JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE
VOLUME 2, No 2, October 2023
e-ISSN: 2828-9986
https://ejournal.unib.ac.id/index.php/jsds/index

# An Analysis of Factors Contributing to Extended Study Duration Among Students of the Faculty of Mathematics and Natural Sciences, University of Bengkulu Using Binary Logistic Regression

**Indah Wahyuliani[1], Pepi Novianti [2*]**

[1] Statistics Study Program, University of Bengkulu
[2] Statistics Study Program, University of Bengkulu

* pie_novianti@unib.ac.id

**Article Info**

**Article History:**
*Received: 23 04 2025*
*Revised: 24 04 2025*
*Accepted: 25 04 2025*
*Available Online: 25 04 2025*

**Key Words:**
*Binary Logistic Regression*
*Oods Ratio*
*GPA*

**Abstract**

Logistic regression is a statistical method used to analyze the relationship between a dichotomous dependent variable and one or more independent variables, which may be numerical or categorical. In this study, binary logistic regression is applied to identify the factors influencing the study duration of students in the Faculty of Mathematics and Natural Sciences at the University of Bengkulu. These factors include both internal and external elements, such as cumulative GPA (Grade Point Average), gender, parents' occupation, scholarship status, and university admission pathway. The results show that GPA significantly affects the length of study, with an odds ratio of 1102.13, indicating that each one-unit increase in GPA greatly increases the likelihood of graduating on time. This study suggests the use of additional statistical techniques, such as bootstrapping, to enhance parameter estimation accuracy and recommends reporting effect sizes, such as odds ratios, for a more comprehensive interpretation of the relationship between independent and dependent variables.

## 1. INTRODUCTION

Regression analysis is a technique used to explain the nature of the relationship between two or more variables, particularly those that involve causal relationships [4]. For categorical data, logistic regression analysis is appropriate. In statistics, logistic regression—also known as the logistic or logit model—is used to predict the probability of an event occurring, using the logit function derived from the logistic curve. This approach predicts a dichotomous dependent variable—such as yes/no, good/bad, or high/low—without assuming that error variance is normally distributed, as it follows a logistic distribution. [3].

This paper focuses on binary logistic regression, a method for determining the relationship between a categorical dependent variable with two categories and one or more predictor variables [1]. The success rate of students in completing their education is influenced by various conditions. Factors influencing educational success originate from both internal (e.g., intelligence, emotional state, psychological condition) and external (e.g., family, community, campus environment, educational facilities, learning motivation) sources [5]. Many of these influencing factors, such as gender, parents' occupation, scholarship status, and university admission pathway, are categorical. Thus, this paper applies binary logistic regression to examine the study duration of students in the Faculty of Mathematics and Natural Sciences at the University of Bengkulu.

### 1.1 Descriptive Statistics

Descriptive statistics involve methods for collecting and presenting data to convey meaningful information. These often include graphical presentations such as histograms, pie charts, ogives, polygons, and stem-and-leaf plots [7].

### 1.2 Test of Independence

The Test of Independence determines the relationship between two variables, assuming homogeneity and mutual exclusivity in data categorization. A nominal scale differentiates categories without implying order, while an ordinal scale indicates rank. The chi-square test statistic is used in this analysis:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \quad (1)$$

Dimana $O_{ij}$ is the observed frequency (actual data) in the cell at row $i$ and column $j$. And $E_{ij}$ is the expected frequency (theoretical value) in the cell at row $i$ and column $j$.

### 1.3 Logistic Regression Analysis

Logistic regression models relationships between a dichotomous (nominal/ordinal scale with two categories) or polychotomous (nominal/ordinal scale with more than two categories) and one or more predictor Logistic regression models include binary, ordinal, and multinomial logistic regression. Binary logistic regression, which this study uses, is suited for a binary (dichotomous) dependent variable ($y$) and continuous predictor variables ($x$) [4].

