



Logistic regression model for identifying factors affecting hospitalization of children with pneumonia

Anwar Fitrianto^{1*}, Wan Zuki Azman Wan Muhamad²

¹ Department of Statistics, Institut Pertanian Bogor, Indonesia.

² Institute of Engineering Mathematics, Universiti Malaysia Perlis, Malaysia.

✉ anwarstat@gmail.com

Article Information

Submitted Dec 13, 2021

Revised April 10, 2022

Accepted July 29, 2022

Keywords

Pneumonia;

Logistic;

Regression;

Maximum Likelihood Estimation.

Abstract

Pneumonia is a lung infection that could happen in babies, children, adults and older people. However, pneumonia in infants and older adults is more serious. Several studies found that infants are more likely to get pneumonia if they live in low-income families. The study aimed to identify factors that cause children to be hospitalized for pneumonia. The binary logistic regression analysis was performed to build a full model regardless of the significance of the variables. The forward selection approach was used to select the significant variables. It was found that the age of the mother, cigarette smoked by the mother during pregnancy, duration (in months) of the children on solid food, and the age when the child had pneumonia with the p-value of 0.0009, 0.0010, 0.0003 and less than 0.0001, respectively. The odds ratio of mother's age, cigarette smoked by mother during pregnancy, how many months the child on solid food, and children's age when they had pneumonia are 0.69, 6.22, 0.40 and 0.60, respectively.

INTRODUCTION

Pneumonia is a lung infection that could happen in babies, children, adults, and older people. However, pneumonia in infants and older adult is more serious. Lamberti et al. researched whether breastfeeding protects infants against pneumonia and if the protection changes with age. The results showed that the babies who were not breastfed were 17 times more likely to get pneumonia than those who were breastfed (Lamberti et al., 2013). Pneumonia can be caused by germs, such as viruses, bacteria, fungi, and parasites. A person might get pneumonia if he breathes in the infected air or the bacteria into the lung. Besides, a person is more likely to get pneumonia after having flu or if the person has asthma, heart disease, chickenpox, or cancer. Furthermore, smokers have a higher chance of getting pneumonia.

The primary treatment of pneumonia is antibiotics. The physician would choose the type and amount of antibiotics based on the patient's age and symptoms. Anyone could prevent or reduce the chances of getting pneumonia by stopping smoking, avoiding skin contact with people with infections, staying away from people with flu or other respiratory infection, and washing their hands frequently.

Studies found that pneumonia is common in infants, such as in Kusahara et al. (2014), Govender et al. (2021), Shakeel et al. (2021), Omer et al. (2018), Yuan et al. (2016), and Enarson et al. (2014). Infants' immune systems are still developing, so they are susceptible to infectious diseases, including pneumonia. Many studies have found that socioeconomic factors contribute greatly to this respiratory disease. In low-income families, the allocation of funds to maintain personal hygiene and the surrounding environment is low, resulting in

several diseases, such as pneumonia. Walker et al., (2013) investigated the relationship between poverty and pneumonia and found that the incidence rates were significantly higher in low-income areas than in high-income areas. She concluded that infants are more likely to get pneumonia if the infants live in a household with low income. Meanwhile, Homaira et al., (2012) investigated the risk factors for pneumonia in children among the urban poor in Dakka, Bangladesh, from 2009 through 2011. Lazzerini et al., (2016) investigated the risk factors for pneumonia for less than two years old Malawian children.

According to Walker et al., (2013), pneumonia is a leading infectious cause of death in children and contributes to a high percentage of all deaths of children under five years old. Children younger than five years of age involve many pneumonia hospitalizations. Researchers have investigated other possible risk factors that cause pneumonia in young children, except for the above factors. This study aimed to identify the factors that lead to the hospitalization of children due to pneumonia by developing a logistic regression model.

METHODS

Data

The data set involved in this study is a dataset of children (not more than one year old) with pneumonia. The dataset consisted of 3470 observations with one dependent and 14 independent variables (Beasley et al., 2016). The detail of the dependent and independent variables is displayed in Table 1. Since the data was not balanced, random sample selection was conducted randomly to select 50 out of 73 observations on the event $Y = 1$ and 100 out of 3397 observations on $Y = 0$ to reduce biases.

