Poverty modeling of regencies/municipalities in the island of Sumatera

D S Rini, D Agustina, I Sriliana, and P Novianti
Universitas Bengkulu, Jl. W.R. Supratman, Bengkulu 38125 Indonesia

E-mail: dyah.setyorini@unib.ac.id

Abstract. Binary logistic regression is a method used to explain the relationship between response variables that have two categories with one or more predictor variables that have a numerical or categorical scale. This research uses binary logistic regression to model and determine the factors that influence poverty in regencies/municipalities of Sumatera Island. In 2016, the percentage of poor people in Sumatera Island was third in Indonesia at 11.03 percent, so this problem needs to be studied further. The data used in this research is obtained from Central Bureau of Statistics in 2016. The response variable is binary where 0 for poor and 1 for not poor and also six predictor variables which include the percentage of Net Enrollment Rate, percentage of Expected Years of Schooling, percentage of GRDP Growth Rates, percentage of Per capita Average Expenditure for Food, percentage of Improved Drinking Water, and percentage of Life Expectancy. Based on this research, binary logistic regression model is giving a percentage of classification of 69.5 percent. Other than that percentage of Life Expectancy significantly affects poverty on the island of Sumatera.

1. Introduction

Sumatera island is the sixth largest island in the world that is located in Indonesia, at 473.481 km². Sumatera is an elongated landmass spanning a diagonal northwest-southeast axis. The Indian Ocean borders the west, northwest, and southwest coasts of Sumatera with the island chain of Simeulue, Nias and Mentawai off the western coast. In the northeast the narrow Strait of Malacca separates the island from the Malay Peninsula, which is an extension of the Eurasian continent. In the southeast the narrow Sunda Strait separates Sumatera from Java. The northern tip of Sumatera borders the Andaman Islands, while off the southeastern coast lie the islands of Bangka and Belitung, Karimata Strait and the Java Sea.The population of this island is around 57.940.351 people [1]. The number of regencies/ municipalities in Sumatera is 154 regencies/municipalities in 10 provinces [2]. Geographical conditions and a large population make Sumatera inseparable from social problems that is poverty.

In September 2016, the percentage of poor people on Sumatera Island reached 11.03 percent. This figure is more than the percentage of Indonesian poverty which is only 10.70 percent. This makes Sumatera Island become the third poor island in Indonesia. Whereas in September 2017, the percentage of poor people on Sumatera Island decreased by 0.59 percent, 10.44 percent. This figure is still more than the percentage of poverty in Indonesia, which is 10.12 percent [3]. The decline in the
percentage of poor people on the island of Sumatera is quite slow, therefore it is necessary to examine the factors that influence poverty on the island of Sumatera.

The factors that influence poverty in Sumatera Island can be known by using binary logistic regression. This method is a method used to explain the relationship between response variables that have two categories with one or more predictor variables that have numerical or categorical scales. Cases of poverty on the island of Sumatera are interesting to study using binary logistic regression, where the response variable has two categories, that is Poor and Not Poor.

Binary logistic regression has been carried out to model poverty and determine the factors that influence poverty. In 2008, logistic regression used in determining trends and determinants of rural poverty in several cities in Punjab [4]. A study using demographic data and health surveys is conducted to identify determinants of poverty with logistic regression [5]. Logistic regression also use to determine factors that can increase poverty [6]. Important factors that influence poverty can be determined by using logistic regression. The result is per capita consumption expenditure that significantly affects poverty [7]. And also, logistic regression is used to venture ratio of the probability occurrence of poverty in Zanzibar with social dimension. The study show that all district in Pemba are on high risk of being enter into poverty [8].

2. Binary Logistic Regression

Logistic regression is a method of data analysis that used to find a dependent variable relationship (nominal or ordinal in two categories) or polychotomus (has a nominal or ordinal scale with more than two categories) with one or more independent variables. Independent variables in logistic regression are continuous or categorical [9].

