Binary logistic regression for identification of high school student interest in Banda Aceh city in continuing study at Universitas Syiah Kuala

Nurhasanah¹, A Rusyana¹-²,* and Fitriana A R¹

1 Department of Statistics, Universitas Syiah Kuala, Banda Aceh, Indonesia 23111
2 Department of Statistics, IPB University, Bogor, Indonesia 16680

*Corresponding author: asep.rusyana@unsyiah.ac.id

Abstract. Universitas Syiah Kuala is the oldest university in Aceh. The campus has the nickname of the jantong hate campus to the Acehnese people, which of course can be interpreted as the campus of pride of the Acehnese people. Universitas Syiah Kuala increased its accreditation status from C to A accreditation awarded by the research, technology and higher education ministry. This showed a very significant quality of improvement. With this condition a study was conducted on the interest of high school students in the Banda Aceh City in choosing Universitas Syiah Kuala as a place to study. This study aims to determine the factors that influence the interest of Banda high school students in continuing their studies to Universitas Syiah Kuala and get a model of interest in the Banda Aceh city students through a binary logistic regression approach. The results show that the factors that influence students' interest in continuing their studies at Universitas Syiah Kuala are the ability to be affiliated with others, goals, and expectations. With the binary logistic regression model obtained, it can be seen the tendency of students to continue their studies at Universitas Syiah Kuala.

1. Introduction

Self-education is carried out by humans from birth to leaving the world. But formally, the level of education starts from the level of kindergarten, then elementary, junior high school, senior high school, and also higher education. The higher education stage is characterized by the availability of diverse choices of department or study program, allowing students to choose according to their talents and interests. Continuing education to college is a dream of most students. So it is not surprising that every student who passed from senior high school will compete to take their favorite college entrance examination test.

Indonesia itself has universities spread throughout its province, not only in Aceh Province. Universitas Syiah Kuala is one of the most desirable universities in Aceh. The high interest of students to continue their studies to Universitas Syiah Kuala certainly does not just happen but is influenced by several factors. One of the factors that is thought to be the reason is influenced by the student's interest. Many choices of departments and their location with high accessibility and Islamic education culture are thought to be the trigger for students' interest in the university.

Students' interest in Universitas Syiah Kuala in this study was divided into two categories, namely interests and not interests, so one of the methods that was used was the binary logistic regression method. Binary regression is a regression model that is used to explain the relationship between the nominal response variable with independent variables in the form of categorical or numeric. With this method,
the model of student interest is obtained in analyzing the factors that influence the interest of students in the city of Banda Aceh in continuing their studies at Universitas Syiah Kuala.

2. Literature review

2.1. Chi-Square test
The Chi Square test is useful for testing the relationship or influence of two variables and measuring the strength of the relationship between one variable and another [1]. The Chi Square test equation is as follows:

\[ X^2 = \sum_{ij} \frac{(X_{ij} - E_{ij})^2}{E_{ij}} \]  \hspace{1cm} (1)

where:
\[ \chi^2 \] = Value of Chi Square,
\[ X_{ij} \] = the frequency observed in the \( i \)-th row and \( j \)-th column,
\[ E_{ij} \] = the expected frequency in the \( i \)-th row and \( j \)-th column,
\[ n_i \] = the number of the \( i \)-th row frequency,
\[ n_j \] = the number of the \( j \)-th column frequency,
\[ n \] = the sum of all frequencies from the contingency table.

2.2. Binary Logistic Regression analysis
The purpose of analyzing categorical data using logistic regression is to get the best and simplest model to explain the relationship between the output of the response variable \( Y \) and its predictor variables \( X \) [2]. The response variable in logistic regression can be either category or quantitative. If the variable \( Y \) is a binary variable or dichotomy in a response variable consisting of two categories namely "success" \( (Y = 1) \) or "fail" \( (Y = 0) \), then the variable \( Y \) follows the Bernoulli distribution which has a probability density function [3]:

\[ f(y_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}; \ y_i = 0,1; \ i = 1,2, ..., n \]  \hspace{1cm} (2)

where:
\[ i = 1, 2, ..., n; \ n = \text{the number of observations} \]
\[ \pi_i = \text{probability of the} \ i-th \text{observation} \]
\[ x_i = \text{value of variable x at} \ i-th \text{observation} \]

so that it is obtained:
For \( y_i = 0 \), then \( f(0) = \pi(x_i)^0(1 - \pi(x_i))^{1-0} = 1 - \pi(x_i) \),
For \( y_i = 1 \), then \( f(1) = \pi(x_i)^1(1 - \pi(x_i))^{1-1} = \pi(x_i) \)

Suppose the probability of the response variable \( Y \) for the given value \( x \) is denoted as \( \pi(x) \). The general model \( \pi(x) \) is denoted as follows:

\[ \pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)} \]  \hspace{1cm} (3)

Equation (3) is called a logistic regression function that shows the relationship between predictor variables and probabilities that are not linear, so to get a linear relationship transformation is often called logit transformation.

