of variables, namely the independent variable and the dependent variable. The independent variable is a variable that describes or affects other variables, while the dependent variable is a variable that is explained or affected by the independent variable.

2.3. Statistics Descriptive

Descriptive statistics are used to present a data set that provides information or describes the data's characteristics. The descriptive statistics data will be outlined succinctly and easily and will contain key information from current data sets.

2.4. Linearity Check

A linearity test is used to assess if the dependent variable and the independent variable have a meaningful linear relationship, according to Permai S D (2018). The linearity test can be done using the R software by running the Terasvirta test. The data is not linear if the p-value is less than 0.05.

2.5. Identification of Response Variable Distribution and Link Function Selection

Identification of response variable distribution needs to be done by fitting distribution. Fitting distribution can be done using R software, by running the fitdist function. The best model can be seen from the smallest AIC value, or by looking at the distribution that has the fittest probability density function (pdf) to the histogram as mentioned by McNeese B. (2016). After getting the best fitting model we can choose the right link function based on Table 1 as summarized by Kabacoff R. I. (2017).

2.6. Modelling for Generalized Linear Model

Generalized linear model modeling will be done by including all independent and dependent variables. This can be done with the help of R software, using the GLM function.

2.7. Best Model Selection

The final stage in the generalized linear model is determining the best model, this can be seen from several criteria, namely Akaike's Information Criterion (AIC), null deviance, and residual deviance. According to Zahro et al (2018) and Lillis D (2017), the best models have the smallest AIC, null deviance and residual deviance values. The best model is the most appropriate model in explaining the relationship between gender and age with total cases and mortality.

3. Result and Discussion

3.1. Data Preparation

This paper use data from total COVID-19 cases and mortality cases that occurred in Indonesia until April 12, 2021, which can be found at covid19.go. The data used consisted of cases based on gender and age that occurred in 34 provinces in Indonesia. 2.5 percent of the overall cases lacked gender information, and 1.8 percent lacked age information. Before using the GLM model in R it is necessary to determine the dependent and independent variables. In this study, gender that consist of male and female also age that consist of child, teenager, young adult, middle aged adult, retired, and elderly are independent variables, while the number of cases and total mortality are the dependent variable. Table 2 shows a summary of COVID-19 cases and mortality by gender, while Table 3 shows a summary by age.

Table 2. Summary of COVID-19 cases and mortality by gender

Type	Cases	Death
Min.	1918	7.00
1st Qu.	4807	84.75
Median	7716	218.50
Mean	23119	624.99
3rd Qu.	18856	514.00
Max.	197039	5528.00

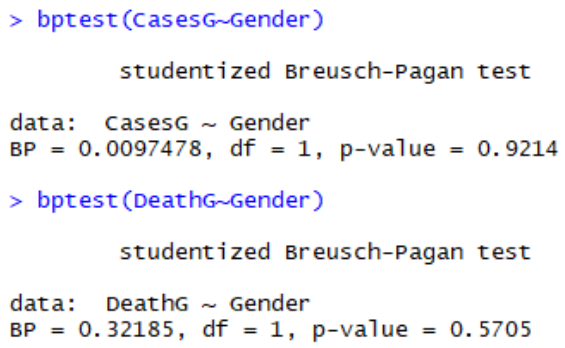
Table 3. Summary of COVID-19 cases and mortality by age

Type	Cases	Death
Min.	77	1
1st Qu.	913	8
Median	2511	36
Mean	7706	221.3
3rd Qu.	5770	127.2
Max.	117515	4622.0

3.2. Linearity Check

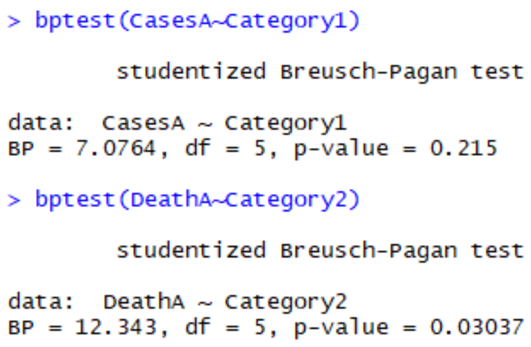
By using the bptest function on R, it can be seen that the relationship between dependent variables with gender is linear because the p-value is greater than 0.05. The relationship between total cases and age also linear, while the relationship between total mortality and age is non-linear since the p-value less than 0.05.