Table 1. The Description of the Studied Variables

Variable	Description	Code Values
Y	Indicator for hospitalization for pneumonia	0=No, 1= Yes
x_1	Mother's age	Years
x_2	Mother's urban environment	0= No, 1=Yes
x_3	Mother's alcohol consumption during pregnancy	0= No, 1=Yes
x_4	Cigarette smoked by mother during pregnancy	0= No, 1=Yes
x_5	Region of the country (Northeast)	0= No, 1=Yes
x_6	Region of the country (North Central)	0= No, 1=Yes
x_7	Region of the country (South)	0= No, 1=Yes
x_8	Mother's poverty level	0= No, 1=Yes
x_9	Normal birth weight (>5.5 lbs.)	0= No, 1=Yes
x_{10}	Mother's education	Years of school
x_{11}	Number of the child's siblings	Number
x_{12}	Month the child was weaned	Months
x_{13}	Month the child on solid food	Months
x_{14}	Child's age when contracted pneumonia	Months

Binary Logistic Regression Models

Binary logistic regression was used to determine the relationship between a dichotomous dependent variable and a set of independent variables, which can be continuous or discrete (Hosmer Jr et al., 2013). It was a non-linear regression model; thus, it did not need to satisfy any assumptions of normality, linearity, homogeneity of variance for the independent variables and constant variance of residuals (Sarkar & Midi, 2010).

A logistic regression model is formulated as follows

$$P(Y = 1 | x) = \pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}. \quad (1)$$

Since the independent variable is not linear, a logit transformation is performed, and the function is defined as follows

$$g(x) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n. \quad (2)$$

The $g(x)$ was a linear function of independent variables; therefore, it was much easier to fit. A link function was needed to transform the probabilities of a response variable from the closed interval (0,1) to an unbounded continuous (Boateng & Abaye, 2019; Chen et al., 2017). Once the transformation had been completed, the relationship between the predictor and response variable could be modelled with linear regression. There are three link functions in binary logistic regression: logit, probit, and compit. The logit function is an odds ratio for a given probability value. Logit transformation was performed by adding log or natural log to the odds, and the link function was defined as follows (Güriş et al., 2011; Stoltzfus, 2011):

$$\logit(p_i) = \ln \frac{p_i}{1 - p_i}. \quad (3)$$

where p_i is the probabilities associated with each response variable.

After the logit transformation, the logistic function is close to the normal integrated curve. The other common link function in binary logistic regression is the probit function (Boateng & Abaye, 2019; Razzaghi, 2013) which is defined as:

$$\text{probit}(p_i) = \varphi^{-1}(p_i). \quad (4)$$

Where $\varphi^{-1}(x)$ is the inverse cumulative distribution function of the normal distribution and p_i is the probabilities associated with each response variable.

Furthermore, the compit function is also known as the complementary log-log function (Hilbe, 2011), which is an asymmetric function with the following expression:

$$\text{compit}(p_i) = \log[-\log(1 - p_i)]. \quad (5)$$

Maximum Likelihood Estimation

The method to estimate the parameters of binary logistic regression is maximum likelihood estimation (MLE). A likelihood function is constructed to perform this method. According to Diop et al. (2011), a simple way to obtain the expression for the likelihood function is as follows:

$$\pi(x_i)^{y_i} [1 - \pi(x_i)]^{1 - y_i}. \quad (6)$$

Since the observations are assumed to be independent, the likelihood function is defined as follows:

$$l(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}. \quad (7)$$

However, Hosmer Jr et al., (2013) claim that this equation is not practical. Hence, a log-likelihood equation is then constructed as follows:

$$L(\beta) = \ln[l(\beta)] = \sum \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}. \quad (8)$$

Next, the $L(\beta)$ is derived with respect to β_0 and β_1 and set equal to zero to estimate the coefficient of parameter, β_i , that maximizes $L(\beta)$. The Newton-Raphson method is used to obtain the derivation (Hilbe, 2011).