The logit form of \( \pi(x) \) is
\[ \text{logit}[\pi(x)] = g(x) = \left( \frac{\pi(x)}{1-\pi(x)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \]  \hspace{1cm} (4)
Equation (4) is a function of the relationship of logistic regression models called multiple logistic regression models [2].

The parameter estimation method that is often used in logistic regression is the Maximum Likelihood Estimation (MLE) method which is used to obtain the maximum settlement of the likelihood function. Each observation for the logistic regression model is a random variable from the Bernoulli distribution. The principle of the Maximum Likelihood Estimation method is to maximize the likelihood function. Suppose that it is taken n samples of random variables \(Y_1, Y_2, \ldots, Y_n\) with \(Y_i = [y_1, y_2, \ldots, y_n]^T\) and the probability of the h-category result is \(p_h(x_i)\), the likelihood function is:

\[
L(\beta_0, \beta_1, \ldots, \beta_k) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}
\]

where:

\(y_i\) = observation on variable \(i\)

\(\pi_i\) = probability for \(i\)-predictor variable

There are several assumptions of the binary logistic regression. One of them is the data which have homoscedastic characteristic. For the heteroscedastic, transformation of data can be conducted or a method is applied to control it. It is like ARCH method in time series [4].

2.3. Statistic test

After parameter estimation, the next step that must be done is to test the significance of the expected parameters. The test used to test the significant \(\beta\) coefficient of the model can use the test simultaneously or partially.

2.3.1. Simultaneous test

Simultaneous testing is done in an effort to examine the role of predictor variables in the model together. Hypothesis:

\(H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0\), (together the predictor variable has no significant effect on the response variable)

\(H_1: \) There is at least one \(\beta_i \neq 0\) \((i = 1, 2, 3, \ldots, k)\), (there is at least one predictor variable which has a significant effect on the response variable)

The test statistic used is the G test statistic or the likelihood ratio test:

\[
G = -2 \ln \left( \frac{L_0}{L_1} \right) = -2 \ln \left( \frac{\prod_{j=1}^{n} \pi_j^{y_j} (1 - \pi_j)^{1-y_j}}{\prod_{j=1}^{n} \pi_j^{y_j} (1 - \pi_j)^{1-y_j}} \right)
\]

\[
G = 2[A - B]
\]

Where:

\[
A = \sum_{i=1}^{n} (y_i \ln(1 - p_j) + (1 - y_i) \ln(1 - p_j))
\]

\[
B = [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)]
\]

Where:

\(n_1\) = number of observations categorized as 1

\(n_0\) = number of observations categorized as 0

\(n\) = number of observations

\(y_j\) = the number of variables the response to the \(j\)-th observation, and where \(j = 1, 2, 3, \ldots, n\).

\(p_j\) = probability value in the \(j\)-th observation

The test statistic of the G follows the chi-square distribution with rejection criteria if \(G > \chi^2_{(df, \alpha)}\) [2].
2.3.2. Partial test
Partial testing is done to determine the significance of the parameters of the dependent variable. Partial or individual test results will show whether an independent variable is feasible to be included in the model or not [5], with the following hypothesis:

H₀: The i-th predictor variable is not a significant factor for the response variable, βᵢ = 0, for i = 1,2,3, ..., k
H₁: The i-th predictor variable is a significant factor for the response variable, βᵢ ≠ 0, for i = 1,2,3, ..., k

Where: i = number of predictor variables

The test statistics are as follows:

\[ Wald (W) = \frac{\beta_i}{SE(\beta_i)} \]  \hspace{1cm} (7)

Rejection ratios generated from test statistics, under the hypothesis will follow the normal distribution [2]. So in obtaining a decision a comparison is made with the standard normal distribution and the criteria for rejecting H₀ if the value W > z₀₂.

2.4. Model conformity test
The conformity test is a test used to see the suitability of a model. One test that can be used to see the suitability of the model is the Hosmer and Lemeshow Test. If the Hosmer and Lemeshow Tests are fulfilled, the model is able to predict the observation value or the model is appropriate.