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} \quad (2)$$

### 1.4 Binary Logistic Regression Analysis

Logistic regression is a method used to examine the relationship between a dichotomous (nominal or ordinal scale with two categories) dependent/responding variable and one or more predictor/independent variables on either a categorical or continuous scale. Binary logistic regression is a data analysis method used to find the relationship between a binary or dichotomous dependent/responding variable and a categorical or continuous predictor/independent variable (Pamungkas, 2017). Each observation is classified as "success" or "failure," denoted by 0 and 1, with the $i$ observation from the sample ($i = 1,2, \ldots, n$). The variable $Y_i$ follows a Bernoulli distribution with parameter $\pi_i$ and has a probability function as shown in Equation (3) below [2].

$$f(y_i;\pi_i) = \pi^{y_i}(1-\pi_i)^{1-y_i}; y_i = 0,1 \quad (3)$$

### 1.5 Parameter Estimation

Parameters are estimated by maximizing the likelihood function, assuming observations follow a Bernoulli distribution. The log-likelihood function is maximized using the Newton-Raphson method to solve for parameters [2]. Let $X_i$ and $Y_i$ be the predictor/independent variable and the response/dependent variable for the i-th observation. It is assumed that each observation is independent of the others, where $i = 1, 2, \ldots, n$. The probability function for each pair is shown in Equations (3) and (4) as follows:

$$\pi(X_i) = \frac{e^{\left(\sum_{j=0}^{p}\beta_j X_j\right)}}{1 + e^{\left(\sum_{j=0}^{p}\beta_j X_j\right)}} \quad (4)$$

When $i = 1, 2, \ldots, n$ and n is the number of observations. Each observation is assumed to be independent, so the Likelihood function is the product of the probability functions of each observation, as shown in Equation (5) below:

$$L(\beta) = \prod_{i=1}^{n} f(x_i) = \pi(X_i)^{y_i}\left(1 - \pi(X_i)\right)^{1-y_i} \quad (5)$$

The Likelihood function is maximized in the form of the log Likelihood, denoted as $L(\beta)$, as shown in Equation (6) bellow:

$$L(\beta) = \sum_{j=0}^{p}\left[\sum_{i=1}^{n} y_i X_{ij}\right]\beta_j - \sum_{i=1}^{n} \ln\left[1 + exp\left(\sum_{j=1}^{p}\beta_j X_{ij}\right)\right] \tag{6}$$

Equation (6) is differentiated with respect to β to find the maximum value of $\beta$, or by taking the derivative with respect to $\beta$ and setting it equal to zero. This leads to the result in Equation (7) bellow:

$$\frac{\partial L(\beta)}{\partial \beta_J} = \sum_{i=1}^{n} y_i X_{ij} - \sum_{i=1}^{n} X_{ij}\left(\frac{e^{\left(\sum_{j=0}^{p}\beta_j X_j\right)}}{1 + e^{\left(\sum_{j=0}^{p}\beta_j X_j\right)}}\right) \tag{7}$$

The derivative of Equation (7) set equal to zero does not yield an explicit result, so a numerical method, namely the Newton-Raphson iteration method, is required to obtain the parameter estimates. The Newton-Raphson method is an iterative approach used to solve nonlinear equations, such as solving the Likelihood equation in a logistic regression model.

### 1.6 Classification Accuracy

Classification accuracy measures the proportion of correctly classified observations. The Apparent Error Rate (APER) is calculated to evaluate the model's classification performance. The value of the Apparent Error Rate (APER) represents the proportion of samples classified incorrectly by the classification function, as shown in Table 2.2 below (Azies, 2017).

**Table 1**. Calculation of Classification Accuracy.

| Observation | Classification | |
|---|---|---|
| | $y_1$ | $y_2$ |
| $y_1$ | $n_{11}$ | $n_{12}$ |
| $y_2$ | $n_{21}$ | $n_{22}$ |

The calculation of the Apparent Error Rate (APER) is the proportion of observations misclassified by the classification function, as shown in Equation (8) bellow:

$$APER = (n_{11} + n_{22}) / n \times 100\% \tag{8}$$

## 2. METHOD

This study uses secondary data from the Academic Office of the Faculty of Mathematics and Natural Sciences, University of Bengkulu, for the 2017 student cohort.

## 3. RESULTS AND DISCUSSION

### 3.1 Descriptive Statistics

Study duration refers to the amount of time required by a student to complete their education, calculated from the time of initial enrollment to the thesis defense.