Likelihood Ratio and Wald Test

A likelihood ratio test is needed to get information about the overall fit of all coefficients in the binary logistic regression model (Fagerland & Hosmer, 2017). The likelihood ratio test can be expressed as follow:

$$G = [(-2 \ln L_0) - (-2 \ln L_1)], \quad (9)$$

where L_0 is the likelihood of the null model and L_1 is the likelihood of the alternate model.

We can also express it as follows:

$$D = -2 \sum_{i=1}^n \left[y_i \ln \left(\frac{\pi_i}{y_i} \right) \ln \left(\frac{1-\pi_i}{1-y_i} \right) \right]. \quad (10)$$

The G follows a chi-square distribution with the degree of freedom equal to the number of predictors added to the model. The null hypothesis is that all the estimated coefficients in the logistic regression model equal zeros. The decision is based on the value of the likelihood ratio test, G , or the corresponding p -value. An alternative way to test the significance of the coefficient in the model is by using the Wald test. The test is similar to the t -test in linear regression, but it does not have a t -distribution and behaves asymptotically normally distributed. Wald statistic can be written as follows:

$$W = \frac{(\hat{\beta}_i)}{[S.E(\hat{\beta}_i)]}. \quad (11)$$

The null hypothesis is rejected if the $|W| > z_{\alpha/2}$, where W is the value of the Wald statistic.

Hosmer-Lemeshow Goodness-of-Fit-Tests

Hosmer-Lemeshow goodness-of-fit-tests is a statistical method to test if the model performs adequately. It evaluates the goodness-of-fit by generating ordered groups of subjects and then

compares the observed value to the predicted value in each group (Fagerland & Hosmer, 2017). The Hosmer-Lemeshow test statistic can be illustrated as:

$$\hat{C} = \sum_{k=1}^g \frac{(O_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}. \quad (12)$$

Based on the formula above, n'_k is the total number of subjects in the k^{th} group, c_k denotes the number of covariate patterns in the k^{th} decile, O_k is the number of responses among c_k covariate patterns, and $\bar{\pi}_k$ is the average estimated probability.

A Chi-square distribution $\hat{C}$ approximates the distribution of C. The high p-value of the Hosmer-Lemeshow test indicates no significant difference between the observed and the predicted value of the response. In other words, the model fits well.

RESULTS AND DISCUSSION

Model Building for the Pneumonia Disease

Out of 150 randomly selected observations in this study, 50 (33.33%) of the children contracted pneumonia, whereas 100 (66.66%) were not. The estimated coefficients of the logistic regression and the standard error of coefficients are displayed in Table 2. The variables x_1 , x_4 and x_{14} have p -values less than 0.05; thus, the three variables were significant. However, the variables x_2 , x_3 , x_5 , x_6 , x_7 , x_8 , x_9 , x_{10} , x_{11} , x_{12} and x_{13} were not as they have high p -values.

Regardless of the significance of each variable, x_3 , x_4 , x_5 , x_6 , x_7 , x_8 , x_9 and x_{11} had positive coefficients, which implied that $P(Y = 1)$ increase when the variables are at a higher value and vice versa.

Table 2. Logistic Regression Analysis for Full Model

Predictor	Coefficient	SE Coefficient	Wald Stat	p-value	Odds Ratio
Intercept	9.3889	3.2693	2.87	0.004	
x_1	-0.4142	0.1452	-2.85	0.004	0.66
x_2	-0.3302	0.6916	-0.48	0.633	0.72
x_3	0.3293	0.7220	0.46	0.648	1.39
x_4	1.4675	0.6850	2.14	0.032	4.34
x_5	0.4640	1.2722	0.36	0.715	1.59
x_6	0.1455	1.1630	0.13	0.900	1.16
x_7	0.4438	1.1456	0.39	0.698	1.56
x_8	1.3039	1.4325	0.91	0.363	3.68
x_9	0.4039	0.6303	0.64	0.522	1.50
x_{10}	-0.0367	0.1809	-0.20	0.839	0.96
x_{11}	0.3342	0.3792	0.88	0.378	1.40
x_{12}	-0.4244	0.7840	-0.54	0.588	0.65
x_{13}	-0.1572	0.9786	-0.16	0.872	0.85
x_{14}	-0.4945	0.0887	-5.57	0.000	0.61

Since many variables existed in the full model, a variable selection procedure was performed to obtain the best-reduced model. The backward elimination procedure using SAS version 9.4 was conducted for this purpose. The procedures were conducted step-by-step by removing insignificant variable one at a time, starting from the insignificant variable with the

largest p-value. The procedure is completed once all independent variables have small p-values (less than 0.05). The results of the reduced model are displayed in Table 3.