The hypotheses for model suitability tests are as follows:

H₀: Model is suitable
H₁: Model is not suitable

The Hosmer and Lemeshow (Ĉ) test statistic which are calculated based on the value y = 1 are:

\[ Ĉ = \sum_{r=1}^{g} \frac{(o_r - \bar{n}_r \bar{p}_{1r})}{\bar{n}_r \bar{p}_{1r} (1 - \bar{p}_{1r})} \]  \hspace{1cm} (8)

where:
\( \bar{p}_{1r} \) = the estimated average probability of success of the r group
\( o_r \) = number of successful occurrence samples in the fourth group.
\( n_r \) = total sample r-group event, with r = 1,2, ..., g.

The test statistic Ĉ approaches the Chi-Square distribution with degree of freedom G-2 [2]. Rejection criteria if Ĉ > χ²(α,G-2).

3. Data and method
The data in this study uses primary data. Data collection in this study was conducted through a survey using a measuring instrument in the form of a questionnaire. This data collection was carried out in senior high schools in Banda Aceh City. The study was conducted from January until April 2019. The type of data obtained in the form of qualitative data. In the measurement of this study there were 23 statement items related to 5 interest indicators and 11 questions related to respondent identity, and there were 5 questions asked additional information about the respondents. Data retrieval of this research was conducted in 6 (six) Public High Schools in Banda Aceh City.

The selection of respondents was done using gradual sampling. The first stage was grouping based on the strata. The strata used are divided into 3 (three), the first strata consisted of state high school students who attend favorite high school level, second strata consisted of public high school students who attend intermediate high school level, and third strata consisted of public high school students who attend ordinary high school level. Table 1 below classifies SMA in 3 (three) levels, namely favorite, intermediate, and ordinary.

Schools are selected using simple random sampling for each stratum. Two (2) high schools were selected for each strata and for each selected high school, two (2) classes were randomly selected, namely class XI and XII for both science and social science majors. Classification of schools according
to the strata can be seen in Table 1. Each strata has a different number of respondents, so that a comparable allocation is used in determining respondents for each strata and to determine the number of respondents taken, we used the Slovin formula.

| No | Name of School       | Rank        | The Number of Students | Total | Sample |
|----|----------------------|-------------|------------------------|-------|--------|
|    |                      |             | Class XI               | Class XII |       |        |
| 1  | SMAN 3 Banda Aceh    | Favorite    | 309                    | 285        | 594    |        |
| 2  | SMAN 4 Banda Aceh    | Favorite    | 270                    | 254        | 524    |        |
|    | Sub Total            |             |                        | 1118     | 231    |        |
| 3  | SMAN 5 Banda Aceh    | Intermediate| 221                    | 208        | 429    |        |
| 4  | SMAN 8 Banda Aceh    | Intermediate| 227                    | 229        | 456    |        |
|    | Sub Total            |             |                        | 885      | 190    |        |
| 5  | SMAN 14 Banda Aceh   | Ordinary    | 38                     | 36         | 74     |        |
| 6  | SMAN 16 Banda Aceh   | Ordinary    | 51                     | 63         | 114    |        |
|    | Sub Total            |             |                        | 188      | 95     |        |
|    | Total                |             |                        | 2191     | 516    |        |

In this study there were 23 statement items related to 5 interest indicators used to compile a binary logistic model. The answer score for each question for each indicator is totaled. The following is an explanation of all the variables used in this study:

| Variable                  | Measurement Scale | Remark                           |
|---------------------------|-------------------|-----------------------------------|
| Student Interest (Y)      | Categorical Scale | 1. Interest 2. No Interest        |
| Need to affiliate with someone else (X_1) | Numeric | Score |
| Action (X_2)              | Numeric           | Score                             |
| Emotional Attitude (X_3)  | Numeric           | Score                             |
| Goals (X_4)               | Numeric           | Score                             |
| Expectation (X_5)         | Numeric           | Score                             |

Besides binary logistic regression, the research goals can also be achieved by Chi-Squared Automatic Interaction Detection (CHAID) method. From the results of data analysis using CHAID method, the model of the classification tree is constructed to obtain factor which influence response variable [6]. However, logistic regression is better in the case that the independent variable has ordinal scale.

4. Results and discussion

4.1. Variable characteristics
The respondents' data in this study are data of high school students in the city of Banda Aceh with the composition showed in Table 3.