Figures 2 and 3 show the linearity test for gender and age, respectively.



```
> bptest(CasesG~Gender)

        studentized Breusch-Pagan test

data:  CasesG ~ Gender
BP = 0.0097478, df = 1, p-value = 0.9214

> bptest(DeathG~Gender)

        studentized Breusch-Pagan test

data:  DeathG ~ Gender
BP = 0.32185, df = 1, p-value = 0.5705
```

Figure 2. Linearity test gender



```
> bptest(CasesA~Category1)

        studentized Breusch-Pagan test

data:  CasesA ~ Category1
BP = 7.0764, df = 5, p-value = 0.215

> bptest(DeathA~Category2)

        studentized Breusch-Pagan test

data:  DeathA ~ Category2
BP = 12.343, df = 5, p-value = 0.03037
```

Figure 3. Linearity test age

3.3. Best Model Selection

When modeling GLM with the GLM function in R, the process of identifying the response variable distribution and selecting the link function can be performed concurrently. This study trying to find is there any relationship between independent variables with dependent variable by looking at p-value, AIC, null deviance, and residual deviance. If any, then find the most appropriate model to explain the relationship between the dependent variables and independent variables from four distributions, which is Gaussian or Normal, Gamma, Inverse Gaussian, and Poisson distribution. Table 4 to Table 7 shows the fitting distribution of each distribution. It can be shown that for independent variable gender there is no significant since almost no variable that has p-value > 0.05. Even though there is at least one variable have p-value < 0.05 in fitting distribution for death-gender by Poisson distribution, but the AIC, null deviance and residual deviance value is to large. So for independent variable gender it is means that there are no relationships with every dependent variable, and there is no need to continue the rest step. While for independent variable age there is a significant since every variable has p-value < 0.05. So, we need to find the most appropriate model for every dependent variable. For dependent variable cases the most appropriate model is by using inverse Gaussian distribution since it has the smallest value for AIC, null deviance, and residual deviance. While for dependent variable death, even though Gamma distribution has the smallest AIC value, but not with null deviance and residual deviance. So, the most appropriate model will be by using the inverse Gaussian distribution that has the smallest value for null deviance and residual deviance. To make it more clear, the QQ plot of the fit distribution for independent variable age in Figure 4 and Figure 5 has a large deviation of the residue from straight line. This implies that the model is non normal. Therefore, it can be concluded that the best model for independent variable age is inverse Gaussian with the link function 1/mu^2.

Table 4. Fitting distribution for cases - gender

Family	AIC	Null Deviance	Residual Deviance	At least one variable have P-Value < 0.05
Gaussian	1641.2	1.1054 e^11	1.1053 e^11	No
Gamma	1505.2	111.08	111.05	No
Inverse Gaussian	1463.6	0.007615	0.0076136	No
Poisson	2715744	2715751	2714987	No

Table 5. Fitting distribution for death - gender

Family	AIC	Null Deviance	Residual Deviance	At least one variable have P-Value < 0.05
Gaussian	1155.1	87354737	86863607	No
Gamma	1007.1	132.11	130.84	No
Inverse Gaussian	982.47	0.52021	0.51816	No
Poisson	79839	80126	79338	Yes

Table 6. Fitting distribution for cases - age

Family	AIC	Null Deviance	Residual Deviance	At least one variable have P-Value < 0.05
Gaussian	4527.8	5.2518 e^10	4.8591 e^10	Yes
Gamma	3955.8	457.78	356.20	Yes
Inverse Gaussian	3937.7	0.19323	0.17049	Yes
Poisson	2740741	3318346	2738757	Yes

Table 7. Fitting distribution for death-age

Family	AIC	Null Deviance	Residual Deviance	At least one variable have P-Value < 0.05
Gaussian	3021.1	81910746	71234908	Yes
Gamma	2068.9	715.72	325.63	Yes
Inverse Gaussian	2134.7	23.741	17.429	Yes
Poisson	82741	132855	81693	Yes

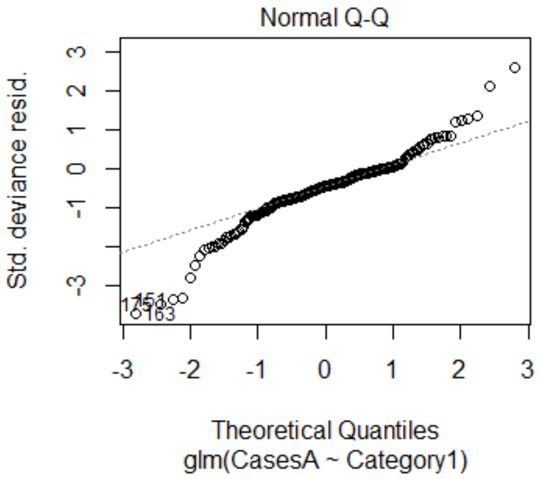


Figure 4. QQ plot inverse Gaussian (Cases-Age)