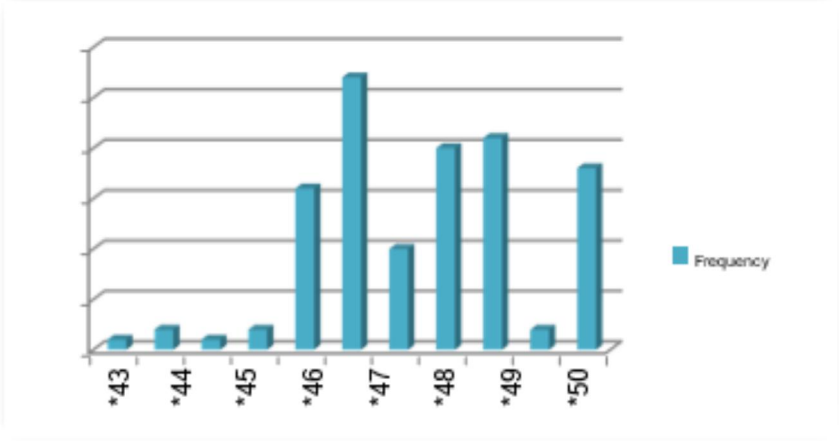


**Figure 1.** Frequency Distribution of Study Duration

Based on Figure 1, the lowest number of students who graduated occurred at 43 months and 45 months, with only 1 student or 0.83%. Meanwhile, the highest number of graduates occurred at 48 months, with 27 students or 22.5%.

### 3.2 GPA

Cumulative Grade Point Average (CGPA) is a number that reflects a student's academic performance or learning progress cumulatively, from the first semester to the final semester.

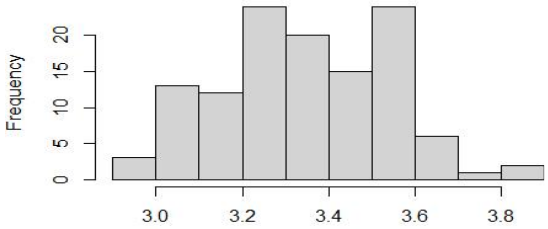


**Figure 2.** Histogram of GPA

Based on Figure 2, the average GPA of FMIPA UNIB students from the 2017 cohort is 3.348, with the highest GPA being 3.86, the lowest GPA being 2.9, and the median GPA being 3.35.

### 3.3 Student Admission Pathways

University admission pathway refers to the selection route through which each student gains entry into the university of their choice. In this study, the admission pathways are categorized into two indicators: independent and non-independent.

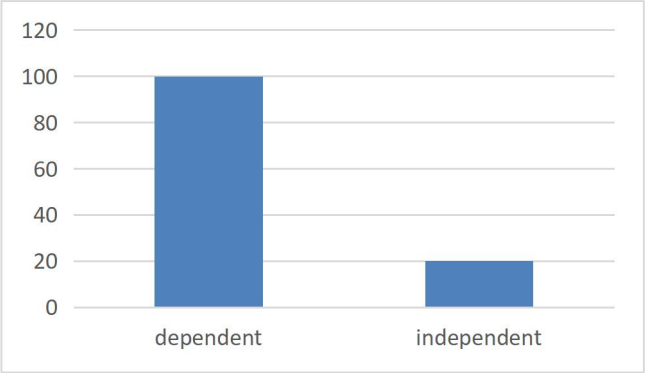


**Figure 3.**Frequency Distrbution of Admission Pathway

Based on Figure 3, the number of students admitted through the independent pathway is 20 students or 16.67%, while those admitted through the non-independent pathway total 100 students or 83.33%.

### 3.4 Scholarships

Scholarships are financial assistance provided to individuals with the aim of supporting their educational continuity (Utomo, 2011). In this study, the indicators used are students who have received a scholarship and students who have never received a scholarship.

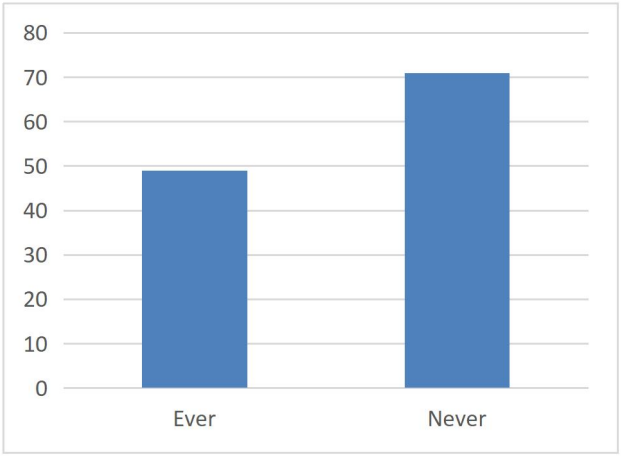


**Figure 4:** Frequency Distribution of Scholarship Recipients

Based on Figure 4.5, the number of students who have received a scholarship is 49 students or 40.83%, while the number of students who have never received a scholarship is 71 students or 59.17%.