The number of students who were respondents in this study amounted to 516 students consisting of 224 male students and 292 female students. With categories of natural science and social science majors, each department consists of 278 students and 238 students. The variables used in this study were 6 variables. The variable consists of one response variable and 5 predictor variables. Figure 1 represents the characteristic of each variable.
Figure 1. Variable characteristics of (a) affiliation with others, (b) actions, (c) emotional attitude, (d) goals, (e) interest and (f) expectations.

Based on Figure 1(a) it can be concluded that in the variable need to be affiliated with other people, respondents indicated that 55% of students answered agree, 39% of students answered very agree, 5% of students answered disagree and 1% of students answered very disagree. Based on Figure 1(b) for the action variable shows that 47% of students answered strongly agree, 43% of students answered agree, 9% of students answered disagree and 1% of students answered strongly disagree. Based on Figure 1(c) it can be seen that the emotional attitude variable responds as much as 51% answer agree, 32% answer strongly agree, 15% answer disagree and 2% answer strongly disagree. Based on Figure 1(d) for the destination variable 44% answered agree, 35% answered strongly agree, 19% answered disagree and
2% answered strongly disagree. According to Figure 1(e) variable of interest in high school students in Banda Aceh, 53% answered agree, 41% answered strongly agree, 5% answered disagree and 1% answered strongly disagree. According to Figure 1(f) on variable of expectations of high school students in Banda Aceh, 67% answered strongly agree, 31% answered agree, 1% answered disagree and 1% answered strongly disagree.

Table 3. Identity of respondents.

| No | Name of School       | Rank of School | Gender | Natural Science | Social Science | Sample |
|----|----------------------|----------------|--------|----------------|----------------|--------|
|    |                      |                | M      | F              |                |        |
| 1  | SMAN 3 Banda Aceh    | Favorite       | 44     | 67             | 55             | 56     | 111    |
| 2  | SMAN 4 Banda Aceh    | Favorite       | 44     | 76             | 62             | 58     | 120    |
|    | Sub Total            |                |        |                |                |        | 231    |
| 3  | SMAN 5 Banda Aceh    | Middle         | 48     | 53             | 56             | 45     | 101    |
| 4  | SMAN 8 Banda Aceh    | Middle         | 46     | 43             | 56             | 33     | 89     |
|    | Sub Total            |                |        |                |                |        | 190    |
| 5  | SMAN 14 Banda Aceh   | Ordinary       | 27     | 32             | 29             | 30     | 59     |
| 6  | SMAN 16 Banda Aceh   | Ordinary       | 15     | 21             | 20             | 16     | 36     |
|    | Sub Total            |                |        |                |                |        | 95     |
|    | **TOTAL**            |                | 224    | 292            | 278            | 238    | 516    |

4.2 Chi-Square test

The initial stage before modeling is the selection of variables involved in the test using the chi square method. The following is an analysis of the selection of these variables.

Table 4. Chi-Square test coefficient.

| Name of Variable | df | \( \chi^2 \) | P-value |
|------------------|----|-------------|---------|
| Interest Vs Gender | 1  | 1.172       | 0.279   |
| Interest Vs Father Education | 4  | 4.210       | 0.378   |
| Interest Vs Mother Education | 4  | 3.342       | 0.502   |
| Interest Vs Father Employment | 5  | 1.466       | 0.917   |
| Interest Vs Mother Employment | 5  | 2.222       | 0.818   |
| Interest Vs Father Income | 42 | 27.109      | 0.964   |
| Interest Vs Mother Income | 31 | 15.185      | 0.992   |
| Interest Vs X_1  | 12 | 122.79      | < 2.2 \times 10^{-16} |
| Interest Vs X_2  | 10 | 126.28      | < 2.2 \times 10^{-16} |
| Interest Vs X_3  | 12 | 55.966      | 1.214 \times 10^{-7}  |
| Interest Vs X_4  | 15 | 69.452      | 5.592 \times 10^{-9}  |
| Interest Vs X_5  | 15 | 120.39      | < 2.2 \times 10^{-16} |

Table 4 displays the results of the chi-square test between interest variables with several other variables. This test is declared reject \( H_0 \) if p-value < alpha, with alpha equal to 0.1. The results in Table 4 show that there are 5 significant tests, namely testing between interest variables with variables \( X_1, X_2, X_3, X_4, \) and \( X_5 \). Therefore, the five variables were included in the interest analysis using the Logistic Regression Analysis.
After data modeling, the next step is to test the parameters of the model obtained. The following parameters are tested simultaneously.

| Table 5. Omnibus tests coefficient. |
|-----------------------------------|
| Chi-Square | Df | Sig  |
|------------|----|------|
| Step       | 47.866 | 5 | <0.001 |
| Block      | 47.866 | 5 | <0.001 |
| Model      | 47.866 | 5 | <0.001 |

The results of the Omnibus Tests in Table 5 show that the Sig values obtained are less than alpha (0.000 < 0.1). This means that there is at least one independent variable tested which affects the model obtained. Then the partial parameters are tested to find out the significant parameters in the model.