Binary logistic regression is a method used to explain the relationship between response variables that have two categories with one or more predictor variables that have a numerical or categorical scale. The binary scale response variable follows Bernoulli distribution with the probability function as follows:

\[ f(Y = y) = \pi^y (1 - \pi)^{1-y}; \ y = 0, 1 \]

where:
\[ \pi \] : probability of success for \( Y = 1 \)
\[ 1 - \pi \] : probability of fail for \( Y = 0 \)

The logistic regression probability model with \( p \)’s predictor variable is formulated as follows:

\[ \pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} \]

where \( \pi(x) \) is probability, \( 0 \leq \pi(x) \leq 1 \)

Based on the previous description, the logit transformation of \( \pi(x) \) is [10]:

\[ g(x) = \ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \]

Binary logistic regression parameters is estimated by using the maximum likelihood. Parameters \( \beta_0, \beta_1, \ldots, \beta_p \) are estimated by maximizing the likelihood function. Observations are assumed to be independent, if \( Y_i, i = 1, 2, \ldots, n \), a random sample has been taken, the likelihood function is as follows:
L(\beta) = \prod_{i=1}^{n} P(Y = y_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}

Maximized likelihood function:

\ln L(\beta) = \sum_{i=0}^{p} \left[ \sum_{i=1}^{n} y_i x_{ik} \right] \beta_k - \sum_{i=1}^{n} \ln \left[ 1 + \exp \left( \sum_{i=0}^{p} \beta_k x_{ik} \right) \right] 

(1)

Parameter \beta_0, \beta_1, ..., \beta_p are obtained from the first and second derivatives of equation (1) through an iterative procedure using the Newton-Raphson method [9].

3. Method

The data used in this study is taken from the Central Bureau of Statistics in 2016. The response variable is the percentage of poor people (Head Count Index / HCI) of regencies/municipalities on the island of Sumatera. The percentage of poor people is categorized into 0 (poor) and 1 (not poor). If HCI for a regencies/municipalities is greater than HCI of Sumatera, then it is categorized as 0 (poor), and vice versa. Predictor variables, thought to affect poverty in Sumatera Island, are the Percentage of Net Enrollment Rate, Percentage of Expected Years of Schooling, Percentage of GRDP Growth Rates, Percentage of Per capita Average Expenditure for Food, Percentage of Improved Drinking Water, and Percentage of Life Expectancy.

The initial step in this study is to describe poverty on the island of Sumatera and conduct multicollinearity tests. If the predictor variables do not contain multicollinearity, data modeling will be carried out using binary logistic regression. In the modeling process, it will be seen predictor variables significantly influence the response variable. The results obtained at this stage are the factors that influence poverty on the island of Sumatera. The next step is to calculate the classification accuracy of the binary logistic regression model. This step aims to see how accurate the classification formed by the model is.

4. Result and Discussion

4.1 Descriptive Statistic of Poverty In Sumatera Island

In 2016, the number of poor people in Indonesia reached 27,76 million people (10.70 percent). Sumatera Island contributes 6.21 million people (11.03 percent) of the poor population of the total number of poor people in Indonesia. This figure makes Sumatera be the third poor island in Indonesia. For region aspect, the number of poor people in urban areas is 2.06 million people (9.07 percent), while the number of poor people in rural areas is 4.15 million people (12.36 percent) [11]. In other words, the number of poor people in rural areas is more than the number of poor people in urban areas.

The poverty status of regencies/municipalities in Sumatera is determined based on the comparison of HCI in regencies/municipalities with HCI of Sumatera in 2016. If HCI of a regency/municipalities is greater than HCI of Sumatera, the regencies/municipalities are classified as poor regencies/municipalities (0). If HCI of a regency/municipality is smaller than HCI of Sumatera Island, then the Regency/municipality is classified as a regency/municipality Not Poor (1). Based on data from BPS in 2016, the number of regencies/municipalities on Sumatera Island classified as Poor...
regencies/municipalities is 78 regencies/municipalities. While the number of regencies/municipalities in Sumatera Island classified as Non-poor regency/municipality is 76 regencies/municipalities. Figure 1 is a poverty status map of regencies/municipalities in Sumatera Island. The map is created using the ArcGIS 10.3 software.

**Figure 1. Poverty Status of Regencies/Municipalities in Sumatera Island**

Descriptive statistics for each predictor variable are presented in the table 1:

| Variable       | $X_1$  | $X_2$  | $X_3$  | $X_4$  | $X_5$  | $X_6$  |
|----------------|--------|--------|--------|--------|--------|--------|
| Mean           | 63,09  | 12,95  | 4,81   | 62,87  | 50,05  | 68,37  |
| Standard Deviation | 12,18  | 1,03   | 1,27   | 5,00   | 21,79  | 2,48   |
| Maximum        | 89,31  | 17,03  | 6,81   | 72,4   | 98,00  | 73,6   |
| Minimum        | 36,05  | 10,92  | -1,35  | 48,17  | 2,32   | 60,44  |
The percentage of Net Enrollment Rate \( X_1 \) has an average value of 63.09. Standard deviation of 12.18 means that the variance of data is quite small. The maximum value of the variable is 89.31 percent, that is, the Regency/Municipality in Sumatera Island which has the highest percentage of Net Enrollment Rate is Bukittinggi, West Sumatera. While Regency/Municipality that have the lowest percentage of Net Enrollment Rate are Lematang Ilir, South Sumatera.