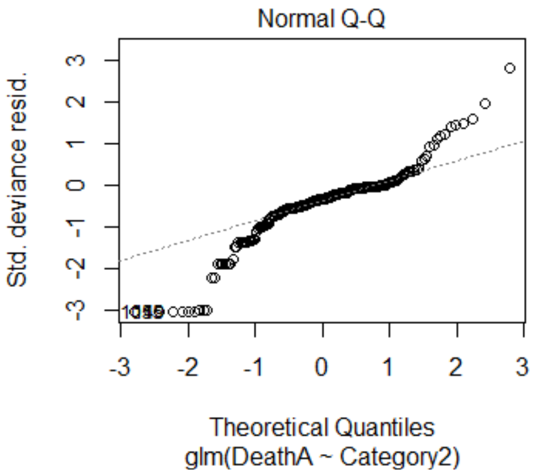


Figure 5. QQ plot Inverse Gaussian (Death-Age)

Table 8 and Table 9 shows summary models for independents variable Age by using Inverse Gaussian distribution. Parameter that has a significant with dependent variable cases are $\beta_2, \beta_3, \beta_4, \beta_5$ and β_6 because it has p-value < 0.05. While parameter that has a significant with dependent variable death are $\beta_3, \beta_4, \beta_5$ and β_6.

Table 8. Inverse Gaussian Cases-Age

Parameter	Estimate	Stand. Error	T-Value	P-Value
β_0	0.0000005837	0.0000002129	2.742	0.00666
β_1	0			
β_2	-0.0000005308	0.0000002157	-2.461	0.01473
β_3	-0.0000005761	0.000000213	-2.705	0.00743
β_4	0.0000005783	0.000000213	-2.716	0.00720
β_5	0.0000005744	0.0000002131	-2.696	0.00762
β_6	0.0000005449	0.0000002147	-2.538	0.01192

Table 9. Inverse Gaussian Death-Age

Parameter	Estimate	Stand. Error	T-Value	P-Value
β_0	0.011664	0.004140	2.818	0.00536
β_1	0			
β_2	-0.001595	0.005471	-0.291	0.77102
β_3	-0.010822	0.004171	-2.594	0.01023
β_4	-0.011616	0.004140	-2.806	0.00556
β_5	-0.011659	0.004140	-2.816	0.00538
β_6	-0.011661	0.004140	-2.817	0.00537

After defining the parameter that has a significant with dependent variable, the GLM fit model by inverse Gaussian can be created.

- Cases – Age

$$\begin{aligned}\hat{\mu} = \exp(&0.0000005837 - 0.0000005308\chi_2 - 0.0000005761\chi_3 \\ &+ 0.0000005783\chi_4 + 0.0000005744\chi_5 + 0.0000005449\chi_6)\end{aligned} \qquad (6)$$

The interpretation for each variable based on this model can be shown from Table 10. Table 10 shows that for every 1 percent addition of the significant variable χ_i will be multiply the average dependent variable cases equal to $\exp(\beta_i)$.Interpretation acording to Nuraeni (2018), in other words, the increasing 1 percent ratio of teenager and young adult will be same as the decrease of average COVID-19 cases in Indonesia around 1. While the increasing 1 percent ratio of middle aged adult, retired, and elderly will be same as the increase of average COVID-19 cases in Indonesia around 1.

Table 10. Interpretation variable to cases average value

	β_2	β_3	β_4	β_5	β_6
$\exp(\beta_i)$	1.00000053	1.0000005761	1.0000005783	1.000000574	1.0000005449

- Death – Age

$$\hat{\mu} = \exp(0.011664 - 0.010822\chi_3 - 0.011616\chi_4 - 0.011659\chi_5 - 0.011661\chi_6) \qquad (7)$$

The interpretation for each variable based on this model can be shown from Table 11. The increasing 1 percent ratio of young adult, middle aged adult, retired, and elderly will be same as the decrease of average COVID-19 cases in Indonesia around 1.01088, 1.01168, 1.01173, and 1.011729 respectively.

Table 11. Interpretation variable to death average value

	β_3	β_4	β_5	β_6
$\exp(\beta_i)$	1.01088	1.01168	1.01173	1.011729

4. Conclusion and Implications

Based on the results of the analysis and evaluation, it can be concluded that there is no relationship between independent variable gender with dependent variables. While there is a relationship between independent variable age with dependent variables. The best model that describe the relationship for independent variable age is inverse Gaussian with the link function 1/mu^2. Parameter that has a significant with dependent variable cases are $\beta_2, \beta_3, \beta_4, \beta_5$ and β_6. While parameter that has a significant with dependent variable death are $\beta_3, \beta_4, \beta_5$ and β_6. The results of the research serve as an evaluation for the government to obtain improve quality in the health sector, especially in developing strategies for dealing with COVID-19 cases, and mortality. Since there is a relationship between the independent variables with dependent variables age, this means that government needs to complete data variables age to help other analysts in developing preventive and other strategies for dealing with COVID-19 or learning about potential events. We suggest to other analysts that maybe want to make study that still related to this case such as the relationship of COVID-19 cases with symptom, and congenital disease.