Table 3. Reduced Model of the Binary Logistic Regression Analysis

Predictor	Coefficient	SE Coefficient	Wald Stat	p-value	Odds Ratio
Constant	9.8269	2.5689	3.82	0.000	
x_1	-0.3683	0.1118	-3.30	0.001	0.69
x_4	1.8272	0.5907	-3.09	0.002	6.22
x_{13}	-0.9218	0.3479	-2.65	0.008	0.40
x_{14}	-0.5046	0.0829	-6.09	0.000	0.60
Log-Likelihood= - 44.213					
Test that all slopes are zero: G= 102.528 DF= 4 p-value=0.000					

Table 3 shows that the variable x_4 has a positive coefficient which implies that $P(Y = 1)$ increases when x_4 is at a higher value, and vice versa. On the other hand, the variable x_1 has a negative coefficient, which implies that $P(Y = 1)$ decreases when x_1 , x_{13} and x_{14} are at higher values. Moreover, the p-values of variables x_1 , x_4 , x_{13} and x_{14} are less than the 0.05 value of alpha. This means that the variables x_1 , x_4 , x_{13} and x_{14} significantly influence the children's pneumonia hospitalization. The fitted final reduced model for the pneumonia disease was

$$P(Y = 1 | x) = \frac{e^{9.82686 - 0.368324x_1 + 1.82715x_4 - 0.921785x_{13} - 0.504557x_{14}}}{1 + e^{9.82686 - 0.368324x_1 + 1.82715x_4 - 0.921785x_{13} - 0.504557x_{14}}}, \quad (13)$$

or

$$\text{Logit}(\hat{Y}) = 9.82686 - 0.36832x_1 + 1.82715x_4 - 0.921785x_{13} - 0.504577x_{14} \quad (14)$$

Variable x_1 , x_4 , x_{13} and x_{14} are the age of the mother, cigarette smoked by the mother during pregnancy, duration (in months) of the children on solid food, and then age when the child had pneumonia. According to Abusaad & Hashem (2014), more than half of the mothers with children who suffer from pneumonia are between 20 to 29 years old. Most young mothers did not have good knowledge about pneumonia disease; hence, they were unaware of the symptoms of pneumonia. Next, the children of smoking mothers are more likely to get pneumonia. Babies younger than six months old are not ready for solid foods (Porter et al., 2009). They usually could not hold their heads up and sit up by themselves. Besides, the mouth muscles that help guide the food into the throat and stomach are not fully developed yet. Thus, the food can drop down into the lungs, not the stomach, when they try to swallow the solid food, leading to pneumonia.

The log-likelihood estimate was conducted to check the overall fit of the reduced model. The G statistic was used to test the null hypothesis that all slopes were zero versus the alternate hypothesis that at least one of the slopes or coefficients was different from zero. The value of the G statistic equals 102.538 (p-value=0.000), indicating that there was sufficient evidence that at least one of the coefficients differed from zero. Thus, the model performed adequately. Furthermore, the values of $|W|$ of variable x_1 , x_4 , x_{13} and x_{14} were larger than

$z_{\alpha/2}=1.96$. Thus, there was sufficient evidence that the coefficients in the model were not equal to zero.

Odds Ratios and Goodness of Fit Test

The odds ratio is displayed in column 6 of Table 3. The odds ratio of the variable x_1 was 0.69, which indicated that a mother one year older was 31% less likely to hospitalize her child for pneumonia. On the other hand, the variable x_4 had an odds ratio of 6.16, which indicated that the event of children hospitalized for pneumonia was 6.16 times more likely to occur if the mother smoked during pregnancy. In other words, children whose mothers smoked during pregnancy tended to have a greater probability of getting hospitalized due to pneumonia than children whose mothers did not smoke during pregnancy. For the variable x_{13} , the odd ratio was 0.40, which indicated that children hospitalized for pneumonia were 60% less likely to occur with one unit increase in the month the child was weaned.

