Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 2 No. 1 2021, 40 - 46
DOI: *https://doi.org/10.22487/27765660.2021.v2.i1.15682*

eISSN: 2776-5660

ANALYSIS OF SKIN DISEASE INFECTION AFTER THE PALU EARTHQUAKE USING BINARY LOGISTIC REGRESSION

Selvia Anggun Wahyuni[1*], Lilies Handayani[2], Muhammad Akriyaldi Masdin[3], Salmia[4]
[1,2,3,4]Tadulako University, Central Sulawesi, Indonesia

***e-mail**: selviaanggun25@gmail.com

ABSTRACT

The incidence of skin disease in Indonesia is still relatively high and is a significant problem. This is evidenced by the 2010 Indonesian Health Profile data which shows that skin and subcutaneous tissue diseases are the third rank of the 10 most common diseases among outpatients in hospitals throughout Indonesia. Skin disease is growing, as evidenced by data from the Indonesian Ministry of Health, the prevalence of skin disease throughout Indonesia in 2012 was 8.46%, then increased in 2013 by 9 %. Palu City is an area that has a high skin disease problem. According to the 2016 BPS of Palu City, skin diseases are among the top 10 diseases in Palu City with a total of 11,363 sufferers. The method used in this research is binary logistic regression. Based on the analysis that has been done, it can be concluded that the best model is formed as follows: $\hat{y} = -5.880 + 4.495\, x_5 + e$. *Based on the best model, it is found that the factors that influence the transmission of skin diseases after the Palu earthquake are genetic factors.*

Keywords: *Central Sulawesi, Infectious Diseases, Skin Diseases, Binary Logistic Regression*

Cite: Wahyuni, S. A., Handayani, L., Masdin, M. A., & Salmia. (2021). Analysis of Skin Disease Infection After the Palu Earthquake Using Binary Logistic Regression. *Parameter: Journal of Statistics*, *2*(1), 40-46.

[*]Received: Nov 2021; Reviewed: Dec 2021; Published: Dec 2021

INTRODUCTION

The skin is an elastic covering that protects the body from environmental influences. One part of the human body that is quite sensitive to various diseases is the skin. A dirty environment will be a source of various diseases, including skin diseases (Harahap, 2000). Skin disease is an infectious disease that attacks the surface of the body and is caused by various causes. This disease is most often found in tropical countries, including Indonesia. This climate facilitates the growth of bacteria, parasites and fungi (Kristiwiani, 2005).

The incidence of skin diseases in Indonesia is still relatively high and is a significant problem. This is evidenced from the 2010 Indonesian Health Profile data which shows that skin and subcutaneous tissue diseases are the third rank of the 10 most common diseases in outpatients in hospitals throughout Indonesia. Skin diseases are growing, as evidenced by data from the Indonesian Ministry of Health, the prevalence of skin diseases throughout Indonesia in 2012 was 8.46% and then increased in 2013 by 9% (Departemen Kesehatan, 2013). Palu City is one of the areas that has a fairly high skin disease problem. According to BPS Palu City 2016, skin diseases are included in the 10 most common diseases in Palu City with 11,363 sufferers.

The earthquake incident on September 28, 2018 that hit Palu City caused many people to lose their homes so that people live in refugee areas until now. The low level of cleanliness, residential density and access to clean water makes it difficult for various skin diseases to occur in people living in refugee areas. The high density of occupancy and interaction or physical contact between individuals facilitates skin diseases. Therefore, the prevalence of skin disease is generally found in environments with high population density and high interpersonal contact.

The data used in this study used data on contracting skin diseases which were used as response variables which were categorized into two, namely contracting skin diseases and not contracting skin diseases. With factors that are considered to influence the transmission of skin diseases, namely age, gender, location of residence, water conditions, and genetic factors. Based on the data categories from the response variables above, the appropriate method used is the binary logistic regression method with the focus of the problem, namely looking for a binary logistic regression model in cases of contracting skin diseases after the Palu earthquake and analyzing what factors influence the transmission of skin diseases after the Palu earthquake.