### 3.5 Inpendence Test

The chi-square test of independence is a statistical method used to determine whether two categorical variables are independent of each other or not. The main purpose of the independence test is to examine whether the distribution of one variable differs significantly based on the values of another variable.

**Table 2**. Inpendence Test

| Variable | Chi-square | P-Value | Results |
|---|---|---|---|
| Duration of Study- GPA | 79.092 | 0.022 | There is correlated about duration of study with GPA. |
| Duration of Study-Scholarship | 0.886 | 0.3465 | There is not correlated duration of study with scholarship |
| Duration of Study-Student admission | 6.5304e-31 | 1 | There is not correlated duration of study with student admission pathways |

The table presents the results of chi-square tests for the relationship between study duration (LS) and several variables. The first row shows that there is a significant relationship between study duration and GPA (IPK), with a chi-square value of 79.092 and a p-value of 0.022. This indicates that study duration affects students' GPA. The second row demonstrates that there is no relationship between study duration and whether a student has received a scholarship (MB), as indicated by a chi-square value of 0.886 and a p-value of 0.3465. Lastly, the third row shows that study duration is not related to the admission pathway (JM), with a chi-square value of 6.5304e-31 and a p-value of 1, suggesting no association between these two variables.

### 3.6 Multicollinearity Test

Multicollinearity detection is one of the assumptions that must be considered in regression analysis. The predictor variables in regression analysis must be independent of each other. This can be observed from the VIF

value, which should be less than 10. This means that no multicollinearity exists among the predictor variables. Below are the results of the multicollinearity test:

**Table 3**. Output of Multicollinearity Test

| Variabel | VIF Value | Results |
|---|---|---|
| IPK ($X_1$) | 1.032 | Multicollinearity. |
| Scholarship ($X_2$) | 1.127 | Not Multicollinearity. |
| Student Admission Pathways ($X_3$) | 1.152 | Not Multicollinearity. |

Table 3 presents the results of the multicollinearity test based on the Variance Inflation Factor (VIF) values for each independent variable. The VIF value is used to detect the presence of multicollinearity, which occurs when independent variables in a regression model are highly correlated with each other. Generally, a VIF value below 10 indicates that multicollinearity is not a concern. In this table, the VIF value GPA (X1) = 1.032, Scholarship Status (X2) =1.127, and for the University Admission Path (X3) variable is 1.152. Since all VIF values are well below the threshold of 10, it can be concluded that there is no multicollinearity among the predictor variables. This means that the independent variables do not exhibit a strong linear relationship with each other and can be reliably included in the regression model without causing issues in parameter estimation.

### 3.7 Partial Test

Partial testing, also known as the partial significance test, is a statistical method used to determine whether an individual independent variable in a regression model has a significant effect on the dependent variable. This test evaluates each predictor variable separately while holding the other variables constant. The most common method for conducting a partial test is by using the t-test for linear regression or the Wald test in logistic regression. If the p-value obtained from the test is less than the chosen level of significance (commonly 0.05), it indicates that the corresponding independent variable significantly influences the dependent variable. In other words, the null hypothesis (which states that there is no effect) is rejected. On the other hand, if the p-value is greater than the significance level, the null hypothesis is not rejected, indicating that the variable does not have a statistically significant effect. Partial testing is essential for identifying which variables contribute meaningfully to the model and helps improve the accuracy and interpretability of the regression analysis.

**Table 4.** Output Statistcs Test of Partial for each variable

| Variabel | $\beta$ | SE | Wald | Db | $P-value$ |
|---|---|---|---|---|---|
| GPA ($X_1$) | 7.105 | 1.458 | 4,870 | 1 | $1.1\times 10^{-6}$ |
| Scholarship ($X_2$) | 0.280 | 0,471 | 0.600 | 1 | 0.551 |
| Student Admission Pathways ($X_3$) | 0.622 | 0.643 | 0.970 | 1 | 0.333 |
| *Constant* | -24.510 | 4.950 | -4.950 | 1 | $7.4\times 10^{-7}$ |

Table 4 presents the results of the partial hypothesis testing. The variable Cumulative GPA (IPK) ($X_1$) shows a decision to reject $H_0$, which means there is a significant effect of the GPA on the study duration of students. For the variable Scholarship Status ($X_2$), the decision is to accept $H_0$, indicating that there is no significant effect of receiving a scholarship on the study duration. Meanwhile, for the variable Admission Pathway ($X_3$), the decision to accept $H_0$ suggests that there is no significant effect of the admission pathway on the study duration of students. Thus, the logistic regression model formed based on the parameter testing results is as follows.