The percentage of Expected Years of Schooling \( X_2 \) has the mean and standard deviation respectively 12.95 and 1.03. In other words, this variable has very small variance. The Regency/Municipality in Sumatera that has the highest percentage of Expected Years of Schooling are Banda Aceh, Aceh (17.03 percent). While Regency/Municipality that has the lowest percentage of Expected Years of Schooling are Palembang, South Sumatera (10.92 percent).

The average value and standard deviation of the percentage of GRDP Growth Rates \( X_3 \) are 4.81 and 1.27 respectively. This variable also has a very small deviation. The largest percentage of GRDP Growth Rates is 6.81 percent owned by Kota Jambi, Jambi. While the smallest percentage of GRDP Growth Rates is -1.35 percent owned by Aceh Timur, Aceh.

The percentage of per capita Average Expenditure for Food \( X_4 \) has an means of 62.87 and a standard deviation of 5.00. In other words, the deviation of data is quite small. The Regency/Municipality with the highest percentage of per capita average expenditure for food is Karo, North Sumatera (72.40 percent). While the Regency/Municipality that have the lowest percentage of per capita average expenditure for food are Pekan Baru, Riau (48.17 percent).

The average values and standard deviations owned by the percentage of Improved Drinking Water \( X_5 \) were 50.25 and 21.79, respectively. Deviation in this variable is greatest when compared to other variables. However, the standard deviation of this variable is still relatively small. The maximum value is owned by the Regency/Municipality of Batam, Riau Islands, which is 98.00 percent. While the minimum value is owned by Regency/Municipality Empat Lawang, South Sumatera, which is 2.32 percent.

The sixth variable \( X_6 \) is the percentage of Life Expectancy. This variable has an average value and the standard deviation that are 68.37 and 2.48 respectively. The value of the standard deviation indicates that the data has a small deviation. The maximum value is in Bukittinggi, West Sumatera (73.6 percent). While the minimum value is owned by Lingga, Riau Islands (60.44 percent).

4.2 Binary Logistic Regression Model
The initial step in the binary logistic regression method is to test the multicollinearity of the predictor variables. Multicollinearity is tested by Variance Inflation Factor (VIF) value. The results of the calculation of the VIF value of the six predictor variables are as follows:

| Variable | \( X_1 \) | \( X_2 \) | \( X_3 \) | \( X_4 \) | \( X_5 \) | \( X_6 \) |
|----------|----------|----------|----------|----------|----------|----------|
| VIF      | 1.49     | 1.62     | 1.06     | 1.19     | 1.22     | 1.36     |
Table 2 shows the results of the VIF calculation of each predictor variable. These results indicate that there are no variables containing multicollinearity, because the VIF value of each variable is smaller than 10. So that all predictor variables can be used in a binary logistic regression modeling.

The next step is testing the regression model simultaneously. The used statistic test is the G test. Based on calculations using SPSS 16, G test statistics is 41.08. If the G test statistic value is compared to the value of \( \chi^2_{(0.05,6)} = 12.59 \), then the G test statistic value is greater than \( \chi^2_{(0.05,6)} \). The conclusion of testing the binary logistic regression model simultaneously is that there is at least one predictor variable that influences the poverty of Regency/Municipality in Sumatera island.

Predictor variables that affect the poverty of Regencies/Municipalities on the island of Sumatera can be seen from the results of partial testing. Estimated parameter and partial test results of binary logistic regression model using the R program and SPSS 16 are presented in the following table:

| Parameter | Estimate | Standard Error | Wald | Odds Ratio |
|-----------|----------|----------------|------|------------|
| \( \beta_0 \) | -33.41 | 8.39 | 15.83 | |
| \( \beta_1 \) | 0.01 | 0.02 | 0.24 | 1.01 |
| \( \beta_2 \) | -0.37 | 0.23 | 2.68 | 0.68 |
| \( \beta_3 \) | 0.27 | 0.15 | 3.16 | 1.32 |
| \( \beta_4 \) | 0.01 | 0.04 | 0.05 | 1.01 |
| \( \beta_5 \) | 0.001 | 0.01 | 0.004 | 1.001 |
| \( \beta_6 \) | 0.52 | 0.11 | 21.39 | 1.69 |