MATERIALS AND METHODS

1. Data Sources

The data uses in this research is primary data obtained using data collection techniques through a digital questionnaire (google form) given to respondents with appropriate criteria for research. In this study, the observation units were temporary shelters for Palu residents after the earthquake, namely in Petobo (South Palu District) and Biromaru (Sigi Regency).

2. Research Variables

The research variables used are as follows: response variable (y) is residents who are infected with skin diseases or not. Predictor variables (x) are age, gender, location of residence, water conditions, and genetic factors.

3. Methods

Regression analysis is a data analysis that describes between a response variable and one or more predictor variables. Logistic regression is a method that can be used to find the relationship between response variables that are dichotomous (nominal or ordinal scale with two categories) or polychotomous (having a nominal or ordinal scale with more than two categories) with one or more predictor variables (Agresti, 1990).

Logistic regression model is a statistical modeling that is applied to model the response variables that are category based on one or more covariates (explanatory variables). Logistic regression models are often used in epidemiology, namely the study of the pattern of disease occurrence and the factors that influence it (Akbar, 2011). In modeling, it is assumed that these binary variables are independent of each other, so that the sum of the binary variables will have a binomial distribution (Rizki, et al., 2015).

Logistic regression can be used to predict a categorical dependent variable on the basis of continuous and/or categorical independents, to determine the effect size of the independent variables on the dependent, to rank the relative importance of independents, to assess interaction effects, and to understand the impact of covariate control variables. The impact of predictor variables is usually explained in terms of odds ratios.

Binary logistics regression method is a data analysis method used to find the relationship between the response variable (y) which is binary or dichotomous with the predictor variable (x) which is polychotomous. The output of the response variable y consists of 2 categories which are denoted by $y = 1$ (success) and $y = 1$ (failure) (Hosmer & Lemeshow, 2000). Based on this explanation, this method is considered suitable to be used to analyze the transmission of skin diseases after the Palu earthquake.

4. Data Analysis

The analytical methods used in this study are:

a. Make a descriptive analysis of the response variables and predictor variables.
b. Detect cases of multicollinearity of predictor variables with VIF test.
c. Perform parameter significance test using:
   i. Overall Test

$$G^2 = -2\ln\left[\frac{\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}}{\prod_{i=1}^{n}[\pi_0(x_i)^{y_{0i}}\pi_1(x_i)^{y_{1i}}\pi_2(x_i)^{y_{2i}}]}\right] \quad (1)$$

   ii. Wald Test

$$W = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \quad (2)$$

d. Perform a pseudo R^2 test.
e. Modeling cases of skin disease infection with binary logistic regression.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_i X_i + e \quad (3)$$

f. Testing the odds ratio, model accuracy, and goodness of fit.
g. Make a conclusion.

RESULTS AND DISCUSSION

1. Descriptive Statistics

In this study, researchers used 30 samples with variables y (skin condition), x_1 (age), x_2 (gender), x_3 (location of residence), x_4 (water conditions), dan x_5 (genetics factor). The following is a pie chart image of variables y (skin condition).

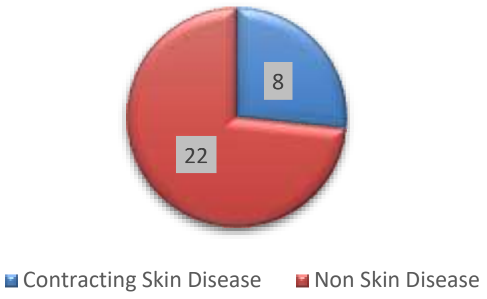

Figure 1. *Pie Chart* Variables Y (Skin Condition)

Figure 1. is a description of the skin conditions of 30 respondents. There were as many as 22 people who did not contract skin disease and there were 8 people who got skin disease so that from the binary logistic regression method it can be known the factors that affect the dependent variable y (skin condition) from the independent variables x known in the study.