Table 4 displays the results of the Pearson, Deviance, and Hosmer-Lemeshow tests. Focussing on the Hosmer-Lemeshow test with the Chi-square value of 6.180 (p-value=0.627) indicated insufficient evidence to say that the model did not fit the data. Table 4 shows the observed and expected frequencies, showing how well the model fits the data by comparing the observed and expected frequencies for $Y = 1$ and $Y = 0$.

Table 4. Goodness-of-Fit Tests

Method	Chi-square	DF	p-value
Pearson	107.121	89	0.093
Deviance	82.314	89	0.679
Hosmer-Lemeshow	6.180	8	0.627

CONCLUSIONS

In this study, The researchers used binary logistic regression to determine the risk factors that cause hospitalization of pneumonia among children. Four significant variables were determined, which were x_1 (age of the mother), x_4 (cigarette smoked by mother during pregnancy), x_{13} (month of the children on solid food) and x_{14} (the age of the child when had pneumonia) with the p-values of 0.0003, 0.0009, 0.0010 and <0.0001 respectively. The variables selection was performed by using forward selection. The obtained final reduced model is as follows:

$$P(Y = 1 | x) = \frac{e^{9.82686 - 0.368324x_1 + 1.82715x_4 - 0.921785x_{13} - 0.504557x_{14}}}{1 + e^{9.82686 - 0.368324x_1 + 1.82715x_4 - 0.921785x_{13} - 0.504557x_{14}}}$$

The odds ratio of x_1 (age of the mother), x_4 (cigarette use by mother during pregnancy), x_{13} (duration (month) of the children on solid food) and x_{14} (the age when the child contracted pneumonia) were 0.69, 6.16, 0.40, and 0.60, respectively. The variable x_4 had the highest positive odds ratio of 6.16, which implied that the event was 6.16 times more likely to occur if the mother smoked during pregnancy. The odds ratios of variables x_1 , x_{13} and x_{14} were less than 1. It implied that the event was less likely to occur when the variables were at a higher value.

AUTHORSHIP CONTRIBUTION STATEMENT

AF contributed to generating research ideas, data acquisition, data analysis and computational statistics, interpretation of results, manuscript drafting, statistical analysis, admin, and technical support. WZAWM contributed to conceptualizing and designing the study, interpreting results, drafting the initial manuscript, revising the manuscript, finalizing the manuscript, and securing funding. WZAWM also helped supervise the project, providing critical feedback and shaping the research, analysis, and manuscript.