A parameter is said to be significant if the value is sig < alpha, with alpha equal to 0.1. Table 6 shows that there are 3 parameters that are significant to the model, namely variables $X_1$ (ability to affiliate with others), $X_4$ (goals), and $X_5$ (expectations). This shows that the three variables influence the model analyzed. Then the feasibility testing of the model using the Hosmer and Lemeshow Test was conducted.

The test shows that Chi-Square value is 0.586 or the sig > alpha value (1 > 0.1). This shows that there is no difference between observation and prediction. This means that the model obtained is significant and feasible to use.

| Table 6. Value of estimated model parameters. |
|-----------------------------------------------|
| Estimation | SE | Wald | Df | Sig  | Exp  |
|------------|----|------|----|------|------|
| Intercept  | -13.486 | 3.768 | 12.811 | 1 | <0.001 | <0.001 |
| $X_1$      | 0.367 | 0.141 | 6.772 | 1 | 0.009 | 1.444 |
| $X_2$      | 0.329 | 0.344 | 0.912 | 1 | 0.339 | 1.389 |
| $X_3$      | -0.013 | 0.284 | 0.002 | 1 | 0.964 | 0.987 |
| $X_4$      | 0.538 | 0.316 | 2.894 | 1 | 0.089 | 1.712 |
| $X_5$      | 0.157 | 0.092 | 2.897 | 1 | 0.089 | 1.170 |

4.3. Binary Logistic Regression

After testing statistics, the mathematical equations from Table 6 can be written as follows.

$$\ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = -13.486 + 0.367X_1 + 0.329X_2 - 0.013X_3 + 0.538X_4 + 0.157X_5$$

The above equation can explain the magnitude of the tendency of high school students in Banda Aceh to continue their education to Universitas Syiah Kuala which is influenced by each variable. If the $X_1$ variable increases by one unit, then the odds will multiply by $e^{0.367}$ assuming other variables are constant. This means that an increase in variable $X_1$ (affiliated with another person) is equal to one unit, students will have a tendency to be interested in continuing their studies to Universitas Syiah Kuala at 1.443 times. If the $X_2$ (action) increases by one unit, then the odds will multiply by $e^{0.329}$ assuming other variables are constant. This means that if the action variable increases by one unit, students will have an inclined tendency to continue their studies to Universitas Syiah Kuala for 1.389 times.

If the $X_3$ (emotional attitude) increases by one unit and the other variables are constant, students will have a tendency to be interested in continuing their studies to Universitas Syiah Kuala at 1.170 times.
5. **Conclusion**

Based on the results and discussion can be concluded as follows:

1. The factors that influence students’ interest in continuing their studies to Universitas Syiah Kuala are the ability to be affiliated with others (X₁), goals (X₄), and expectations (X₅).

2. The binary logistics model can explain the tendency of high school students in Banda Aceh to continue their education to Universitas Syiah Kuala which is influenced by each of these variables.

6. **References**

[1] Mason R D, Lind A D and Marchal W G 1999 *Statistical Techniques in Business and Economics* (McGraw Hill)

[2] Hosmer D W . and Lemeshow S, 2000 *Applied Logistic Regression* (Canada: John Wiley and Sons)

[3] Casella G and Berger R L 2002 *Statistical Inference* (USA: Duxbury)

[4] Rusyana A, Ferdhiana R and Putri M E 2019 Application of ARCH model on nutmeg price forecasting in South Aceh District *IOP Conf. Ser. Mater. Sci. Eng.* **523** 1 012003

[5] Agresti 1990 *Categorical Data Analysis* (New York: John Wiley Sons Inc.)

[6] Fitriana A R, Aida J, Salwa N and Rusyana A 2018 Classification of the length of study based on the student characteristics and academic performance in FMIPA Unsyiah in *Journal of Physics: Conference Series* **1116** 2 022009