2. Multicollinearity Test

The initial stage carried out in this research is multicollinearity testing. It is expected there will be no multicollinearity between independent variables. The hypotheses are:

H_0 : Data does not occur multicollinearity
H_1 : Data occur multicollinearity

Multicollinearity test uses the VIF value as a conclusion. Reject H_0 if the VIF value > 10. The VIF value can be seen in Table 1.

Table 1. VIF Value of Each Independent Variable

Independent Variables	VIF Value
x_1	1.324
x_2	1.369
x_3	1.681
x_4	1.374
x_5	1.162

Based on Table 1, it is known that the VIF value of the independent variable is < 10, so it can be concluded that it failed to reject H_0 or the data did not occur multicollinearity.

3. Parameter Significance Test

a. Overall Test

The overall test is carried out to see the effect of the independent variable (x) on the dependent variable (y) stimultaneously. The hypothesis are:

$\mathrm{H}_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (there is no effect of the independent variable on the dependent variable)
H_1: there is at least one $\beta_k \neq 0; k = 1,2,\dots,p$ (there is an effect of the independent variable on the dependent variable)

The criteria for rejection of the overall test hypothesis is reject H_0 if the value of Sig. < 0.05. The value of Sig. can be seen in Table 2.

Table 2. Overall Test

	B	S.E.	Wald	df	Sig.	Exp(B)
Constant	-1.012	0.413	6.004	1	0.014	0.364

Based on Table 2, the Sig. value of 0.014 is obtained which is smaller than the value of 0.05 (Sig. < α) meaning that it rejects H_0. So it can be concluded that the independent variable (x) has a significance effect on the dependent variable (y).

b. Wald Test

The overall test is carried out to see the effect of the independent variable (x) on the dependent variable (y) partially. The hypothesis are:

$H_0\colon \beta_k = 0$

$H_1: \beta_k \neq 0; k = 1,2, \ldots, k$

Wald test result can be seen in Table 3.

Table 3. Wald Test

Variable	$\boldsymbol{\beta}$	S.E.	Wald	df	Sig.	Exp($\boldsymbol{\beta}$)
X_1	1.402	1.121	1.564	1	0.211	4.063
X_2	2.957	1.634	3.275	1	0.070	19.241
X_3	-3.164	1.676	3.565	1	0.059	0.042
X_4	-0.990	2.321	0.184	1	0.668	0.370
X_5	4.495	1.830	6.031	1	0.014	89.568
Constant	-5.880	3.291	3.192	1	0.074	0.003

In Table 3, by looking at the comparison of each Wald value against the value of ${\chi^2}_{tabel}$ or the value of Sig. α (0.05) it was found that there was only one significant variable, namely the variables X_5 which had a Sig. value (0.014) < α (0.05) so it can be concluded that the variable x_5 (genetic factor) has a significant effect on the dependent variable (skin condition).

4. Pseudo R^2

This stage is the stage to see how the percentage of the independent variable (x) can explain the dependent variable (y). The value of pseudo R^2 can be seen in Table 4.

Table 4. Pseudo R^2 Test

-2 *Log likelihood*	*Nagelkerke R Square*
20.297	0.56

From Table 4, the Negelkerke R Square value is 0.56, which means that 56% of the independent variable factors (x) can explain the diversity of the dependent variable (y), while the remaining 46% are other factors outside the research model determined by researcher.

5. Logistic Regression Model

Based on the analysis, the best logistic regression model was obtained. Variable x_5 (genetic factor) has a significant effect on the dependent variable (skin condition). The value of the variable parameter x_5 (genetic factor) can be seen in Table 3. Thus, the best logistic regression model is obtained as follows: $\hat{y} = -5.880 + 4.495\, x_5 \ + e$.