REFERENCES

- Abusaad, F. E., & Hashem, F. S. (2014). [Mothers learning needs assessment regarding pneumonia among children less than five years at Saudi Arabia](#). *Int Res J*, 3, 85–93.
- Beasley, W., Rodgers, J., Bard, D., Hunter, M., Mason, S., & Beasley, M. W. (2016). *Package 'NlsyLinks.'*
- Boateng, E. Y., & Abaye, D. A. (2019). [A review of the logistic regression model with emphasis on medical research](#). *Journal of Data Analysis and Information Processing*, 7(4), 190–207.
- Chen, K., Cheng, Y., Berkout, O., & Lindhiem, O. (2017). [Analyzing proportion scores as outcomes for prevention trials: a statistical primer](#). *Prevention Science*, 18(3), 312–321.
- Diop, A., Diop, A., & Dupuy, J.-F. (2011). [Maximum likelihood estimation in the logistic regression model with a cure fraction](#). *Electronic Journal of Statistics*, 5, 460–483.
- Enarson, P. M., Gie, R. P., Mwansambo, C. C., Maganga, E. R., Lombard, C. J., Enarson, D. A., & Graham, S. M. (2014). [Reducing deaths from severe pneumonia in children in Malawi by improving delivery of pneumonia case management](#). *PloS One*, 9(7), e102955.
- Fagerland, M. W., & Hosmer, D. W. (2017). [How to test for goodness of fit in ordinal logistic regression models](#). *The Stata Journal*, 17(3), 668–686.
- Govender, K., Msomi, N., Moodley, P., & Parboosing, R. (2021). [Cytomegalovirus pneumonia of infants in Africa: A narrative literature review](#). *Future Microbiology*, 16(18), 1401–1414.
- Güriş, S., Çağlayan, E., & Ün, T. (2011). [Estimating of probability of homeownership in rural and urban areas: Logit, probit and gompit model](#). *European Journal of Social Sciences*, 21(3), 405–411.
- Hilbe, J. M. (2011). Logistic regression. *International Encyclopedia of Statistical Science*, 1, 15–32.
- Homaira, N., Luby, S. P., Petri, W. A., Vainionpaa, R., Rahman, M., Hossain, K., Snider, C. B., Rahman, M., Alamgir, A. S. M., & Zesmin, F. (2012). [Incidence of respiratory virus-associated pneumonia in urban poor young children of Dhaka, Bangladesh, 2009–2011](#). *PloS One*, 7(2), e32056.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.
- Kusahara, D. M., Enz, C. da C., Avelar, A. F. M., Peterlini, M. A. S., & Pedreira, M. da L. G. (2014). [Risk factors for ventilator-associated pneumonia in infants and children: a cross-sectional cohort study](#). *American Journal of Critical Care*, 23(6), 469–476.

- Lamberti, L. M., Zakarija-Grković, I., Walker, C. L. F., Theodoratou, E., Nair, H., Campbell, H., & Black, R. E. (2013). [Breastfeeding for reducing the risk of pneumonia morbidity and mortality in children under two: a systematic literature review and meta-analysis](#). *BMC Public Health*, 13(3), 1–8.
- Lazzerini, M., Seward, N., Lufesi, N., Banda, R., Sinyeka, S., Masache, G., Nambiar, B., Makwenda, C., Costello, A., & McCollum, E. D. (2016). [Mortality and its risk factors in Malawian children admitted to hospital with clinical pneumonia, 2001–12: a retrospective observational study](#). *The Lancet Global Health*, 4(1), e57–e68.
- Omer, S. B., Clark, D. R., Aqil, A. R., Tapia, M. D., Nunes, M. C., Kozuki, N., Steinhoff, M. C., Madhi, S. A., & Wairagkar, N. (2018). [Maternal influenza immunization and prevention of severe clinical pneumonia in young infants](#). *The Pediatric Infectious Disease Journal*, 37(5), 436–440.
- Porter, R. S., Kaplan, J. L., Homeier, B. P., & Albert, R. K. (2009). *The Merck manual home health handbook*. Merck Research Laboratories.
- Razzaghi, M. (2013). [The probit link function in generalized linear models for data mining applications](#). *Journal of Modern Applied Statistical Methods*, 12(1), 19.
- Sarkar, S. K., & Midi, H. (2010). [Importance of assessing the model adequacy of binary logistic regression](#). *Journal of Applied Sciences*, 10(6), 479–486.
- Shakeel, S., Iffat, W., Qamar, A., Ghuman, F., Yamin, R., Ahmad, N., Ishaq, S. M., Gajdács, M., Patel, I., & Jamshed, S. (2021). [Pediatricians’ compliance to the clinical management guidelines for community-acquired pneumonia in infants and young children in Pakistan](#). *Healthcare*, 9(6), 701.
- Stoltzfus, J. C. (2011). [Logistic regression: a brief primer](#). *Academic Emergency Medicine*, 18(10), 1099–1104.
- Walker, C. L. F., Rudan, I., Liu, L., Nair, H., Theodoratou, E., Bhutta, Z. A., O’Brien, K. L., Campbell, H., & Black, R. E. (2013). [Global burden of childhood pneumonia and diarrhoea](#). *The Lancet*, 381(9875), 1405–1416.
- Yuan, X., Qian, S.-Y., Li, Z., & Zhang, Z.-Z. (2016). [Effect of zinc supplementation on infants with severe pneumonia](#). *World Journal of Pediatrics*, 12(2), 166–169.