$$\pi(X) = \frac{exp\big(g(X)\big)}{1 + exp\big(g(X)\big)}$$

$$g(X) = -24.510 + 7.105X_1 + 0.280X_2 - 0.622X_3.$$

$$\pi(X) = \frac{exp(-24.510 + 7.105X_1 + 0.280X_2 - 0.622X_3)}{1 + exp(-24.510 + 7.105X_1 + 0.280X_2 - 0.622X_3)}$$

- Odds $exp(7.105) = 1218.042$
  Interpretation: An odds ratio of 1218.042 means that there is a 1218-fold increase in the likelihood of an event occurring in the dependent variable for every one-unit increase in the independent variable.
- Oddsratio= $exp(0.280) = 1.324$
  Interpretation: An odds ratio of 1.324 indicates that there is a 32.4% increase in the likelihood of an event occurring in the dependent variable for every one-unit increase in the independent variable.
- Odds ratio= $exp(0.622) = 1.861$
  Interpretation: An odds ratio of 1.861 indicates that there is an 86.1% increase in the likelihood of an event occurring in the dependent variable for every one-unit increase in the independent variable.

### 3.8 Clasificcation Accuracy

Here is the classification accuracy percentage for the study duration category of the 2017 FMIPA cohort at the University of Bengkulu.

**Tabel 5.** Output of Clasification Accuracy

| Observation | | Prediction | | |
|---|---|---|---|---|
| | | Study Duration Category | | Percentage Correct |
| | | Graduated on time | Graduated not on time | |
| Category | Graduated on time | 33 | 11 | |
| | Graduated not on time | 16 | 60 | |
| Total Percentage | | | | 77.5% |

Table 5 presents the output of classification accuracy in predicting student graduation timeliness. The table shows that out of the students who graduated on time, 33 were correctly classified, while 11 were misclassified as not graduating on time. Similarly, out of the students who did not graduate on time, 60 were correctly classified, and 16 were misclassified as graduating on time. The model achieved a 77.5% classification accuracy for predicting student graduation timeliness. The calculation of the Apparent Error Rate (APER) or classification accuracy for the male baby weight category is as follows.

$$APER = \frac{n_{11} + n_{22}}{n} \times 100\%$$

$$APER = \frac{33 + 60}{120} \times 100\%$$

$$APER = \frac{93}{120} \times 100\% = 77.5\%$$

## 4. CONCLUSION

Based on the results obtained, it can be concluded that the GPA on the study duration of students affects the study duration of students. It may be inferred from the odds ratio that a student's chances of graduating on time increase with their GPA.

## REFERENCES

[1] A. Agresti, Categorical Data Analysis, 2nd ed., New York: John Wiley & Sons, 2002.

[2] H. A. Azies, "Analisis Pelaku Hidup Bersih dan Sehat (PHBS) Rumah Tangga Penderita TB di Wilayah Pesisir Kota Surabaya Menggunakan Pendekatan Regresi Logistik Biner," Tugas Akhir, Fakultas Vokasi, Institut Teknologi Sepuluh November, Surabaya, 2017.

[3] D. N. Gujarati, Dasar-dasar Ekonometrika, Jakarta: Erlangga, 2007.

[4] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, New Jersey: John Wiley & Sons, Inc., 2000.
[5] N. A. J. Putra, P. K. Nitiasih, N. Adil, and G. Gunatama, "Identifikasi Faktor-Faktor Yang Mempengaruhi Lama Masa Studi Mahasiswa Di Fakultas Bahasa Dan Seni Undiksha," 2014.
[6] Y. L. D. Putri and B. Bustami, "Mengidentifikasi Faktor-Faktor Yang Mempengaruhi Lama Masa Studi Mahasiswa Menggunakan Regresi Cox Proportional Hazard," 2021.
[7] H. B. Rochmanto, "Analisis Faktor-Faktor Yang Mempengaruhi Petani Bawang Merah Untuk Menabung Di Bank Menggunakan Regresi Logistik Biner (Studi Kasus Di Desa Ngrami, Kecamatan Sukomoro, Kabupaten Nganjuk)," Tugas Akhir, Fakultas Vokasi, Institut Teknologi Sepuluh November, Surabaya, 2017.