6. Odds Ratio

Odds ratio is an opportunity divided by other odds. In this case, looking for the odds ratio aims to determine the tendency of the value of variable y (skin condition) if the x_5 variable (genetic factor) increases or there is a heredity skin disease. The value of odds ratio can be seen in Table 5.

Table 5. Nilai Odds Ratio

Variable	Odds Ratio
x_1	4.063
x_2	19.241
x_3	0.042
x_4	0.370
x_5	89.568
Constant	0.003

In Table 5. the odds ratio value for the x_5 variable (genetic factor) is 89.568 so it can be concluded that if the x_5 variable (genetic factor) increases or there is a hereditary skin disease, the tendency of respondents in cases to contract skin disease is 89.568 times. This shows that the x_5 variable (genetic factor) has a significant influence on the condition of a person who will contract skin disease after the Palu earthquake.

7. Model Accuracy

At this stage the aim is to find out how many percent of the model's accuracy is in classifying observations. The results of the accuracy of the model can be seen in Table 6.

Table 6. Model Accuracy

Observation		Prediction		
		Condition		Percentage
		No Skin Disease	Contracting Skin Disease	
Condition	No Skin Disease	22	0	100%
	Contracting Skin Disease	3	5	62.5%
Number of Percentage				90%

In Table 6, the results of the classification of the model's accuracy in classifying observations are 90%. This means that from 30 observations there are 27 observations that are rightly classified by the model. These results indicate that the model used is correct because it has a high percentage of accuracy.

8. Goodness Of Fit

This stage is the stage for the model fit whether it is sufficient to explain the data or vice versa. The hypothesis are:

H_0 : the model sufficient to explain the data (goodness of fit)
H_1 : the model does not sufficient to explain the data

The criteria for rejection of the goodness of fit hypothesis is reject H_0 if the value of Sig. < 0.05. The value of goodness of fit can be seen in Table 7.

Table 7. Hosmer and Lemeshow Test

Chi-Square	df	Sig.
6.79433	7	0.4506

Based on Table 7. the Sig. value is 0.4506 > α (0.05), so it failed to reject H_0. So it can be concluded that the model is sufficient to explain the data (goodness of fit).

CONCLUSION

Based on the analysis that has been done, it can be concluded that the best model formed is as follows: $\hat{y} = -5.880 + 4.495\, x_5 \ + e$. Based on the best model, it is found that the factors that influence the transmission of skin diseases after the Palu earthquake are genetic factor.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.

Akbar, M.A. (2011). *Analisis Regresi Logistik Multinominal Untuk Mengetahui Faktor-Faktor Utama yang Mempengaruhi Keputusan Mahasiswa Matematika UNM Setelah Selesai S1*. Makassar: Universitas Negeri Makassar.

Departemen Kesehatan RI. (2013). *Riset Kesehatan Dasar*. Jakarta: Badan Penelitian dan Pengembangan Kesehatan Kementrian Kesehatan RI.

Garson, G.D. (2021, December 29). *Logistic Regression*. Retrieved from https://faculty.chass.ncsu.edu/garson/pa765/logistic.htm

Harahap, M. (2000). *Ilmu Penyakit Kulit.* Jakarta: Hipokrates.

Hosmer, D.W., & Lemeshow, S. (2000). *Applied Logistic Regression.* New York: John Wiley & Sons, Inc.

Kristiwiani, D. (2005). *Hubungan Antara Praktik Kebersihan Diri Dengan Kejadian Skabies Pada Anak SD di SD Bandarhaja I Semarang.* Semarang: Universitas Diponegoro.

Rizki, F., Widodo, D. A. A., & Wulandari, S. P. (2015). Citation: Faktor Risiko Anemia Gizi Besi Pada Ibu Hamil di Jawa Timur Menggunakan Analisis Regresi Logistik. *Institut Teknologi Sepuluh Nopember Journal*, Vol. 4, No. 2.