References

Ma, J. (2020, March 13). Coronavirus: China's First Confirmed Covid-19 Case Traced Back To November 17. *South China Morning Post*. https://www.scmp.com/news/china/society/article/3074991/coronavirus-chinas-first-confirmed-covid-19-case-traced-back

Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H. K., Morgan, R., and Klein, S. L. (2020). Impact of sex and gender on COVID-19 outcomes in Europe. *Biology of Sex Differences*, 11, 1-13. https://doi.org/10.1186/s13293-020-00304-9

Klein, S. L., Dhakal, S., Ursin, R. L., Deshpande, S., Sandberg, K., & Mauvais-Jarvis, F. (2020). Biological sex impacts COVID-19 outcomes. *PLoS pathogens*, 16(6), 1-5.

https://doi.org/10.1371/journal.ppat.1008570

Ahrenfeldt, L. J., Otavova, M., Christensen, K., and Lindahl-Jacobsen, R. (2020). Sex and age differences in COVID-19 mortality in Europe. *Wiener kinische Wochenschrift*, 1-6. https://doi.org/10.1007/s00508-020-01793-9

Singh, R., and Adhikari, R. (2020). Age-structured impact of social distancing on the COVID-19 epidemic in India. *arXiv preprint arXiv:2003.12055.* Retrived April 3, 2020, from https://arxiv.org/abs/2003.12055

Davies, N. G., Klepac, P., Liu, Y., Prem, K., Jit, M., and Eggo, R. M. (2020). Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nature medicine*, 26(8), 1205-1211. Retrived April 3, 2020, from https://www.nature.com/articles/s41591-020-0962-9?fbclid=IwAR0DM-WoHX5tcEp5WU_UzKW8BwFjcXWXa7aGsFdzzmGeCt1ir3LpM8lt1r4

Idris, M. (2021, January 26). Ini Jumlah Penduduk Indonesia yang Lahir Sebelum Kemerdekaan 1945. *Kompas.com*. https://money.kompas.com/read/2021/01/26/153811526/ini-jumlah-penduduk-indonesia-yang-lahir-sebelum-kemerdekaan-1945?page=all

Agresti, A. (2015). *Foundations of linear and generalized linear models.* John Wiley and Sons. https://books.google.co.id/books?id=dgIzBgAAQBAJ&lpg=PR11&ots=70wtz4JCRd&dq=Foundations%20of%20Linear%20and%20Generalized%20Linear%20Models%20Wiley%20Series%20in%20Probability%20and%20Statistics&lr&hl=id&pg=PR1#v=onepage&q=Foundations%20of%20Linear%20and%20Generalized%20Linear%20Models%20Wiley%20Series%20in%20Probability%20and%20Statistics&f=false

Kabacoff, R. I. (2017). *Generalized linear models*. Quick-R. https://www.statmethods.net/advstats/glm.html

Zahro, J., Caraka, R. E., and Herliansyah, R. (2018). *Aplikasi generalized linear model pada R* [eBook edition]. Innosain. https://myjobstreet-id.jobstreet.co.id/application/application-status.php?view=1&x=99ts7bjodljth2qp8epb07osp1

Permai, S. D. (2018, December 8). *Uji Linieritas Menggunakan R*. BINUS university school of computer science. https://socs.binus.ac.id/2018/12/08/uji-linieritas-menggunakan-r/
McNeese, B. (2016, October). *Deciding which distribution fits your data best*. SPC for excel. https://www.spcforexcel.com/knowledge/basic-statistics/deciding-which-distribution-fits-your-data-best

Lillis, D. (2017). *Generalized Linear Models in R, Part 2: Understanding Model Fit in Logistic Regression Output*. The Analys Factor. https://www.theanalysisfactor.com/r-glm-model-fit/

KPCPEN. (2021). *Peta Sebaran COVID-19* [Infographic]. Covid19.go.id. https://covid19.go.id/peta-sebaran-covid19

Nuraeni (2018). *Pemodelan Jumlah Kematian Bayi di Provinsi Sulawesi Selatan Menggunakan Regresi Poisson Inverse Gausian*. [Universitas Negeri Makassar]. https://core.ac.uk/download/pdf/154762921.pdf