Table 3 shows the results of Wald test statistics for each predictor variable. A predictor variable has a significant effect if the statistical value of the Wald test is greater than \( \chi^2 \)-table, \( \chi^2_{(0.05,1)} = 3.84 \). Predictor variables that significantly influence the poverty of Regencies/Municipalities in Sumatera Island at the 5 percent significant level are the percentage of Life Expectancy. The binary logistic regression model is

\[ \pi(x) = -33.41 + 0.01X_1 - 0.37X_2 + 0.27X_3 + 0.01X_4 + 0.001X_5 + 0.52X_6 \]

Table 4 presents the results of calculating the deviance value using the R program. Deviance value is a statistic test that is used to test the goodness of the binary logistic regression model. The statistical value of the deviance test is 172.38 which is less than the value of \( \chi^2_{(0.05,145)} = 174.101 \). The conclusion for testing the goodness of the model at the level of 5 percent is a binary logistic regression
model according to the case of poverty in Regencies/Municipalities on the island of Sumatera (there is no difference between the results of observations with possible predictions of the model).

| Table 4. Deviance Value of Model |
|----------------------------------|
| Deviance | Db | $\chi^2_{(0.05;145)}$ |
|----------|----|---------------------|
| 172.38   | 147| 174.101             |

The last step in binary logistic regression is determining the classification accuracy. This step aims to find out how much the percentage of observations can be classified correctly by the model. The results of the classification accuracy of the binary logistic regression model can be seen in the following table:

| Table 5. Classification Accuracy |
|----------------------------------|
| Observed | Predicted | Percentage Correct (percent) |
|----------|-----------|-----------------------------|
| Poor (0) | 53        | 67.9                        |
| Not Poor (1) | 22 | 71.1                        |
| Overall Percentage | 69.5 |

Table 5 shows that Regencies/Municipalities in Sumatera that have poor status are correctly classified by 67.9 percent, while 32.1 percent are not properly classified. Regencies/Municipalities in Sumatera that have a non-poor status are correctly classified by 71.1 percent, while 28.9 percent are not properly classified. The overall percentage of Regencies/Municipalities in Sumatera Island is correctly classified by 69.5 percent. This shows that the binary logistic regression model can classify 108 Regencies/Municipalities from 154 Regencies/Municipalities in Sumatera Island correctly according to the poverty status of each Regency/Municipality. Based on these results it can be concluded that the logistic regression model is quite good in the case of poverty in Regencies/Municipalities on Sumatera Island.

5. Conclusion
The results of this study are binary logistic regression models giving a percentage of classification accuracy of 69.5 percent. As well as percentage of Life Expectations affect poverty significantly on Sumatera Island.

References
[1] Wikipedia 2018 Sumatera (https://en.wikipedia.org/wiki/Sumatera)
[2] BPS 2017 Data dan Informasi Kemiskinan Tahun 2016 (Jakarta: BPS)
[3] BPS 2018 Berita Resmi Statistik Profil Kemiskinan di Indonesia September 2017 (Jakarta: BPS)
[4] Hashmi A A, Sial M H and Hashmi M H 2008 Trends and Determinants of Rural Poverty: A Logistic Regression Analysis of Selected Districts of Punjab The Pakistan Development Review 47 909-23
[5] Achia T N O, Wangombe A, Khadioli N 2010 A Logistic Regression Model to Identify Key
Determinants of Poverty Using Demographic and Health Survey Data *European Journal of Social Science* **13** 38-45

[6] Rusnak Z 2012 *Logistic Regression Model in Poverty Analyses* (Wroclaw: Wroclaw University of Economics) **1** 9-23

[7] Xhafaj E and Nurja I 2014 Determination of the Key Factors that Influence Poverty Through Econometric Model *European Scientific Journal* **10** 65-72

[8] Makame I D and Mzee S S 2014 Determinant of Poverty on Household Characteristics in Zanzibar: A Logistic Regression Model *Developing Country Studies* **4** 188-95

[9] Agresti 2002 *Categorical Data Analysis* (New York: John Wiley & Sons)

[10] Hosmer D W and Lemeshow J S 2000 *Applied Logistic Regression* (Canada: John Wiley & Sons Inc.)

[11] BPS 2017 *Berita Resmi Statistik Profil Kemiskinan di Indonesia September 2016* (Jakarta: BPS)